HANDBOOK OF EPISTEMOLOGY

HANDBOOK OF EPISTEMOLOGY

Edited by

ILKKA NIINILUOTO
University of Helsinki, Finland

MATTI SINTONEN
University of Tampere, Finland

and

JAN WOLEŃSKI
Jagiellonian University,
Cracow, Poland

SPRINGER-SCIENCE+BUSINESS MEDIA, B.V.

A C.I.P. Catalogue record for this book is available from the Library of Congress.

ISBN 978-94-015-6969-9 ISBN 978-1-4020-1986-9 (eBook)
DOI 10.1007/978-1-4020-1986-9

Printed on acid-free paper

All Rights Reserved
© 2004 by Springer Science+Business Media Dordrecht
Originally published by Kluwer Academic Publishers in 2004
Softcover reprint of the hardcover 1st edition 2004

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

TABLE OF CONTENTS

INTRODUCTION

PART I: SOURCES OF KNOWLEDGE AND BELIEF

PART II: KNOWLEDGE ACQUISITION

PART III: TRUTH AND JUSTIFICATION

PART IV: EPISTEMOLOGY AND AREAS OF KNOWLEDGE

PART V: SPECIAL TOPICS

FOREWORD

Epistemology or theory of knowledge has always been one of the most important - if not the most important - field of philosophy. New arguments are constantly brought to bear on old views, new variants are marshalled to revive ancient stands, new concepts and distinctions increase the sophistication of epistemogical theories.

There are a great many excellent textbooks, monographs as well as anthologies consisting of articles in epistemology. Similarly, there are useful philosophical dictionaries which contain a great number of relatively short entries, and general philosophical handbooks which also touch epistemological issues. This volume of 27 essays grew out from the interest to see a handbook which is devoted entirely to the historical roots and systematic development of theory of knowledge. It is not intended to compete but to supplement the already existing literature. It aims at giving both beginners and more advanced students as well as professionals in epistemology and other areas of philosophy an overview of the central problems and solutions of epistemology. The essays are self-contained and stil often rather extensive discussions of the chosen aspects of knowledge. The contributions presuppose very little familiarity with previous literature and only a few of them require the mastery of even elementary logical notation. This, we hope, makes the volume also accessible to the philosophically interested wider audience.

The contributors were asked to provide substantial, up-to-date, self-contained and balanced surveys of the various subareas and more specific topics of epistemology, with reference to literature. It was also suggested that each entry should be initiated with a short historical introduction to the problem area. Although the authors were asked to give a fair treatment to views which they themselves do not favour, they were also asked to voice their own views. This can be seen in the current volume, and as editors we have not even tried to create consensus where none exists. This way of proceeding is of course inevitable in philosophy. We have welcomed discussion and even passionate views, and only wish that we are not held responsible for the views of the contributors.

The volume starts with an historical introduction to epistemology, and there are sections for such traditional systematic topics as the sources of knowledge and belief, knowledge acquisition, truth and justification. Apart from these we stil wished to ive plenty of space for the various areas in the kingdom of knowledge, such as science, mathematics, the humanities and the social sciences, religion, and language. Similarly, we stil wanted to give voice to some traditional and more recent special topics in epistemology, such as evolutionary epistemology, relativism, the relation between epistemology and cognitive science, sociology of knowledge, epistemic logic, knowledge and art, and feminist epistemology.

This volume has taken a long time to complete. We want to thank all contributors not just for insightful entries but also for their patience and understanding during the process. Our special thanks goes to Professor Jaakko Hintikka who has supported the project from the start, and to Professor Robert Audi who not only agreed to write one of the key entries but also gave valuable advice on the entire project.

When editing this book for publication we have been assisted by a number of graduate students from the Departments of Philosophy and of Moral and Social Philosophy at Helsinki University. We would like to thank George Gebhard, Tomi Kokkonen, Taneli Kukkonen, Erika Mattila, Sami Paavola, Timo Viitala and Juhani Yli-Vakkuri in particular for their help in correspondence, desk editing and more generally in preparing the handbook for publication.

Helsinki, August 2003

Ilkka Niiniluoto, Matti Sintonen, Jan Woleński

CONTRIBUTORS

Robert Audi is Charles J. Mach Distinguished Professor of Philosophy at the University of Nebraska, Lincoln. He is the author of numerous books and articles in epistemology, ethics, action theory, and philosophy of religion. Most recent book is *The Architecture of Reason*, Oxford UP, January, 2001.

David Bloor is Director of the Science Studies Unit at the University of Edinburgh, and holds a personal chair in the Sociology of Science. He has published extensively on the sociology of scientific knowledge, and also on the philosophy of Wittgenstein.

Michal Bradie is Professor of Philosophy at Bowling Green State University. He has published extensively on evolutionary epistemology and evolutionary ethics. He is the author of The Secret Chain: Evolution and Ethics. He is currently working on a project involving the role of models and metaphors in scientific theorizing. (Email: mbradie@bgnet.bgsu.edu)

Marian David graduated from the University of Arizona, and is currently Associate Professor of Philosophy at the University of Notre Dame. He is the author of *Correspondence and Disquotation: An Essay on the Nature of Truth* (Oxford 1994) and co-editor of *Grazer Philosophische Studien.* Current research interests include the philosophy of language/mind and the role of truth in epistemology. (e-mail: David.1@nd.edu)

Susan Haack is Cooper Senior Scholar in Arts and Sciences, Professor of Philosophy and Professor of Law at the University of Miami. The author of *Philosophy of Logics*; *Evidence and Inquiry*; *Deviant Logic, Fuzzy Logic*; and *Manifesto of Passionate Moderate.* Haack works on logic, language, pragmatism, metaphysics, epistemology, philosophy of science, and scientific evidence in court.

Sven Ove Hansson is professor of Philosophy at the Royal Institute of Technology, Stockholm. He is the author of *Setting the Limit.* (Oxford University Press 1998), *A Textbook of Belief Dynamics. Theory Change and Database Updating* (Kluwer 1999), and *The Structures of Values and Norms* (Cambridge University Press, in press).

Paul Humphreys is Professor and Chairman in the Corcoran Department of Philosophy at the University of Virginia. His research interests lie primarily in the philosophy of science, metaphysics, and epistemology, and include computer modelling, strategic reasoning, emergence, causation, and explanation. He is currently completing a book on how contemporary instruments and computation have extended the domain of empiricism. He is also editor of the series Oxford Studies in the Philosophy of Science. (e-mail:pwh2a@virginia.edu)

Kent Johnson, Philosophy, Rutgers University, works primarily in philosophical linguistics and logic.

Kevin T. Kelly is Professor of Philosophy at Carnegie Mellon University. His research interests include epistemology, philosophy of science, formal learning theory and computability. (e-mail: kk3n@andrew.cmu.edu)

Markus Lammenranta is Docent and Senior Assistant of Theoretical Philosophy at the University of Helsinki. His current research interests include classical problems of epistemology and metaphysics. (email: markus.lammenranta@helsinki.fi)

Kathleen Lennon is senior lecturer in philosophy at the University of Hull UK. Her research interests are philosophy of mind and action, feminist epistemology and gender theory. (email: K.Lennon@phil/hull.ac.uk)

Wolfgang Lenzen has been Professor of Philosophy at the University of Osnabrück, Germany, since 1981. He has worked on Philosophy of Science ("Theorien der Bestätigung", 1972), on Philosophical Logic ("Recent Work in Epistemic Logic", 1978; "Glauben, Wissen und Wahrscheinlichkeit", 1980), on Leibniz ("Das System der Leibnizschen Logik", 1990), and on Applied Ethics ("Liebe, Leben, Tod", 1999). His current field of research is Philosophy of Mind.

Ernest LePore is Director of Center for Cognitive Science, Rutgers University, and works primarily in philosophy of language and mind. He is co-author, with Jerry Fodor, of *Holism* (Blackwell, 1991) and author of *Meaning and Argument* (Blackwell, 2000).

Joseph Margolis is Laura H. Carnell Professor of Philosophy at Template University. His current interests include the analysis of materialism and realism and their alternatives and the systematic differences between the physical and human sciences.

Roman Murawski is professor at Faculty of Mathematics and Computer Science of Adam Mickiewicz University in Poznan, Poland. He is the head of the Department of Mathematical Logic. His research areas are: mathematical logic and the foundations of mathematics, philosophy of mathematics and history of logic. (e-mail: rmur@math.amu.edu.pl)

Ilkka Niiniluoto has worked since 1977 as Professor of Theoretical Philosophy at the University of Helsinki. He graduated in mathematics, and made his PhD dissertion in 1973 on inductive logic. His current research interests include truthlikeness and critical scientific realism. (e-mail: ilkka.niiniluoto@helsinki.fi)

David Novitz is Reader in Philosophy at the University of Canterbury, New Zealand. He has published extensively in the philosophy of art, and has a special interest in epistemology and art.

Frederick F. Schmitt is Professor of Philosophy at the University of Illinois at Urbana-Champaign and Indiana University. His research interests include metaphysics and epistemology. He is the author of *Knowledge and Belief* (Routledge 1992) and *Truth: A Primer* (Westview 1995), and the editor of *Socializing Epistemology* (Rowman and Littlefield 1994).

Robert K. Shope is Professor of Philosophy at the University of Massachusetts in Boston. His research interests include epistemology, philosophy of mind, philosophy of psychoanalysis and issues concerning meaning and representing. (e-mail: robert.shope@umb.edu)

Harvey Siegel is Professor of Philosophy at the University of Miami. Current research interests include topics in epistemology, philosophy of science, and philosophy of education. (email: hsiegel@miami.edu)

Stephen Stich is Board of Governors Professor of Philosophy and Cognitive Science at Rutgers University. He is author of *From Folk Psychology to Cognitive Science* (MIT Press, 1983), *The Fragmentation of Reason* (MIT Press, 1990) and *Deconstructing the Mind* (Oxford University Press, 1996).

Tom Stoneham is Lecturer in Philosophy at the University of York and was formerly a Fellow of Merton College, Oxford. He has published on the philosophy of mind and has just finished a book about Berkeley (OUP, 2001). (e-mail: stoneham@ukonline.co.uk)

Paul Weirich is Professor of Philosophy at the University of Missouri-Columbia. Current research interests include epistemological issues in decision theory and game theory. Recent books are *Equilibrium and Rationality: Game Theory Revised by Decision Rules* (CUP, 1998) and *Decision Space: Multidimensional Utility Analysis* (CUP, forthcoming). (e-mail: weirichp@missouri.edu)

Jan Woleński is Professor at Jagiellonian University, Institute of Philosophy, Department of Epistemology. He is interested in the history of logic and analytic philosophy, in particular logic in Poland, philosophical logic, in particular, applications of formal methods in philosophy, and epistemology. Main publications: *Logic and Philosophy in the Lvov-Warsaw School* (1989), *Essays in the History of Logic and Logical Philosophy* (1999).

Keith E. Yandell (PhD., Ohio State) is Professor of Philosophy at the University of Wisconsin, Madison, U.S.A. where he has taught since 1966. He has published *Basic Issues in the Philosophy of Religion* (Allyn and Bacon), *Christianity and Philosophy* (Eerdmans), *Hume's "Inexplicable Mystery"* (Temple U. Press), *The Epistemology of Religious Experience* (Cambridge U. Press), and *The Philosophy of Religion* (Routledge). He is under contract to write *The Comparative Philosophy of Religion* (Kluwer) and *The Soul* (Ashgate).

Introduction

JAN WOLEŃSKI

THE HISTORY OF EPISTEMOLOGY

1 INTRODUCTION

Although there are many different classifications of philosophical problems, the division of philosophy into ontology (or metaphysics), epistemology, and axiology (ethics and aesthetics) still seems the most efficient and general one. Thus, epistemology belongs to the main parts of philosophy. However, the terms which now denote this field, namely 'epistemology' and 'theory of knowledge', appeared not very long ago, later than terms indicating metaphysics, ethics, aesthetics or even ontology. As late as in the 17th century there was no single word referring to epistemology. At that time as well as in the 18th century, epistemological problems were considered in books like (I give the English titles) *Rules for the Direction of Mind* (René Descartes), *An Essay Concerning Human Understanding* (John Locke), *A Treatise Concerning the Principles of Human Knowledge* (George Berkeley), *An Enquiry Concerning Human Understanding* (David Hume), *New Essays on Human Understanding* (Gottfried Leibniz) or *Critique of Pure Reason* (Immanuel Kant). Kant placed his central epistemological views under the label 'transcendental aesthetic', following the meaning of *aisthesis* as referring to cognition by senses. As a matter of fact, Kant also used (in his *Critique of Aesthetic Judgement)* the term 'aesthetics', more precisely, its German counterpart *Aesthetik*, in a more contemporary fashion, i.e., to denote matters of beauty. Earlier, Alexander G. Baumgarten in his *Sciagraphia encyclopaediae philosophicae* (1769) proposed the word *gnoseologia*, which gained some popularity and is sometimes employed even now. The German word 'Erkenntnistheorie' (theory of knowledge) became popular after Eduard Zeiler's influential paper "Bedeutung und Aufgabe der Erkenntnistheorie" (1862), but this name and its cognates were used earlier. Thomas Krug's, *Allgemeine Handwörterbuch der philososophischen Wissenschaften* (1827) proposed the label 'Erkenntnislehre'. Ernst Reinhold (the son of Karl L. Reinhold, a leading post-Kantian philosopher) in *Versuch einer neuen Theorie der menschlichen Vorstellung-svermögen und Metaphysik* (1832) had the term "Theorie der Erkenntnis". It was James E Ferrier who introduced the label 'epistemology' in his *Institutes of Metaphysics* (1854). Other words were also proposed to baptize our field: 'Wissenschaftslehre' (Johann G. Fichte, Bernard Bolzano), 'Wissenschaftstheorie' (Eugen Dühring), 'criterology' (Neo-Thomists), and 'noetics' (also Neo-Thomists). However, the words 'epistemology' and 'Erkenntnistheorie' (as well as their translations into other languages) are most popular nowadays.

The terminological variety noted above is not incidental and displays different ideas attached to epistemological concern. If epistemology is understood extensively, it covers everything that focuses on knowledge or cognition:

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 3-54.
© 2004 *Kluwer Academic Publishers.*

psychology, sociology, logic, history, physiology, pathology, axiology, metaphysics, and several other things. On the other hand, epistemology conceived more restrictively investigates the sources, values (cognitive), principles, and limits of knowledge. This general characterization can be made more detailed by further explanations, for example:

> "[Epistemology] [...] The theory of knowledge. Its central questions include the origin of knowledge, the place of experience in generating knowledge, and the place of reason in doing so; the relationship between knowledge and certainty, and between knowledge and the impossibility of error; the possibility of universal [...] scepticism; and the changing forms of knowledge that arise from new conceptualizations of the world. All of these issues link with other central concerns of philosophy, such as the nature of truth and the nature of experience and meaning. It is possible to see epistemology as dominated by two rival metaphors. One is that of building or pyramid, built on foundations. In this conception it is the job of the philosopher to describe especially secure foundations, and to identify secure modes of construction, so that the resulting edifice can be shown as to be sound. This metaphor favours some idea of the 'given' as a basis of knowledge, and of a traditionally defensible theory of confirmation and inference as a method of construction [...] The other metaphor is that of a boat or fuselage, that has no foundations but owes its strength to the stability given by its interlocking parts. This rejects the idea of a basis in the 'given', favours ideas of coherence and [...] holism, but finds it harder to ward off [...] scepticism." (S. Blackburn, *The Oxford Dictionary of Philosophy,* Oxford University Press, Oxford 1994, p. 123).

The typical epistemological problems are like the following: What is knowledge?; Is knowledge based on senses or reason? Is certainty attainable? What is truth? Are there ultimate limits of knowledge? Although it is difficult to delimit sharply both ways of understanding epistemology, classical epistemological questions form a relatively stable tradition which can be sufficiently identified through history.

This chapter is intended as a historical survey of epistemology, basically in its restrictive understanding, but taking into account its relationships with other philosophical disciplines and fields outside philosophy. Since the size of this text is limited, the history of epistemology given here must be concise. I will try to stress those facts from the history of epistemology which had a real historical significance, especially for contemporary discussions, in particular within the analytical turn of philosophy. Hence, I must omit many interesting details as well as positions belonging to other philosophical traditions (this restriction is perhaps the most relevant with respect to the last chapter). However, it does not mean that non-analytical epistemological thinking is entirely neglected, also because the borderline between analytical and non-analytical philosophy is imprecise in many respects. I will particularly focus on post-Cartesian philosophy. Here is the reason: One can ask which part of philosophy should be taken as *the* starting point for the whole philosophical enterprise. According to the tripartite division of philosophy into ontology, epistemology, and axiology, three possibilities appear, and, in fact, each of them has been executed in the history of philosophical thought. Leaving aside axiologically oriented philosophy (although, as we will see, it was sometimes very important in the history of epistemology), the development of philosophy can be divided into two periods. Roughly speaking and admitting some exceptions which I will not mention here, pre-Cartesian philosophy was definitely ontologically oriented, but post-Cartesian thought became largely preoccupied with epistemology. In this sense, Descartes is the father of modern philosophy. In fact, *cogito, ergo sum,* whatever it is (a principle, inference or performance), clearly suggests that an ontological statement (I am) is based on an epistemological datum (I think). Similarly, Berkeley's *esse = percipi* may be interpreted as an attempt to define an ontological category (existence, being) by an epistemological one (perception). Although it is possible that post-Kantian philosophy is more balanced with respect to the relation between ontology and epistemology, Descartes' philosophy certainly is an important turning point in the history. This justifies my plan.

Let me also note that my survey does not cover the latest period (roughly speaking, after World War II, except the later British analytic philosophy) of the development of epistemology. It means that fairly recent epistemology is out of the scope of this chapter. The reason is that I do not intend to interfere with historical remarks made by other authors of this volume which, as a whole, is simply a report on epistemology at the present stage of its development. Finally, let me say a word about references to this chapter. All references occur in the main text and mention only the titles and dates of some work; most titles are given in English. In order to make the history of epistemology better accessible from this survey, I include dates of lives of most philosophers mentioned. The bibliography at the end of this chapter lists a selection of works exclusively devoted to the entire history of epistemology, its particular periods or major problems; no works about the views of particular thinkers are mentioned. The reader interested in further details can consult bibliographies attached to particular items in *The Encyclopedia of Philosophy.* ed. by P. Edwards, Macmillan, New York 1967 and *The Routledge Encyclopedia of Philosophy,* ed. by E. Craig, Routledge, London 1998, as well as references in other papers in this volume. Moreover, everybody interested in the development of philosophical ideas should consult *Historisches Wörterbuch der Philosophie,* ed. by J. Ritter et al., v. 1-10, Benno Schwabe-Wissenschaftliche Buchgesellschaft, Basel-Darmstadt 1971-1999, further volumes in preparation. Useful historical information is also included into *A Companion to Epistemology,* ed. by J. Dancy and E. Sosa, Blackwell, Oxford 1992. In many cases I use terminology derived from Kazimierz Ajdukiewicz, *Problems and Theories of Philosophy,* trans. by H. Skolimowski and A. Quinton, Cambridge University Press, Cambridge 1973 which seems to me the clearest introduction to philosophy. Last but not least, I want to express my debt to Timothy Childers (the Czech Academy of Sciences, Institute of Philosophy) for careful reading of this text and extensive comments, which lead to numerous fundamental revisions and improvements, stylistic and substantial as well.

1. ANCIENT PHILOSOPHY

All accessible sources document that the first Ionian philosophers were almost wholly dominated by problems of metaphysics and cosmology. The first epistemological remarks can be found in Heraclitus (the dubious value of senses), the Pythagoreans (the theories of direct cognition) and the Eleats (Parmenides: the identity of thinking and being) in the 6th century B. C. Perhaps the invention of the deductive method by the Pythagoreans and the Eleats had the most significance for the further development of epistemology; due to this discovery, Parmenides could develop a very radical rationalism. Epistemology was further pushed by Empedocles (ca.490-ca.430) in his conception of sense-cognition and Anaxagoras (ca.500-

ca.428/27) who introduced the concept of *nous* (reason, but rather global than individual). The first more complete conceptions of knowledge were elaborated by Democritus (ca.460-ca.360; a version of critical realism, together with the distinction between primary and secondary qualities) and the Sophists, particularly Protagoras (481-411; relativism, conventionalism, pragmatism, elements of scepticism). Finally, Socrates (469-399) stressed the role of general concepts in knowledge.

It was Plato (427-347) who derived far-reaching consequences from Socrates' view about generality. First of all, Plato defined knowledge *(episteme)* as a true justified belief which was contrasted by him with a mere opinion *(doxa)* This distinction explained the possibility of error, because only an opinion, not knowledge, may be erroneous. Plato's theory of knowledge was closely related to his ontological conception. Knowledge as a distinguished cognitive state has its own object, namely Forms. Thus, knowledge must be general, because Forms are such; knowledge is of course absolutely certain. Opinion is generated by senses, concerns changing things and is at most probable, never certain. The famous metaphor of the cave in the *Republic* illustrates well Plato's view about the cognitive situation of human beings. We have two different worlds and two different ways of access to them: by reason to Forms and by senses to things. As a matter of fact, only the world of Forms is truly real, and this essentially contributes to Plato's view that only *episteme* is genuine knowledge. Plato's view on opinion in *Republic* was more complex. He distinguished ignorance which is objectless and belief with the sensible world (something between existence and non-existence) as its object. Thus, belief has some shadow of knowledge.

Plato divided *episteme* into two kinds: intuitive *(noesis)* and discursive *(dianoia)*. The latter is modelled by mathematics and it is somehow restricted in its value. *Noesis* is the highest form of knowledge; it is the faculty which leads us to Goodness. It shows that Plato, although strongly influenced by mathematics, did not attribute it the highest cognitive value. Plato understood that this picture required completion by a conception of the origin of knowledge. He proposed a myth of metempsychosis. According to Plato, the soul which is the real knowing subject is immortal and embodied in various human beings. The soul, when it is outside a body, lives in the world of Forms and has direct cognitive access to them. The epistemic acts of human beings consist in recollection *(anamnesis)* of the knowledge acquired by souls due to their participation in the world of Forms. This situation is illustrated in the dialogue *Meno* in which a slave boy discovers a geometrical theorem without any prior knowledge of geometry.

Plato introduced several important epistemological insights. Even if the dualism of senses and reason was present before Plato, he completed this distinction, made it sharp and derived far-reaching consequences from it. Plato developed the first full-blooded radical rationalism. This view appeared in Plato in two dimensions: as apriorism (methodological rationalism), i.e., the view that only intellectual cognition is valuable, and as nativism (genetic rationalism), that is, the view that knowledge is inborn. He was also a radical foundationalist. On the other hand, Plato's epistemological views gave rise to several difficult questions which are ever now discussed. The definition of knowledge as true justified knowledge is a source of constant trouble in epistemology. To be fair, Plato himself was conscious of difficulties arising in this context, in particular, that it leads to the rejection of opinion as a form of knowledge: this was his reason for introducing the distinction between ignorance and belief. Since most contemporary epistemologists do not agree with this restrictive view of knowledge, we encounter a very characteristic ambiguity of the word 'knowledge'. On the one hand, many philosophers divide knowledge into *episteme* and *doxa,* but, on the other hand, some try to defend Plato's definition of knowledge and apply it to *doxa.* It usually leads to serious difficulties. Plato was also a predecessor of irrationalism, because his concept of *noesis* can be (in fact, it was) interpreted as referring to a kind of contemplation.

Aristotle (384-322), a student of Plato, rejected the apriorism and nativism of his teacher. Instead, the Stagirite developed instead an empirical account of knowledge. According to Aristotle, knowledge always begins with sense experience concerning particular substances and is *a posteriori.* This view as well as methodological empiricism (aposteriorism) concerning cognition was conjoined by Aristotle with the moderating role of reason. It was necessary because Aristotle retained Plato's idea of *episteme.* For Aristotle, knowledge (as *episteme)* has forms (as components of substances, not as Platonic independent entities) as its objects. Aristotle used here his theory of substances as entities composed of matter and form. Since forms are always in individual substances, we grasp things as instances of general essences. In more recent terminology, we perceive particulars as instances of universals. Basically, the process of grasping consists in abstraction performed by an active capacity to judge which is imposed upon passive perception.

The complex structure of cognitive acts explains, according to Aristotle, how it is possible to form general propositions on the base of *a posteriori* knowledge. First evident principles are the starting point of theoretical (scientific) knowledge which proceeds by chains of logical deductions based on syllogisms (recall that Aristotle invented logic as an independent science and developed the theory of the syllogism). Thus, science as a result of knowledge forms an assertive-deductive system with evident axioms at its beginning. Here we have the picture of scientific method that was dominant until the Renaissance, i.e., for almost twenty centuries. Euclides' *Elements,* Ptolemey's *Almagest* were perhaps the highest applications of Aristotle's methodological ideas. Although it is true that Aristotelian methodology became an obstacle for the development of science in the Middle Ages and later, Aristotle himself cannot be accused as being guilty for this situation; as a theoretician of science, he did his best, perhaps even more. Of other epistemological views of Aristotle's, his conception of truth was particularly important. He defined truth in many places, but two statements became the most influential:

(a) "To say of what is that it is not, or of what is not that it is, is false, while to say of what is that its is, and of what it is not that it is not, is true." (*Metaphysics* 1011 b)

(b) "[...] he who thinks the separated to be separated and the combined to be combined has the truth, while he whose thought is in a state contrary to that of the objects is in error." *(Metaphysics* 1051 b).

Other epistemological problems considered by Aristotle concerned, for example, the value of probable inferences (in *Topics)* and practical knowledge *(phronesis).* The Stagirite essentially contributed to epistemology as a genuine part of philosophy. He not only offered new solutions (empiricism) and new devices (mature logic), but also widened the scope of epistemology (practical knowledge, probable arguments). Although Aristotle's empiricism is at odds with Plato's rationalism, both great Greeks had something in common in their epistemology, namely the concept of *episteme.* Aristotle distinguished four kinds of cognitive activities: perception, memory, experience, and scientific knowledge, i.e., just *episteme* starting from evident principles and proceeding by deduction. Contrary to Plato, he did not denigrate perception, memory or experience as cognitively devoid of any value, but, on the Aristotelian view, *episteme* was essentially superior to any other kind of cognition.

The post-Aristotelian hellenistic schools, which appeared in the end of the 4th century B. C., were preoccupied with ethics, particularly with the question of the availability of happiness. For Epicureans, Stoics, and Sceptics other philosophical problems were simply subordinated to the ethical enterprise. The Epicureans developed a very radical sensualist epistemology consistent with their materialism. The Stoics were also empiricists, but they admitted intuition *(katalepsis).* Scepticism was perhaps the purest epistemological current of ancient philosophy. Ironically, although sceptical maxims proposing ways to achieve happiness are only of historical significance, the sceptical challenges have become one of the most stable ingredients of epistemology. For scepticism, epistemology was simply an introduction to ethics and had no intrinsic value. In order to show that happiness requires abstaining from decisive statements, the ancient sceptics invented several arguments intended to prove that knowledge is impossible. Knowledge is gained either directly or indirectly. Direct knowledge (knowledge by perception) is impossible for ten reasons (so called *tropes,* according to Aenesidemus): (1) different living creatures perceive objects differently; (2) different human beings perceive objects differently; (3) different senses give different perceptions; (4) the same senses give different perceptions depending on various circumstances; (5) perception sometimes depends on distances and locations of perceived objects; (6) perception of an object is often mediated by some other objects, for example air; (7) perception can depend on quantitative properties of perceived objects and their composition; (8) perception is involved in several relations between the perceiver and the perceived objects; (9) perception can depend on expectations; (10) perception can depend on social factors, for example, education or religion. Now consider the status of indirect knowledge, that is, knowledge gained by inference. It can be deductive or inductive. However, deductive inference is plagued either by *petitio principii* or *regressus ad infinitium* or must appeal to premises justified by direct knowledge. Results burdened by *petitio* or *regressus* cannot be regarded as genuine knowledge, because both situations are logically unacceptable. Any appeal to direct knowledge is equally dubious because of the tropes mentioned. Thus, deductive inference does not provide knowledge. The situation of inductive inference is not better, because its premises do not provide full justification of inductive conclusions. Since direct knowledge, indirect knowledge by deduction, and indirect knowledge by induction exhaust all available generators of knowledge, scepticism considers the thesis asserting the impossibility of knowledge as demonstrably proved. The classical Sceptic, for example, Pyrrho of Ellis (ca.376-ca.286) recommends the following strategy. All possible statements are isostenic (have an equal epistemic value); in particular, if *A* is a statement, not-*A* is a statement, both are isostenic. The best we can do consists in abstaining from decisive opinions. Thus, the Sceptic says: I do not know that *A* and I do not know that non-*A*, and contrasts this attitude with that of dogmatic and academic philosophers. We find the *locus classicus* of this view in the following words of Sextus Empricius (2nd century B. C.):

> "The natural result of any investigation is that the investigators either discover the object of search or deny that it is discoverable and confess it to be inapprehensible or persist in their search. So, too, with regard to the objects investigated by philosophy, this is probably why some claimed to have discovered the truth, others have asserted that it cannot be apprehended, while others again go on inquiring. Those who believe they have discovered it are the 'Dogmatists', specially so called – Aristotle, for example, and Epicurus and the Stoics and certain others; Cleitomachus and Carncades and other Academics treat it as inapprehensible; the Sceptics keep on searching." *(Outlines of Pyrronism* I, 1-4.)

It is important to see the difference between the Academician, for example, Carneades (214-129) and the Sceptic. The latter does not assert anything, the Academician asserts that nothing is true; yet both protest against the claim of the Dogmatist that truth is attainable, and this is the reason why both are counted as sceptics in the wide sense.

There is a standard objection against scepticism and academism, raised by Clement of Alexandria (ca.l50-ca.215), one of the first Christian philosophers. Let *S* be a statement expressing the sceptical or academic view. Now we ask: what about *S* itself? If it is isostenic with respect to negation (scepticism) or asserted as not true (academism) it looses its strength, according to Clement and similar critics. The Pyrronists explained that they expressed their views as guesses or posits. The answer given by Carneades was that all statements should be regarded as merely probable. In fact, the thesis that no truth is attainable is consistent with probabilism. Moreover, the ancient sceptics sometimes accepted statements *de se* as epistemically legitimate. Thus, according to this view, scepticism is not a thesis about the external world, but about perceiving human beings. This point shows an interesting aspect of scepticism. It seems that ancient scepticism criticized the concept of *episteme* as something concerning the external world, but admitted considerations of cognitive activities from a subjective point of view. Putting it in another words: scepticism rejects knowledge as *episteme.* but tolerates epistemology. Thus, on the sceptical view, everybody who admitted knowledge of the external world as *episteme* was dogmatic, independently of whether it was generated by reason or senses: Plato, Aristotle, the Epicureans, and the Stoics belonged, according to scepticism, to the dogmatic variety. It shows how the concept of *episteme* was widespread in ancient philosophy. Scepticism is important not only because it challenged epistemologists. From the contemporary point of view, the main merit of scepticism consisted in introducing an alternative epistemology with relativism, probabilism, anti-foundationalism, and coherentism as the main points. Thus, both epistemological metaphors mentioned by Blackburn can be applied to ancient epistemology: dogmatism falls under the metaphor of a pyramid with solid foundations, but scepticism favours the allegory of a boat. However, a warning is here in order: the difference between the two positions, particularly as described in modern terms, was not perceived in this way by ancient philosophers, because they were more interested in practical issues than in theoretical epistemology. The axiological orientation of scepticism probably prevented its representatives from a fuller development of fallibistic epistemology.

The fall of Antiquity brought the irrationalism of Neo-Platonism, particularly in Plotinus (ca.205-270). It is not surprising if we recall that Plato himself was close to irrationalism. Plotinus replaced *noesis* by non-verbalized contemplation directed to the One, a counterpart of the World of Forms. Since Neo-Platonism strongly influenced early Christian philosophy, elements of irrationalism became its standard ingredients. Augustine of Hippo, the last great ancient philosopher (or the first medieval philosopher, if you like) completed the Platonic version of Christianity. An important point derived by Augustine (354-430) and all later Christian philosophers was that human beings could be successful in knowledge of God, and that the revelation had to be accepted as an unquestionable source of knowledge. Augustine followed Plato and Neo-Platonism in nativism, although with some modifications demanded by religious principles: *ideae innatae* are the ultimate results of God's creation. Augustine proposed the conception of *illuminatio* (enlightenment) as a condition of knowledge. *Illuminatio* is a result of the God's free grace, something which human beings obtain or not, independently of their merits. However, Augustine was not a radical irrationalist. He rejected Tertulian's dictum *credo qua absurdum* (I believe, because it is an absurd) as a correct account of the status of religious belief. Augustine, although he considered faith as something infinitely superior to reason, aimed at an agreement between both. Thus, he was a predecessor of the view expressed by a famous formula *fides quaerens intellectum* (faith looking for understanding). Augustine was also a predecessor of some important later views. In particular, he introduced a voluntarist account of judging, and his syllogism *dubito, ergo sum* anticipated Descartes' *cogito, ergo sum,* but this connection was not perceived for a long time.

2. MEDIEVAL PHILOSOPHY

Most philosophical problems considered in the Middle Ages directly or indirectly concerned relations between faith and reason or theology and philosophy. The early solution proposed by Dionysius the Pseudo-Aeropagite (ca.500) was fairly Neo-Platonic with a mystical stance. According to him, human capacities are too restricted in their cognitive powers to produce knowledge of divine matters. This view resulted with an idea of negative theology: human beings can know what God is not, but positive knowledge exceeds their cognitive capacities. Dionisius' views became popular due to translation of his work into Latin by John Scot Eriugena (ca.810-ca.877), the most remarkable thinker of the so called Dark Ages (5th century-10th century). Eriugena retained the main principle of negative theology, but, contrary to Dionysius, tried to reconcile reason and faith by a pantheistic view that human beings are manifestations of God. Thus, Eriugena made a step toward a more rationalistic account of the relation between theology and philosophy.

The Augustinian principle *fides quaerens intellectum* was reintroduced by Anselm of Canterbury (1033-1109). Anselm was fully convinced that there was harmony and coherence between theology and philosophy. However, according to his view, a full understanding of theology requires the intervention of reason. Perhaps Anselm's ontological proof of the existence of God is the most impressive sign of his theological rationalism. Anselm's optimism concerning the natural harmony between theology and philosophy was doubted by Peter Abelard (1079-1142). He was also a rationalist but of a different kind than Anselm. Abelard was the first great practicioner of the scholastic method understood as consideration of contradictory opinions in order to achieve a proper solution. His work *Sic et Non* lists several incoherences from Holy Scripture as well as earlier writings of theologians and philosophers. According to Abelard, nothing is free of rational doubt. Hence, it is not proper to assume an initial validity of theological authorities. He demanded rational solutions of the contradictions he raised. Perhaps it was the first example of critical rationalism. Moreover, Abelard offers conceptualism as a solution of the problem of universals. The problem of universals principally belongs to ontology, but it always had a definite significance for epistemology (see the section on Aristotle below).

European medieval philosophy was strongly influenced by Muslim philosophers, because Arabs transmitted a great deal of ancient philosophy to Europe, but for other reasons as well. Islamic philosophers became influential commentators on and interpreters of ancient masters, particularly Aristotle. Ibn Sina (Avicenna) (980-1037) and Ibn Rushd (Averroes, the Commentator) (1126-1198) were important for the development of medieval Aristotelism with a more empirical flavour. Moreover, Averroes formulated the thesis of the superiority of philosophy over theology; this view was later transformed into the so called Latin Averroism (the theory of double truth). In order to complete this brief excursion into Islamic philosophy, let me mention that Al-Farabi (ca.870-950) defended the priority of faith over reason. Thus, the principal solutions of the problem how theology was related to philosophy were developed inside Arabic philosophy.

The 13th century was the golden age of medieval philosophy. Albert the Great (ca. 1200-1280) and Thomas Aquinas (1225-1274) succeeded with a synthesis of Christianity and Aristotelism. Thomas followed Aristotle's epistemology in all essential points. In particular, he accepted genetic empiricism which was captured by a Latin formula *nihil est in intellectu, quod non prius fuerit in sensu* (nothing is in the intellect unless it first appeared in the senses). Perhaps the important innovation was that the first principles were necessary propositions. This view was rather a strengthening of Aristotle than a rival account. However, Aquinas could not use directly the Stagirite in solving the theology/philosophy problem, because this question did not exist in ancient philosophy. Aquinas' solution is as follows. There are theological truths which are inaccessible for rational demonstration, for example the dogma about *creatio ex nihilo,* On the other hand, we have theological truths which are logically provable, for example the existence of God. Thomas' Five Ways of proving God's existence differ from Anselm's ontological argument. Aquinas proofs are basically Aristotelian in their spirit: they start from premises which assert something about the world (for example, every being has its cause), then proceed by metaphysical principles (for example, any series of causes must terminate), and end with the statement asserting that God exists. These proofs reveal Aristotelian empiricism by their appeal to premises about the real world. However, it is also clear why Aquinas insisted so strongly that the first principles should be necessary. Finally, we have also truths of reason which have no theological import. Now, it is a question why a genuine contradiction between theology and philosophy is impossible. According to Aquinas, it is so because philosophy and theology are given by God who cannot create inconsistencies. Hence, any alleged contradiction is merely temporary and sooner or later will be resolved by the human mind. Although this view is not literally Aristotelian, its general spirit is such, because it aims at a compromise between theology and philosophy.

Jan Fidanza (Bonaventure) (ca.1217-1274), Siger of Brabant (ca. 1240- ca.1284) and Roger Bacon (ca.1214-1292) were other important figures in epistemology of the 13th century. Bonaventure was opposed to rationalism and empiricism. He favoured mysticism and defended the necessity of revelation without any conditions. The mystical orientation was later continued in Germany by Meister Eckhart (ca.1260-1327) and Nicholas of Cusa (1401-1464), who came back to negative theology. Siger was strongly influenced by Averroes. In particular, Siger transformed the Commentator's view about the superiority of philosophy over theology into the theory of double truth. There are theological truths and philosophical truths. Both belong to different epistemological orders and cannot remain in any logical conflict. This theory welcomed an allegoric interpretation of religious truth in order to solve alleged inconsistencies between products of faith and reason. Latin Averroism influenced Dante Aligheri (1265-1321) and Marsilius of Padova (ca. 1275-1342) in their political philosophy. It was also important for later scholasticism, and became a philosophical basis of contemporary fideism, a kind of religious philosophy considering faith as a purely subjective matter. Bacon based his philosophy on his practice as a scientist. He proposed an empiricistic epistemology based on experiment and mathematics.

Thomas Aquinas (a Dominican father) and Jan Fidanza (a Franciscan father) were personal friends. However, Franciscan philosophers later became opponents of Thomism. Duns Scotus (ca. 1266-1308) and William of Ockham (ca. 1285-1349) became the most important representatives of this stream. Scotus, famous for his subtle conceptual distinctions, was mostly interested in metaphysics and theology. His doctrine about *haeceitas* (the individual essence) justified the necessity of knowledge. He also revived the voluntarism of Augustine. Although Scotus was a realist with respect to universals, his theory of *haecceitas* was a step toward nominalism, radically developed by Ockham. This ontological view resulted in a rejection of cognition of natural kinds (species). Cognition was restricted to particulars and consisted in abstraction from properties of singular things. For Ockham, universals were reduced to signs; this view culminated in the "terministic" tendency of logic which began in the 13th century. Ockham was a radical empiricist and came close to phenomenalism and scepticism. He defended the theory of double truth with its separation of theology and philosophy and considered metaphysical problems as more connected with will than intellect.

Ockham's philosophy helped in the rise of a new scientific methodology in the end of scholasticism. Jean Burridan (ca. 1295-1358), a student of Ockham, Thomas Bradwardine (?-1349), Nichole of Autrecourt (ca. 1300-1369), and Nichole Oresme (ca.1325-1382) relied on experience and mathematics. The theory of impetus, an inner disposition of moving bodies, was the most significant result of this new scientific outlook. The theory of impetus broke with Aristotelian mechanics based on teleology and the principle that any movement must be caused by another movement. This change prepared the ground for the Galilean revolution in physics.

There is one particular epistemological idea with which the Schoolmen were particularly preoccupied, namely that of truth. The most famous description of the concept of truth was given by Thomas Aquinas:

(c) "Veritas est adequatio intellectus et rei, secundum quod intellectus dicit esse quod est vel non esse quod non est." *(De Veritate* 1,2).

The first of this formulation defines truth as the agreement (*adequatio*) of thought or mind (*intellectus*) and thing (*rei*), whereas its second part formulation (from the word '*secundum*') basically repeats the content of (a) of Aristotle. Thomas Aquinas attributed the adequatio formula to Isaac Israeli, a Jewish philosopher. However, this reference is erroneous because the term *adequatio,* crucial in (c), does not occur in Israeli. This term was introduced by Wilhelm of Auvergne in his comments on Avicenna.

Then, it was used by Albert the Great and adopted by Aquinas who also used the words *confomitas, convenientia* and *correspondentia.* Anyway, (c) became the standard account of the theory of truth which is labelled "the classical (or correspondence) theory of truth". It is remarkable that no term in Aristotle can be literally translated as *adequatio* although (b) contains and idea of correspondence. Also Abelard commented on Aristotle's definition of truth in various places, and his statements can be summarized by the following formula: "a sentence 'p' is true if and only if it refers to an existing state of affairs". Some authors claim that Abelard anticipated the semantic definition of truth.

It is true that medieval philosophy was governed by principles derived from theology and religion. However, we very often encounter a common error that consists in looking at the philosophy developed in this long period (about 1000 years) as completely uniform. It was quite the opposite: inside a general religious framework, many mutually conflicting views arose. On the other hand, it is also true that this pluralism was suspicious for religious authorities. In fact, the Church used administrative means to block the development of some ideas. Eriugena, Abelard, Bacon, Siger, Ockham and numerous other philosophers were officially condemned by the Church authorities; some of them were also personally repressed. Even Aristotelian philosophy was regarded as somehow heretical before Aquinas succeeded in Christianizing it. This fact allows for a better understanding of Averroism, which was an attempt to achieve a peaceful co-existence of faith and reason. However, this proposal was rejected, and the hostility of the Church toward the rise of modern science became paradigmatic for a long time: the trial of Galileo was a symptom of this situation. In general, the main merit of medieval epistemology lies in giving a variety of epistemic foundations for religion as a human phenomenon. Although the Church chose a particular solution, namely Thomas' view, based on Aristotelian empiricism, of the natural coherence between faith and reason, all other possibilities, for example, the theory of double truth, were proposed. Since scholasticism became a negative pattern of philosophy and its method in the Renaissance, medieval philosophy was almost completely rejected in the beginning of modern times. Perhaps this fate did not give full justice to many very subtle scholastic views, but, on the other hand, modern history began with a clear demand: do philosophy in an opposite manner than the Schoolmen did!

3. MODERN PHILOSOPHY SINCE DESCARTES TO KANT

The Renaissance was very good for art, literature, and science, but less favourable for philosophy. This might have been caused by the fact that the Renaissance men very often identified scholasticism with philosophy. Since scholasticism was rejected, philosophy did not gain a particularly strong interest. Scholasticism was also linked, correctly, of course, with Aristotle's thought. Hence, the Renaissance attitude against the scholastic style of doing philosophy opened the door for other philosophies neglected during the last period of medieval philosophy. Platonism, Stoicism, and Epicurean hedonism became much more popular than Aristotelism philosophy. They were mainly employed as the foundation for a new philosophical anthropology, more centered around human matters and cultivating aesthetic values. Renaissance philosophers were not afraid of eclecticism, and this attitude did not help in discovering new original views. However, some features of Renaissance culture and some particular philosophical events were of certain importance for the future development of philosophy, in particular epistemology. The general climate of the Renaissance liberated science and philosophy from the bounds of theology. Moreover, the scientific revolution triggered by Copernicus and Kepler, and advanced by Galileo, sooner or later, influenced philosophy. Copernicus presented himself as a mathematician. In fact, the achievements of mathematics in science as well as the development of mathematics itself in the 16th century (especially, the origin of 'symbolic' algebra in Viete and Cardano) partly prepared the ground for the Cartesian revolution in philosophy. Two philosophical facts were of special importance for the development of epistemology: French scepticism (Michel Montaigne, 1533-1592, Pierre Charron, 1541-1603) and Francis Bacon's (1561-1626) empiricism in England; the former influenced Descartes, but the latter was continued by the great British empiricists.

As I have already mentioned (more than once, in fact), René Descartes (Renatus Cartesius) (1596-1650) became a revolutionary philosopher. Speaking most generally, he radically changed the priorities of philosophy, because, according to him, philosophy should and could find its starting point in epistemology. Moreover, he conceived of philosophy as completely autonomous, in particular independent of theology. He trusted reason and considered philosophy as *mathesis universalis* on which all other knowledge was based. Guided by these views, Descartes created a philosophical system which was new and original, not comparable with anything else since Plato and Aristotle. It was outlined in many books of which the following are of special importance for epistemology: *Rules for the Direction of Mind* (1618-1629), *Discourse on the Method for Properly Conducting Reason and Searching for Truth in the Sciences* (1637), *Meditations on First Philosophy* (1641). Descartes, being himself a distinguished mathematician (he discovered analytic geometry), demanded that philosophy had to be based on a proper method. He looked for the fundamental starting point, obvious and free of any doubt. He demonstrated the importance of this point by probing so called methodological scepticism. Descartes took seriously the sceptical challenge, although he did not believe that scepticism could be true. Thus, methodological scepticism was a way to overcome scepticism of the ancient kind. The main idea of methodological scepticism consisted in doubting everything that could be subjected to a reasonable, i.e., coherent doubt. Can we doubt that God exists? We can – says Descartes. Can we doubt that the real world exists, provided that God exists? We can – says Descartes. In particular, we cannot exclude that there is the malicious demon who constantly deceives us. However, if we are doubting, we are thinking, and if we are thinking, we are. Thus, Descartes says, if I am thinking, therefore I am (= I exist). *Cogito, ergo sum* – this phrase became one of the most famous philosophical sentences. We must accept *cogito, ergo sum* as correct, because its denial is contradictory. Assume that I say: it is not true that if I am thinking, I exist. By simple rules of sentential logic, we obtain: I am thinking and I do not exist. However, existence of an act of thinking without a subject who is thinking, seems impossible. Thus, we have an indubitable axiom for any further philosophical proceeding. *Cogito, ergo sum* was not new (see the section on Augustine above), but Descartes derived from it new consequences.

Descartes supplemented his refutation of scepticism *via* methodological scepticism by a special insight of the proper method. Thinking about the method, Descartes was very strongly influenced by mathematics. He considered mathematics as a collection of indubitable truths, necessary, logically interconnected truth and independent of experience. Thus, mathematics and its method provides a pattern of knowledge in its most perfect sense. The proper method is governed by several rules the most important of which are: (a) never accept anything true without having self-evident knowledge of its truth; (b) encountering a difficulty, divide its examination into as many factors as possible; (c) always start with the simplest elements and proceed from them to more complex wholes; (d) be sure that your enumerations and catalogues of problems are complete, that is, nothing has been left out; (e) always proceed step by step, in particular, reduce complicated and obscure propositions to simpler items, then come back, checking everything by intuition of the basic simples. Analysis, reduction to the simples and checking by intuition are the most fundamental features of Descartes' method, and their mathematical provenience is obvious. Descartes believed that this method provided an access to *clarae et distinctae ideae,* clear and distinct ideas. And he defined truth as a proposition consisting of clear and distinct ideas. The *cogito* argument and the analytic method are instruments which produce certain knowledge; the quest for certainty was the most essential Cartesian claim. On the other hand, certainty is not automatically achieved by reason. Mind has not only the intellectual faculty, but is also equipped with freedom of will (the faculty of choice); in particular, sensations are passive and cannot constitute ideas by themselves. Hence, judgments are made by cooperation of intellect and will. Since the latter has a larger extent than the former, the result (a judgment) can be erroneous, that is, ideas involved in it are unclear or not distinct. It is Descartes' explanation of how error is possible. Let me note that voluntarism was another, probably unconscious, Augustinian motive in Descartes' philosophy.

Descartes, armed with *cogito, ergo sum* and the method of analysis, began to construct his system. Roughly speaking, he intended to obtain philosophy as metaphysics and theology, general natural philosophy, and various more specialized fields, like medicine or mechanics. He conceived of his system as hierarchically organized, proceeding from more abstract things to less abstract ones. Comparing his methodological scepticism with his idea of a complete system of knowledge, we see that Descartes liked to rebuild everything that he rejected on his way on *cogito, ergo sum.* He wanted to prove that God exists, that we exist, and that we are living surrounded by existing things. He understood perfectly that he could not build his system from *cogito, ergo sum* only. Since we have, *via cogitationes.* an access to the content of our consciousness, we must try to find clear ideas which populate our mind. Descartes found three ideas of this kind: God, soul and body. They are *ideae innatae,* innate ideas, and thereby are clear and distinct. God, soul and mind are substances with special attributes: God is infinite, soul (mind) thinks *(res cogitans),* and body is extended *(res extensa).* Now, Descartes' main task was to prove that particular substances exist. Roughly speaking (the matter is more ontological than epistemological), he proved the existence of God *via* pointing out that He must exist for His perfection; it is a version of the so called ontological proof (see the fragment about Anselm above). The existence of *res cogitantes* and *res extensae* were established on the basis of God's existence; details must be neglected here. A special problem was connected with the mutual relation of soul and body. For Descartes, they were generally independent (in this respect, he accepted dualism) with the exception of human beings where *res cogitans* and *res extensa* are interconnected. Descartes formulated the so called psychophysical problem (the mind-body problem), one of the most frequently debated philosophical questions. Although it is mainly an ontological problem, it often influenced epistemology so strongly (for example, the foundations of cognitive science) that it should be mentioned on this occasion.

Descartes' epistemology was based on two fundamental views: radical apriorism and nativism. Both were not new, we encountered them in Plato (see above). However, Descartes justified his principles entirely by epistemological analysis, and without any direct appeal to ontology or metaphysics. In fact, his metaphysical theses were secondary to epistemology. Thus, he began a tradition in which ontological views are consequences of epistemological analysis; it does not mean that every post-Cartesian philosopher executed this pattern, but many did. Another feature of Descartes' epistemology consists in psychologism. *Cogito, ergo sum* and the conception of innate ideas are based on psychological analysis; for example, the rejection of genetic empiricism appeals to the passivity of sensations. Another symptom of Descartes' psychologism is his use of the word 'idea'. Contrary to the older tradition, although not without justification in the linguistic usage of the Latin word, ideas were for him concepts in the psychological sense. Descartes' trust in reason was derived from the indubitable accessibility of the content of mind. Thus, he was a foundationalist: he believed in the ultimate foundation of knowledge consisting in accessibility of clear and distinct ideas. His philosophy also had important methodological consequences. First, he, like nobody before him, defended the perfection of mathematical method. Secondly, his claim that extension was the only real attribute of corporeal bodies justified the geometrization of physics, a program to which Descartes contributed himself. However, his physics lost to the

Galileo-Newton project based on another idea of quantitative properties, not only geometrical.

Descartes' philosophy raises several problems and doubts. Is Cartesian dualism a proper solution, and, in particular, is his interactionism consistent with dualism? Or more specifically: how can such different substances mutually interact? These questions concern the mind-body problem. The next doubt question concerns the non-circularity of his ontological arguments derived from epistemology. For example, Descartes proved the existence of God starting from the thesis that we possess the idea of God which had to be created by Him. Still another question results from the analysis of *cogito, ergo sum.* Is it an inference or a statement? The occurrence of the word *ergo* suggests the first view. Now, if *cogito, ergo sum* is an inference, it is an enthymeme. What about the lacking premise which seems to be needed in order to conclude 'I am'? The full reasoning seems this: Since (a) if I am thinking, I am, and (b) I am thinking, therefore (c) I am. The danger of circularity is clear. These and other problems have been discussed since Descartes' works appeared. And they are still being discussed. However, Descartes' influence was enormous. He decided the future course of European philosophy. In particular, he began the great tradition of modern European rationalism occurring in almost every domain of philosophy and science. For example, the Port-Royal School in grammar used Cartesian views in developing the so called rational grammar based on the assumption that fundamental grammatical categories were innate (this idea was recently revived by Noam Chomsky who called his linguistics 'Cartesian'). 'Axiomatic' systems of natural law, initiated in the Netherlands by Hugo Grotius, are another part of Cartesian heritage. The French style of doing science, which consists in looking for ultimate simple conceptual elements, is still another example. And, of course, the philosophical ideas of Pascal, Spinoza and Leibniz would be difficult to understand without an appeal to Descartes' heritage. The same concerns his opponents. Thus, René Descartes deserves the name of 'the father of modern philosophy'.

Descartes sent his *Meditations on First Philosophy* to several philosophers. They formulated objections to which Descartes prepared extensive answers. Thomas Hobbes (1588-1679) was among the critics. In his objections, Hobbes agreed with Descartes about the paradigmatic character of mathematics, in particular geometry. Hence, he also shared the Cartesian view that extension as a geometrical property was an attribute of bodies. However, Hobbes questioned Descartes' analysis of *cogito, ergo sum.* because it did not prove the independent character of mind. For Hobbes, the statement 'I am thinking' did not exclude that the thinking subject was corporeal. Hobbes was interested not only in epistemology and ontology, but also, and even more, in political philosophy. Continuing the tradition going back to Roger Bacon and William of Ockham, and revived by Francis Bacon, Hobbes based his philosophy on empiricism. In a sense, he laid the ground for the golden period of British empiricism in the 17th and 18th centuries. This formation began with John Locke (1632-1704).

Locke directed his epistemology, elaborated in an extensive treatise *An Essay Concerning Human Understanding* (1689/1690), explicitly against Descartes, particularly against the nativism of the latter. Instead, he developed genetic empiricism, a theory which claims that the mind is a *tabula rasa* (a pure blackboard), unless experience writes some signs on it. However, there is something common in Locke and Descartes, namely the trust in the accessibility of mental contents. Even more, Locke, like Descartes, believed in the indubitability of the results of the direct knowledge of mental contents. Thus, Locke was also a foundationalist, but of a different kind, connected with empiricism. As a matter of fact, foundationalism was a common view shared by rationalists as well as empiricists from Descartes to Kant. For Locke, experience was the only source of knowledge, and there were two kinds of it: sensation and reflection. The former was 'outer', but the latter 'inner' and provided access to mental contents. Let us start with reflection and its role in generating knowledge. According to Locke, reflection is a conscious awareness of our mental activities and their results. Due to reflection, the mind is able to acquire the direct intuitive knowledge that consists in apprehending ideas without any mediation of other ideas; Locke, like Descartes, uses the word 'idea', in the psychological sense. Intuitive knowledge is certain. The mind does not need to prove or check it, because this kind of knowledge is accommodated, similarly as light by eyes, only by a directed activity. For example, we know that something which is red is also not blue by experiencing colours in this way. Intuition also operates in the domain of memory; in fact, Locke was the first philosopher who seriously analyzed memory as a cognitive faculty. It is important that everybody is subject to cognition of this kind. In particular, we intuitively grasp whether ideas mutually agree or not. Thereby, intuitive knowledge is common and indubitable; such is the Lockean argument against scepticism. The intuition described is decisive for the certainty and evidence of our knowledge, in the domain of logic and mathematics as well. Both represent demonstrative knowledge as knowledge based on mediating ideas. For example, when we prove the sum of the angles of a triangle is equal to two straight angles, we must also use other ideas, because we cannot directly compare the ideas involved in our demonstration. The certainty of demonstrative knowledge recurs to the certainty of knowledge generated by intuition, although there is an important difference between both kinds of cognition. Intuition does not require any activity, it is present or not. On the other hand, demonstration is always active. It happens that demonstration is a remedy for doubts this situation never occurs in the case of intuition. Nevertheless, grasping the agreement or disagreement of ideas also constitutes the foundation of demonstrative knowledge. Intuitive and demonstrative cognition forms knowledge which is certain. Intuition plays one more important role. It provides simple ideas. All others are complex and achieved by association or abstraction. In order to complete the remarks on reflection, it is interesting to note its close conceptual relation to introspection, which became the main method of psychology in the first period of the development of this field.

Sensation provides knowledge about particular individual objects. Two problems arise here. First, it is indubitable by intuition that we have ideas in our minds. However, it is not certain whether ideas have real counterparts. Locke defended the view that they had, appealing to the causal theory of perception. According to Locke, our senses are not able to produce sensations without external causes: since sensations are momentary, and senses themselves are almost always potentially subjected to sensations, the explanation of the latter has to recur to external objects and their effects on our senses. Further, Locke argued that there is an interesting difference between ideas given in memory and those given by sensations. We can influence our acts of remembering, but sensations occur involuntarily. It means that sensations have outer causes. We often feel pleasure or annoyance connected with sensations, but these feelings do not arise in related rememberings, although ideas occurring in sensations and memory are common. It also indicates that we must appeal to external causes in order to explain the genesis of sensations. Finally, the credibility of the existence of outer reality is strengthened by the evidence of various senses. Thus, our firm belief that real counterparts of our ideas actually exist, although it does not reach the level of absolute certainty, provides the practical certainty which means more than mere probability. Thus, we have in Locke three levels of knowledge, according to its certainty. Intuitive knowledge is on the highest level and has the maximum degree of credence. Then, we have demonstrative knowledge which is certain by demonstration. Finally, outer sensations provide practically certain knowledge. Moreover, we have several modes of beliefs assessed by probability, for example, conjecturing, guessing, doubting, etc. To some extent, Locke reintroduces the ancient distinction of *episteme* and *doxa,* although the scope of the former was for him broader than for Plato and his followers.

The second problem concerns the question of how far the representation of objects by ideas fits real properties of represented things. Here, we must point out the ambiguity of the word 'idea' in Locke, similar to that in the writings of Descartes, but much more dangerous in the case of Locke. As I have already remarked, Locke used the word 'idea' in its psychological sense as referring to presentations or concepts as products of mental acts. On the other hand, ideas represent objects as they are grasped under this or that aspect. Hence, if we say that ideas are objects of reflection, it means that we contact things as-represented-so-and-so by ideas. This immediately raises the question of whether all constituents of representations actually represent real properties of outer things. Locke's celebrated answer was negative and consisted in his famous distinction between primary and secondary qualities. Roughly speaking, only primary qualities correspond to objective properties, but secondary qualities are subjective products of senses. The list of primary qualities is rather narrow and covers, in its minimal version, shape, size and mobility; in other places Locke also adds number, solidity, and texture. The rest, in particular, colour, taste, sound, smelling, etc. belongs to secondary qualities. It is important that Locke's distinction was strongly rooted in the physics of the 17th century. In fact, Locke was the first philosopher who heavily used the physical ideas of Galileo and Newton in philosophy. Physically speaking, primary qualities correspond to properties which are quantitatively expressed in physical equations. Consequently, secondary qualities are purely qualitative and have no mathematical meaning, unless we relate them to primary (now, we can say expressible by mechanics) attributes. By the way, we find here another sharp contrast between Locke and Descartes; for the latter, only geometrical attributes are fully objective. Locke was inspired by physics, but his argumentation was not limited to a repetition of arguments derived from mechanics. His basic aim consisted in the epistemological legitimization of the primary/secondary qualities distinction. He achieved this task by pointing out that primary qualities were experienced by all senses. Besides, he maintained that touch is the distinguished sense and it informs us about objective properties better than the other senses. The view based on the distinction between primary and secondary qualities is called critical realism. It is usually contrasted with naive realism, that is, the view that the world is just such as it appears. Using Lockean language, the difference between critical and naive realism is this: for naive realism, all qualities are primary, and all represent objective features of the world, but for critical realism, some qualities are primary and other secondary, and only the former correspond with objective properties. Critical realism was proposed by Democritus in antiquity. However, Democritus was led by very vague intuitions, derived mainly from metaphysics. It was Locke who argued for critical realism by epistemological arguments. Locke was not original in his genetic empiricism either, because Aristotle and Thomas Aquinas preceded this view. In fact, the formula *nihil est in intellectu quod non prius fuerit in sensu* was shared by the Stagirite and Aquinas as well as by Locke. The latest differed from his honourable predecessors concerning the scope of the sources of knowledge. Locke's empiricism, contrary to that of Aristotle or Thomas, was not sensualistic, because it admitted reflection as a device for the production of ideas. However, the most important difference between Locke and the older empiricists concerned method. In contrast with Descartes, Locke followed the father of modern philosophy in general methodological flavour. Locke's method was psychological, because mental contents were the starting point of his further investigations. In particular, he claimed that the relations between ideas and objects could be conceptually characterized only after taking into account object-as-represented-so-and-so by ideas. Locke, like Descartes, proceeded by the reconstruction of the genesis of our ideas, and analytically by extracting simple and evident ingredients of cognition. And, last but not least, Locke, similarly as Descartes, executed metaphysics by deriving metaphysical conclusions from epistemologically justified assumptions. Perhaps Locke's deep ideological connections with the Cartesian revolution caused that all problems and difficulties of modern empiricism were related rather to him, not to his empiricist predecessors. Almost everything that is discussed in the contemporary theory of perception, directly or indirectly goes back to Locke's philosophy of sensations. Also his broadening of the concept of knowledge essentially determined later discussions concerning this concept.

The difficulties of Lockean empiricism are numerous and serious. They are conceptual, for example, the ambiguity of the word 'idea'. They are also substantial, for example, the difference between hallucinations and sensations with real counterparts: the causal analysis of sensation does not explain hallucinations. George Berkeley (1685-1753) was the first to raise objections against Locke. It was a quarrel inside the same family because Berkeley also belonged to the empiricist camp. Berkeley (his main work *Principles of Human Knowledge* appeared in 1710) radicalized Locke's view in two respects. Berkeley proposed an empiricism based on sensualism and nominalism. Sensualism meant the rejection of reflection. Thereby, all ideas were conceived as directly or indirectly derived from sensory experience. Nominalism denied the existence of general ideas obtained by abstraction; thus, for example, the general triangle as an idea, admitted by Locke as neither right, nor acute, nor obtuse disappeared under Berkeley's assumptions. Mathematics was about sensible objects, for instance, shapes drawn on a blackboard. These radicalizations entailed further solutions. Locke himself had some doubts concerning the concept of substance: he was not certain whether existence of the substance, understood as the substratum of all things, was consistent with empiricism or not. The problem of the objectivity of some properties was solved by Locke with the help of the distinction of primary and secondary qualities. However, he was inclined to regard the substance in the above sense as something unknowable, but still existing. Berkeley argued that this view was a piece of redundant metaphysics; there was no place for substance independent of sensations in his picture of the world. For Berkeley, Locke's arguments for the existence of external things were insufficient. Neither the distinction between reflection and sensation nor an appeal to agreement of evidence provided by different senses proved that external things correspond to ideas; note that Berkeley, in comparison with Locke, restricted the scope of the term 'idea' to particular sensations and their complexes. There is no property of our ideas which could serve as a base for concluding that our sensations correspond to actual things.

Berkeley found the solution in subjectivism: he considered all qualities as secondary in the Lockean sense, but without any appeal to external things. We experience only mental contents, and they are the only reality accessible to sensations. Immaterialism was the next consequence of Berkeley's views about perception: there are not material things. Everything in Berkeley was summarized by his famous dictum: *esse* = *percipi* ('to exist' means no more than 'to be perceived); the full content of this slogan is obvious if one remembers that, for Berkeley, the scope of experience consists of particular sensations. Berkeley understood, of course, that his views were at odds with ordinary beliefs on which other people and numerous things existed. He did not want to advance a view which would be so inconsistent with the common sense. He tried to overcome his subjectivism by an appeal to God who sees everything constantly and introduces stability into the flexible world of sensations. Additionally, he pointed out that this was the cause of why many people had the same sensations. Since Berkeley's explanation of the objectivity of sensations belongs rather to metaphysics than to epistemology, it can be left here without further comments. However, Berkeley's *esse* = *percipi* deserves more attention. Traditionally, *esse* is an ontological or metaphysical concept, and *percipi* belongs to the vocabulary of epistemology. Thus, the basic Berkeleyan equality is simply a reduction of ontology to epistemology, perhaps the most complete in the whole history of philosophy. Berkeley also understood that his epistemological theory went against suggestions of mathematical natural science. He criticized fundamental concepts of mathematics, in particular the concept of infinitesimals (infinitely small quantities). Some commentators say that Berkeley correctly recognized logical unclarities in this concept, but it is only partly truth; in fact, he was guided by his nominalism which forced the view that every line must have a limited number of points. He also rejected Newtonian mechanics for absoluteness of space and time, the concept of force and admission of causal relations between phenomena. For Berkeley, there was nothing in our sensory experience that could justify these categories. Berkeley's philosophy, similarly as radical scepticism, is usually considered as oddity. His extreme sensualism and subjectivism (subjective idealism) did not find many defenders. But, on the other hand, Berkeley provided a challenge for everybody who wanted to derive the objectivity of knowledge from the analysis of sensory perception. He did this by showing difficulties in Locke's epistemology form the point of view its coherence with science and ordinary beliefs.

The controversy between Berkeley and Locke showed that the relation between mental contents and external entities was the main issue. In Locke's opinion this relation obtains, i.e., there exist, at least in some cases, its second terms, namely objects. Berkeley, on the other hand, radically denied this position. The same problem was considered by David Hume (1711-1776), the third of the great British ('British' because Hume was Scottish, and Berkeley Irish; only Locke was English). His epistemology was extensively elaborated in *Treatise on Human Nature* (1739-1740) and once more explained with some refinements in *Enquiries Concerning Human Understanding and Concerning the Principles of Morals* (1748). Hume's general solution was equally anti-Lockean and anti-Berkeleyan: he regarded the problem as pointless. According to Hume, the only problem for epistemology consists in investigations concerning the empirical correctness of our mental contents. The rest is and must be silence. Hume did not deny that belief in the existence of the external world is of considerable practical importance for human beings. What he denied was theoretical possibility of solving the Locke--Berkeley controversy.

The method applied by Hume followed that used by Locke and Berkeley, and consisted in a genetic analysis of our mental contents. Hume distinguished impressions and ideas. Roughly speaking, the former correspond with sensations in Locke's sense. Seeing, hearing, etc. consists in having impressions; hence, they are direct sensory experiences; in fact, Hume spoke also about impressions generated by passions, but this topic, as related to his moral philosophy, can be omitted here. On the other hand, ideas arise when we remember, think, imagine, etc. Clearly, ideas represent indirect sensory cognition. It should be noted that Hume restricted the extension of the term 'idea' still more than Berkeley did. Hume rejected not only abstract, but also complex ideas, that is, ideas consisting of many impressions; if an idea is general, it is so in given circumstances and for a given subject, never automatically by a faculty called abstraction. Consider thinking. Presumably, it consists in connecting or disconnecting impressions. This process generates ideas. Thus, ideas are rather products of thinking than complexes of impressions. However, there is a close genetic link between impressions and ideas, because the former are pictures of the latter. The development of cognition can be described in the following way. Everything starts with impressions. Mental operations give rise to ideas. Now, ideas concur in some constant associations. Hume distinguished three principles of such associations: (a) by similarity; (b) by proximity in space and time; (c) by causal inference.

Now, the crucial question for Hume is: which ideas and claims based on them are empirically legitimate? Hume divided the objects of cognition into two groups: relations of ideas and matters of facts. The relations of ideas are not subjected to factual claims, but they constitute the domain of mathematics. Mathematics is certain, but completely devoid of any factual content. Hence, the problem of empirical legitimation of our statements is restricted only to claims concerning matters of facts. If such claims were reducible to impressions, the questions would be very simple, because cognitions based on impressions are certainties. Unfortunately, factual statements exceed the impressional base, because they go beyond currently given experiences. Hume examined two ideas which were usually taken as the base for factual claims, namely the idea of causality and the idea of substance. Contrary to Berkeley, Hume took seriously Newtonian physics which seemed to appeal to causality and substance. Hence, he regarded his analysis as important for the foundations of science. Hume did not deny that we observed constant successions of events, more strictly ordered sequences of impressions or ideas. However, he pointed out that we have no logical reason to conclude that succesive connections are necessary; according to Hume's impressive dictum: *post hoc* does not mean *propter hoc.* Thus, the traditional account of causality as a necessary and universal relation fails, because there are no logical or empirical grounds for claiming that the causal nexus has such properties. Logical derivation of necessity and universality of causality is impossible, because logical demonstrations are restricted to the domain of relations between ideas. Moreover, the empirical justification is here obviously insufficient, because the causal connection between phenomena usually exceeds its experiential base. In particular, we cannot logically exclude that future data will provide evidence forcing the rejection of a given connection, earlier regarded as universal. Neither can we appeal to the principle that every event has its cause, because this argument is simply circular: it assumes the thesis which is to be proved. What remains is reducible to already observed successions grasped by the mind. In particular, instincts are responsible for grouping ideas into more or less stable sequences of ordered items. In fact, we can only expect that regularities will also appear in the future. The criticism of the concept of substance was similar. This concept arises as an effect of linking of coexisting ideas into stable wholes. Also this operation is legimitated neither by logic nor by empirical evidence, but finds its explanation in instincts. Hume's attack on substance was stronger than the Berkeleyan criticism, because it dispensed with all kinds of substance, while Berkeley had at least acknowledged a spiritual reality. The belief in causality and substance had its justification, or rather explanation, in practical reason, but no theoretical import. Thus, we should say goodbye to the concepts of causality and substance: they are metaphysical redundancies in science and in daily life. Clearly, Hume's criticism solves the problem of the relation between mental contents and their alleged objects. The solution points out that the problem itself is meaningless: it admits neither a positive nor a negative way out. This view is faithful to empiricism and introduces elements of scepticism.

Hume refined the former empiricism in many respects. He rejected Bacon's belief that experience could justify necessary connections between facts, Locke's critical realism in favour of pure phenomenalism, and Berkeley's *esse* = *percipi.* Briefly, he purified empiricism of metaphysical (in his understanding) elements like causality and substance that have no empirical justification. Hume introduced into epistemology two important novelties, namely, the view that logic and mathematics are devoid of factual knowledge, and the new approach to knowledge consisting in a departure from Plato's conception of *episteme.* Since factual knowledge is never certain, we must admit that it is only probable. This feature was not considered by Hume as a pejorative mark, because it exhibited the fundamental feature of knowledge produced by experience. Both Humean novelties opened new epistemological perspectives. In particular, he discovered a new kind of empiricism, namely, moderate methodological empiricism, connected with full genetic empiricism: all knowledge is genetically empirical, but its part is devoid of factual content. Thus, Hume tried to incorporate the traditional virtues of mathematics, namely universality, certainty and necessity into empiricism. This line was followed by logical empiricism in the 20th century which also shared Hume's rejection of metaphysics. Hume's criticism set down challenges for friends of causality and substance, and they had to deal with his arguments. There are still other points in Hume's philosophy which became important for later epistemology. He was a radical naturalist and thereby a pioneer of naturalized epistemology. His thesis that is-sentences do not imply ought-sentences, and his theory of moral sense are central for the epistemology of ethics. On the other hand, Hume's theses give rise to several doubts and questions. Is his picture of the development of mental contents correct? Do we really have no arguments for the reality of causal connections and substances? What does it mean that factual knowledge is probable? Is the probability objective or subjective? Thus, challenges stemming from Hume's criticism touch not only friends of the view criticized, but also his own solutions. In general, all problems of causality, induction, physical (natural, real, etc.) necessity, laws of nature and logical probability are, so to speak, surrounded by Hume's views. Perhaps this variety of questions stimulated by Hume is the best measure of his enormous significance for the further development of epistemology.

We left rationalism with Descartes. This tradition was developed by his continental followers, partly as a response to British empiricists. However, there was one notable exception among the post-Cartesians, namely, Blaise Pascal (1623-1662). He was a brilliant mathematician and physicist, but these circumstances did not influence his philosophy. Pascal turned his attention to practical matters and argued that knowledge which satisfied high Cartesian standards was not able to deal with the real problems of human life. Thus, reason is useless for human worries. These issues require insights taken from the 'heart' and faith. Pascal stands out because the irrationalism expounded in the age of reason. He was a predecessor of existentialism in many respects, but he did not find followers until in the 19th century in Sören Kierkegaard.

The epistemological and metaphysical motifs of Descartes' philosophy were adopted by Benedict Spinoza (1632-1677), a Dutch philosopher of Jewish origin. Spinoza, like Descartes, believed in the power of reason. In particular, he maintained that the world was intelligible and thereby accessible to rational knowledge. This belief is evident from the title of his main work *Ethics Demonstrated in a Geometrical Manner* (1677) which directly indicates the application of a rational mathematical method to philosophical problems. The word 'ethics' is misleading, because Spinoza's *opus magnum* is basically an ontological-epistemological treatise. Both aspects are closely related. In particular, Spinoza argued that his improvement of Cartesian dualism led to a specific account of knowledge. Spinoza replaced the dualism of *res cogitans* and *res extensa* by a kind of monism based on an identification of God and nature. Thinking and extension became *modi,* rather parallel than interconnected, of unified substance in Spinoza's ontological model. Spinoza, like Plato, distinguished knowledge and opinion. The latter is provided by senses and concerns particular *modi* taken in separation. Moreover, we need knowledge referring to the relations between both modes, and to substance as such. The domain of reason produces rational knowledge. Thus, for Spinoza, knowledge and rational knowledge are the same. Nature (substance) appears as *res cogitans* or *res extensa,* but ontologically it is united. This unity is *causa sui* and its existence is intuitively certain. Nothing more can be said about substance. In particular, substance is not determined; it has no particular properties. According to Spinoza, *omnis determinatio est negatio,* that is, every predication about substance turns it into its opposite. This is a mysterious aspect of Spinoza's philosophy, especially of his rationalistic epistemology, because it seems that the indeterminacy of substance blocks its rational description. Hence, intuition in Spinoza's sense is quite different from Cartesian intuition, because the latter leads to clear and distinct ideas, while the former is more similar to mystical contemplation than to knowledge modeled by mathematics. Perhaps this is the main reason that Spinoza's epistemology is considered to be much less interesting than his metaphysics. Moreover, his appeal to mathematics in *Ethics* is rather declarative and verbal than substantial.

Gottfried Wilhelm Leibniz (1646-1716) was the greatest successor of Descartes. He did for rationalism something similar to what Hume did for empiricism: both offered much more perfect versions of related epistemologies than their predecessors. In stressing that Leibniz was a rationalist, one should realize that his interests, like in the case Spinoza, were more metaphysical than epistemological; yet, both philosophers tried to develop epistemologies closely connected with their metaphysical views. One of Leibniz's major works has an interesting title, namely *New Essays Concerning Human Understanding* (1704, published in 1765). This immediately recalls Locke's treatise *An Essay Concerning Human Understanding.* This similarity is not incidental. In fact, Leibniz's work is a systematic polemic against Lockean empiricism; it is interesting that Leibniz, having received information about Locke's death, decided not to publish his book. Leibniz was inclined to accept Locke's account of knowledge as a description of the mechanism of cognition. For example, Leibniz agreed with Locke that we encountered concrete ideas earlier than abstract ones. However, according to Leibniz, Locke did not explain the nature of knowledge and overlooked this issue when he pointed out that the external world causally influenced our mental contents. In general, empiricism cannot solve the problem of the nature of knowledge, because a careful analysis of the soul proves that it cannot be dependent on something else. Leibniz modified the empiricist maxim already mentioned *nihil est in intellectu, quod not prius fuerit in sensu* by adding *nisi intellectus ipsae* (except the reason itself). Hence, the source of knowledge must reside in the soul itself, namely in reason and its faculties. Leibniz argued also for the autonomy and self-sufficiency of the soul in a metaphysical way, using his theory of monads, but this aspect of his system can be omitted here. We have a direct access to our soul (mental reality) and this is the only thing that is directly accessible to us. The soul possesses ideas and sensations; although the origin of sensations is causal, the soul has the material for all sensations in advance. Ideas were understood by Leibniz as immediate direct objects expressing the essence and properties of things. Leibniz was a nativist, but, contrary to Descartes, he did not consider all contents existing in the soul to be clear or even conscious. In particular, we can have empty ideas that do not refer to anything, for example the fastest motion (Leibniz did not know that the fastest notion is the speed of light). Sensations are usually unclear and confused. According to Leibniz, we always have the so-called small perceptions which are mostly unconscious, but constantly influence our soul. It is the soul, as an active force, which forms clear and distinct ideas from dark elements. The soul acts in a mathematical way. Hence, mathematics is a pattern of rationality. Leibniz was a great mathematician and there is nothing strange in the fact that his rationalism was mathematically oriented. Leibniz, who had a great understanding of logical matters, suggested a special language *(calculus ratiocinator, lingua universalis)* suitable for expressing and solving every problem. The solutions would be purely combinatorial, strict rules being assumed in advance. Lebniz hoped that future philosophers, instead of conducting interminable discussions, would calculate and thereby reach agreement.

Leibniz distinguished two kinds of truth: truths of facts and truths of reason. They correspond to two general principles of our thinking. The first principle says that everything has sufficient reason for its existence. This is the principle of sufficient reason and governs the domain of the truths of facts. Every truth of reason is reducible by a finite combinatorial procedure (resolution) to the principle of contradiction, which is the second general principle of thought. The truths of reason can be known with absolute certainty of mathematics is once more an example. Every truth is *a priori,* because its predicate is contained in its subject. Basically, every truth is necessary, true in all possible worlds (this important idea was actually introduced by Leibniz) and can be established *qua* truth by analysis of its subject and predicate. However, this is possible only for God who knows everything in advance. Human beings must regard truths of facts as contingent, i.e., not true in all possible worlds, and *a posteriori.* but this is related only to our epistemic capacities, not to the nature of things. Our cognitive situation requires some instruments of cognition in order to assess the credibility of empirical judgements. Leibniz was the first philosopher who clearly observed that truth, necessity, apriority and the objective certainty are coextensive properties of knowledge on the radical rationalistic account. On the other hand, he saw equally clearly that these properties were, except for mathematics, not accessible to human beings. In particular, doing empirical science requires modest epistemic qualifications. Thus, Leibniz outlined perhaps the most radical rationalism, but, on the other hand, his epistemology gives justice to empirical scientific practice. At least one point in Leibniz could bother other rationalists, namely that the virtues and faculties of reason became essentially dependent on various very strong metaphysical hypotheses, particularly on the existence of God. Leibniz's rationalism became influential due to its popularization by Christian Wolff (1669-1764) who set down the foundations of German general philosophical education known as *deutsche Schulphilosophie* (German schoolphilosophy).

The philosophers of French Enlightenment, called *les philosophes,* were mainly interested in social and political matters. Most of them accepted empiricism and criticized religion. These attitudes lead to rationalism, but in another sense than Cartesianism. The new rationalism (perhaps the word 'anti-irrationalism' is a better label) also trusted the natural faculties of reason, but, contrary to Descartes and his followers, rational knowledge was not understood as departing from experience. It does not mean that *les philosophes* rejected Cartesianism at all. For example, they liked methodological scepticism, but as a way of excluding irrationalism. In general, the rational attitude consists, according to *les philosophes,* in the reasonable and sound use of experiential data. The famous *Encyclopedia, that is, Rational*

Dictionary of Sciences, Arts and Crafts, usually called *The Great French Encyclopedia,* 1751-1753, is perhaps the most complete account of the spirit of anti-irrationalism. Three epistemological views of particular French thinkers of the Enlightenment are worth mentioning. Etienne Bonnot de Condillac (1715-1780) developed a version of sensualism. He intended to improve Lockean empiricism by combining it with rigorous Cartesian method. Condillac believed that empiricism, Cartesian method and Newtonian physics could save philosophy. His sensualism was presented in his book *Treatise on Sensations* (1754) in which a well-known metaphor of the statue was elaborated. The statute served as a model of empirical knowledge. At the beginning, the statute had no active senses. Then, its senses were activated step by step beginning with smell. Condillac tried to show how the knowledge of the statute grew, depending on the new capacities acquired by new senses. Jean le Rond D'Alembert (1717-1783), a mathematician and philosopher, was (together with Denis Diderot) one of the main editors of *Encyclopedia,* He wrote an introductory essay to it, in which he expressed his main philosophical views. According to D'Alembert, scientific knowledge is the only one that deserves to be regarded as *the* knowledge. It must be certain, but it can be certain if it refers only to facts. D'Alembert did not reject psychic facts, but he denied that science could be based on them. Thus, real science investigates external facts. D'Alembert's position was not only empiricist; it was also positivistic. As a positivist, D'Alembert dismissed all metaphysical problems concerning the very essence of the world. In that, he was similar to Hume. On the other hand, D'Alembert's positivism was different than that of Hume, because D'Alembert believed in the certainty of factual knowledge, while Hume was a probabilist. Thus, two kinds of positivist philosophy originated in the 18th century. In general, French philosophers of the Enlightenment trusted science very much: scientism prevailed, and D'Alembert was its main apostle. However, there was a notable exception: Jean-Jacques Rousseau (1712-1778). Although he shared the main social and political ideas of *les philosophes,* he did not agree with their theoretical philosophy. Rousseau accused civilisation of degenerating and corrupting people. He was against science as a pattern of knowledge and favoured intuition. Hence, contrary to the main stream of the Enlightenment, Rousseau was an irrationalist. This feature of his philosophy made him a predecessor of Romanticism.

Immanuel Kant (1720-1804) claimed that he carried the Copernican revolution in philosophy. Kant was disappointed by rationalism as well as by empiricism. He intended to provide a synthesis of both main currents of epistemology. More specifically, Kant wanted to achieve a compromise between Hume's empiricism (Kant once said, Hume awoke him from dogmatic slumbers) and Leibniz's rationalism. He was educated in the tradition of Wolff and *deutsche Schulphilosophie.* Kant developed his theoretical philosophy in his *opus magnum Critique of Pure Reason* (1781, sec. ed. with important revisions 1787). It was the first of three critiques (others are *Critique of Practical Reason,* 1788 and *Critique of Judgment.* 1790), and, hence, Kant's mature philosophy is called 'critical' in contradistinction to his early philosophy, termed 'precritical'. Kant's project was to answer the fundamental philosophical problems concerning knowledge, existence, and values by a critique of reason. Thus, perhaps more than anyone, Kant realized the Cartesian project of building philosophy on the basis of epistemology.

Kant introduced two divisions of propositions. First, he distinguished analytic and synthetic propositions, according to their logical form. The structure '*S* is *P*' represents propositions in general. Now it can happen that the content of the predicate-concept *P* is contained in the subject-concept *S*. Any such proposition is analytic. On the other hand, if the content of *P* exceeds the content of *S*, we are dealing with a synthetic proposition. This logical distinction was supplemented by an epistemological division: a proposition is *a priori* if its truth is independent of experience, and it is *a posteriori* if its truth cannot be established without an appeal to experience. As a result we have four mutually exclusive categories of propositions: analytic *a priori,* analytic *a posteriori,* synthetic *a priori,* and synthetic *a posteriori.* Now the distinction inside analytic propositions into *a priori* and *a posteriori* is redundant, because, according to Kant, analytic propositions are *a priori.* Thus the primary purely combinatorial division into four categories is reduced to three kinds of propositions: analytic, synthetic *a priori* and synthetic *a posteriori.* Propositions which are *a priori* can be either analytic or synthetic, propositions which are synthetic can be either *a priori* or *a posteriori.* There is one obvious weakness in Kant's definition of analytic propositions. It does not apply to negative propositions. Kant was aware of this situation. Although he remarked that his criterion of analyticity could be extended to negative propositions, he never explained fully how to do this. Probably he maintained that his proposal that, all analytic propositions are reducible to the principle of contradiction, solves the problem of negative analytic propositions.

Historically speaking, Kant's distinctions among propositions help us understand some important positions in epistemology. For example, Plato's *episteme* consists of *a priori* propositions. With some further assumptions, one can identify discursive *episteme* in Plato's sense with the corpus of true analytic propositions, although intuitive *episteme* covers true synthethic *a priori* propositions. Hume's position admits only propositions that are either analytic (concerning relations of ideas) or synthetic *a posteriori* (matters of facts). Leibnizian truths of reason are of course *a priori,* and his truths concerning facts are synthetic *a posteriori.* However, this is so only from the human epistemic perspective, since for God all truths are analytic, and therefore *a priori.* In general, aposteriorism (methodological empiricism) and apriorism (methodological aposteriorism) can occur in radical or moderate versions. Under Kant's distinctions, radical apriorism (Plato, Leibniz) admits only *a priori* truths, i.e., either analytic or synthetic *a priori* as legitimate pieces of knowledge; radical aposteriorism (Locke, Berkeley) only *a posteriori,* that is, synthetic *a posteriori,* moderate apriorism (a possibility discovered by Kant himself) admits all kinds of truth, that is, analytic, synthetic *a priori,* and synthetic *a posteriori,* while moderate aposteriorism (Hume) admits only analytic and synthetic *a posteriori.* Now, the comparison of various possibilities immediately points out what Kant meant as the compromise between empiricism and rationalism (the Copernican revolution): it is moderate apriorism. If we look at Kant's philosophy from this perspective we easily understand his objections to Hume and Leibniz. Kant did not agree with Hume's rejection of propositions which are synthetic *a priori,* but Leibniz, according to Kant, went too far in apriorism, because he considered all truths *a priori* as analytic. It is clear why the question 'How are synthetic *a priori* propositions possible?' became central for Kant. Since only synthetic propositions extend our knowledge, Kant's main problem can also be formulated in the following way: how is it possible to extend our knowledge by steps which are *a priori.* that is, without any appeal to experience? And still another formulation: how is it possible to have universal and necessary truths (universality and necessity were considered by Kant as attributes of *a priori* propositions) referring to facts?

For Kant, there was no problem with analytic propositions and synthetic *a posteriori* propositions. Logic is analytic, singular empirical propositions are examples of the synthetic *a posteriori.* The real problem concerns the synthetic *a priori.* Kant was convinced that synthetic *a priori* propositions occurred in our knowledge. He intended to explain their possibility, not existence. Kant maintained that to explain possibilities of something, one must undertake transcendental arguments, i.e., proceed by transcendental deduction. These arguments consist in assuming premises necessary for understanding the factuality of something. Thus, we know that synthetic *a priori* propositions exist. However, something must be assumed in order to show that unless it were true, the synthetic *a priori* would be impossible. Clearly, the existence of synthetic *a priori* propositions cannot be derived directly from experience. On the other hand, one must analyze knowledge, including experience itself, in order to find elements which legitimize the synthetic *a priori,* Kant realized his task in three parts. First, he developed transcendental aesthetics considered as the theory of forms of pure sensory intuition. Then, he passed to transcendental analytic, that is, the theory of categories used by reason, and, finally, he ended the *opus* with transcendental dialectics which, according to Kant, proved the impossibility of metaphysics.

For Kant, sensory experience is the starting point of knowledge. However, experience is always ordered by temporal and spatial relations. Time and space manifest themselves as universal and thereby necessary ingredients of all experiences. Since universality and necessity cannot be *(pace* Hume and Leibniz) derived from experience alone, we must recognize that time and space are *a priori:* they are rational (existing in reason) forms of sensory intuition. If we cancel all empirical content with our experiences, space and time still remain. This is an outline of Kant's derivation of space and time as transcendental conditions of sensory experience. Mathematics is the repertoire of synthetic *a priori* propositions based on time and space. Arithmetic recurs to temporal relations, geometry appeals to space. Since every synthetic proposition is an effect of synthesis of subject and predicate, space and time are responsible for the synthesis of mathematical propositions.

Theoretical physics is for Kant another domain in which syntethic *a priori* propositions occur. But truths of theoretical physics have a different nature than propositions of arithmetic and geometry. The new kind of synthetic *a priori* truths, discovered by transcendental analytic, is based on categories as *a priori* concepts possessed by reason. Due to categories like causality, quality, modality, etc. (Kant distinguished 24 categories, related to 24 forms of propositions), we can form (synthesize) theoretical principles of natural science, for instance, the basic law of the Newtonian mechanics. Mathematics, which consists of synthetic *a priori* truths based on space and time, and natural theoretical science, i.e., synthetic *a priori* truths based on categories, exhaust the proper knowledge in which the synthetic *a priori* is involved. However, reason also has a tendency to transgress its own cognitive faculties and enter into metaphysical speculations. Kant, using ingenious arguments, tried to prove that metaphysics is in principle impossible. For Kant, traditional metaphysics consisted of rational psychology, rational cosmology and rational theology. Each part of metaphysics produces antinomies, because we can prove and disprove that the soul is simple and complex (showing that rational psychology is impossible), that the universe is finite and infinite (showing that rational cosmology is impossible), and that God exists and does not exist (proving that rational theology is impossible). The only metaphysical thesis that can be established theoretically, that is, by analysis of pure reason, is that there are phenomena and noumena *(Dinge an sich).* The latter are unknowable, but, *via* transcendental arguments, we must accept their existence in order to explain the existence of the former: noumena are causes of phenomena. Kant did not reject the existence of the immortal soul and God, but he argued about these matters practically, not theoretically: the soul and God were derived from postulates of practical reason. Yet Kant very strongly separated being *(Sein)* and oughtness *(Sollen).*

Kant influenced the further course of epistemology in many respects. He introduced a new epistemological position, namely moderate apriorism. His distinction of analytic and synthetic propostions became one of the most important conceptual devices. The view that logic is analytic, but mathematics synthetic *a priori* became one of the main positions in the foundations of formal science; in particular, mathematical intuitionists share this view. Transcendental arguments, phenomenalism with simultaneous acceptance of the existence of the real world as something outside phenomena, the critique of the ontological argument for the existence of God based on the observation that existence is not a predicate, a sharp border between being and oughtness, not only logical as in the case of Hume, but also ontic and epistemological or practical justification of ethical categories became the constant subject of philosophical debates. Kant had devoted defenders and radical critics. Perhaps no philosopher was so extensively commented upon and perhaps the formula 'every post-Kantian is proper-Kantian' describes Kant's influence more properly than similar statement in the case of any other philosopher. But maybe, the following remark is of a special importance. Kant sharply distinguished *de jure*-questions and *quid facti*-questions. While the former refer to the problem of legitimacy, e.g., of the synthetic *a priori* and require transcendental arguments, the latter concern factual questions, e.g., whether there are such and such phenomena. Kant, like nobody else, considered epistemology as concerning *de jure*-questions. The real importance of this view is obvious for any attempt to naturalize epistemology, especially today, when the theory of knowledge is often regarded as a part of cognitive science. When one stresses Kant's influence, it does not mean that his philosophy has no weak points, independently of its historical importance. What does it mean that *Dinge an sich* exist, but that they are unknowable? Is not it Kant's account of things-in-themselves is a piece of obscure metaphysics? What are transcendental arguments? Are they logical in the usual sense or do they require a special transcendental logic? How does Kant's theory of space deal with non-Euclidean geometries? What is reason in Kant's sense? It is individual reason or something transcendental? Was Kant a psychologist? These and other questions as well as different answers to them decided that Kant's heritage became extremely diversified.

It is clear that British empiricism was attacked by rationalism, and why it was so. However, it was done not only from the rationalistic point of view. Thomas Reid (1710-1796), a Scottish philosopher, shared empiricism, but did not agree with the account of knowledge developed by Locke, Berkeley, and Hume. Reid wrote three important treatises *Enquiry into the Human Mind on the Principles of Common Sense* (1764), *Essays on the Intellectual Powers of Man* (1785) and *Essays on the Active Powers of Man* (1788). The title of the first book indicated the direction of Reid's epistemology and his attack on his predecessors. The expression 'common sense' is the key term here. According to Reid, British empiricists, mainly Berkeley and Hume, advanced a philosophy that was completely at odds with common sense. Reid particularly strongly attacked the view that ideas in the psychological sense are media between our minds and things. He agreed that sensory experience was the primary source of knowledge. However, according to Reid, knowledge concerns not ideas but things and their qualities directly. Thus, ideas are natural signs referring to qualities and there is no reason to distinguish primary and secondary qualities. In this function, ideas are similar to words which signify directly something. This analogy shows that mind transcends itself and reaches the real world. Reid was a direct realist: our knowledge is acquired directly without any mediation of ideas. Moreover, the principles of common sense must be obeyed, because their abandonment leads to absurdities as in the case of Berkeley or unjustified scepticism in the style of Hume. Reid, who was not properly appreciated in his time, began a new kind of epistemology, namely a commonsense empiricism based on direct realism. His influence was local, mainly limited to Scotland. However, in a more general perspective, he was a predecessor of G. E. Moore and the theory of direct perception.

I conclude this section with a few remarks about the concept of truth from Descartes to Kant. The phraseology of *adequatio* introduced by the Schoolmen was very popular. Descartes spoke about the conformity of thought and object, Locke about the agreement of ideas with existence in nature, Spinoza about the convenience of ideas with ideated objects, Leibniz about the correspondence between propositions and something else, Wolff about the consensus of propositions and things represented by them, and Kant about the adequacy *(Übereinstimmung)* of knowledge to its object. However, these explanations marked quite different contents. In fact, Descartes defended a kind of the evidence theory: a judgment is true if it is evident, namely composed of clear elements. In their rationalistic systems, Spinoza, Leibniz, and Wolff developed the coherence or identity theory of truth, rather than its classical account. Kant attacked very strongly the correspondence theory of truth, because, under his general views, it was impossible to compare the content of knowledge with its independent object. Since the object of knowledge is construed by reason from experiential material with the help of *a priori* ingredients, the formula about the adequacy of knowledge and its object must mean something different than in Thomas Aquinas. It is a remarkable fact that foundationalist epistemology from Descartes to Kant was supplemented by such varying theories of truth.

5. MODERN PHILOSOPHY AFTER KANT

(A) German idealism and Neo-Kantianism

As I have already noted, Kant's philosophy elicited numerous responses, including critical ones. The most influential response came from German transcendental idealists, notably Johann Gottlieb Fichte (1762-1814), Friedrich Wilhelm Schelling (1775-1854) and Georg Friedrich Wilhelm Hegel (1770-1831). Fichte rejected Kantian *Dinge an sich* and argued that the independence of things in their relation to consciousness is merely an appearance stemming from an unavoidable distinction between subject and the object. According to Fichte, this distinction is correct, because it is constituted by the first act of the Self (the Absolute). However, things are nothing more than presentations, though not of natural consciousness, but of the transcendental Self. Since the Self creates the subject and the objects, both are of the same nature. Although he deliberately strengthened transcendental elements much beyond Kant's horizon, Fichte came back to the classical rationalistic tradition on which deduction was the only source of necessary truths. His method was simply deductive, recurring everything to the first principle of the Self. Fichte's main work was *Grundlage der gesamten Wissenschaftslehre* (1794-1795). It was just this treatise that introduced the term 'Wissenschaftslehre' as a label for epistemology.

If we agree that the relation of the subject to the object became the central problem of post-Kantian philosophy, Fichte must be recognized as a philosopher who maximized the role of the Self. Schelling tried to come back to a more balanced picture of the relation between both categories. Hence, Schelling's philosophy was much more concerned with the philosophy of nature than Fichte's theory. For Schelling, the Absolute transcends the Self and the Nature. This view determined Schelling's metaphysics, which was similar to ontological ideas of Spinoza. In epistemological matters, Schelling followed Fichte and admitted autonomous knowledge acquired by the mind itself *(intuitus intellectualis,* intellektuelle Anschauung). However, it must be remembered that, like in the case of Fichte's Self, the mind is not to be identified with particular human faculties: it is the transcendental Mind. Schelling's philosophy, although very speculative, influenced the philosophy of nature, even the science of the 1st half of the 19th century (Henrik Steffens in Norway). In general, Schelling's ideas were welcomed by the representatives of Romanticism in all areas of this important cultural movement.

It was Hegel who created the most powerful system of German transcendental idealism, perhaps even the strongest system of idealism since Plato. For Hegelian epistemology, the principle that thought is identical with being was of the utmost importance. Since being changes all the time, the same concerns thought. Changes are not accidental, but regular and evolutionary: higher stages appear after less advanced ones, according to general necessary rules. Those rules are logical in character and, hence, Hegel's idealism is sometimes termed as 'logical'. However, Hegel's logic was conceived as something opposed to traditional formal logic. In order to stress this difference, Hegel spoke about dialectic as the new logic which was regarded as the proper method of dealing with changeable reality. In particular, the dialectic does not preserve the rule of contradiction. In general, dialectic of thought and reality proceeds by three stages: thesis, antithesis and synthesis. The last unifies the former two, which are contrary. Thus, if *A* is a thesis, not-*A* is its antithesis, the dialectical synthesis embraces *A* and not-*A*. Therefore, every synthesis unifies contrary elements in the real whole.

Hegel's philosophy was radically rationalistic. Everything real is rational, logical and necessary. However, the dialectic method is not deductive: when Hegel speaks about entailment, it must be understood as a special intensional relation referring to dialectic principles, like, for example, the rule of the development by successive negations or the rule of the unity of oppositions. In fact, Hegel offered a new philosophical method which was practized by many later philosophers. One more aspect of Hegel's epistemology is worth mentioning. For Hegel, reality is essentially historical. The later stages of evolution cover the former ones: reality brings its own history, according to Hegel's famous view. This view was novel and influential, especially for the development of the epistemology of humanities. It also inspired a fundamental question: is objectivity possible, independently of a historical perspective? Hegel's answer was affirmative: it is possible, but only if we take into account the transcendental perspective which finds its ultimate realization in the last stage of the cosmic evolution, that is, in the Absolute Spirit. Anyway, the individual subject is always subordinated to a totality: state, nation, etc., also from the epistemological point of view. These views constantly reappeared in the post-Hegelian philosophy, particularly in Marxism and several other currents interested in social matters.

The orthodox Hegelianism came to its end in Germany very soon after Hegel's death. It was revived by Francis Herbert Bradley (1846-1924) in England. He shared a general Hegelian idea about the conformity of reality and knowledge. Since Reality forms a whole whose which parts are interconnected, the same concerns its knowledge. Every conceptualization of Reality which results with its structuring into particular separate facts is a simplification and produces only partial and fragmentary knowledge; in essence, facts must be regarded as our constructs. These ideas led Bradley to the coherence theory of truth. We can speak of the correspondence of knowledge and reality as the relation between two integral wholes. The correspondence theory of truth is correct in this perspective but not with respect to particular propositions and pieces of reality. Elements of knowledge may be only partly true, that is, at most they may possess a degree of truth. Since facts are our constructs, it is impossible to compare them with concrete propositions. We can only compare propositions with other propositions and investigate internal coherence inside the whole which constitutes knowledge; consistency and comprehensiveness are marks of coherence. In the ultimate perspective, coherence and correspondence are the same: Knowledge and Reality as integrities remain in mutual correspondence and both are internally coherent, that is, consistent and comprehensive. Bradley was a very influential philosopher. His ideas attracted many British philosophers, so that at the turn of century, the British Neo-Hegelian School arose (John McTaggart, Harald Joachim, Bernard Bosanquet, and others); Brand Blanshard, an American philosopher also joined this group.

Although transcendentalism was the most important form of German idealism of the first half of the 19th century, there other idealistic tendencies also appeared of which ideas of Friedrich Schleiermacher (1768-1834) and Arthur Schopenhauer

(1788-1866) became particularly important for the development of epistemology. Schleiermacher, who was also a poet, was strongly influenced by the ideology of German Romanticism. He translated of all Plato's dialogues into German. His literary and translatory work became an important source of his general philosophical ideas. In particular, Schleiermacher intended to set forth rules according to which we acquire knowledge. Since thoughts are dressed in linguistic forms, the ways of operating words were of the utmost importance for Schleiermacher's epistemology. Thus, the role of language in knowledge led Schleiermacher to the idea of hermeneutics as the general theory of understanding. Formerly, that is, before Schleiermacher, hermeneutics was understood as an art of interpretation in theology, philology and jurisprudence; it should be rather said that these disciplines had their own separate hermeneutical rules. Schleiermacher was the first to propose a general hermeneutics aimed to be a method applicable to any thought expressed in language. Moreover, Schleiermacher observed that the understanding of a text was also involved in the situation usually called the hermeneutical circle: in order to understand a given text, we need to understand something else, for example another text or culture which also is symbolic in its essence. The problem whether it is possible to break the hermeneutical circle in order to obtain a fully objective basis of understanding or whether every hermeneutical interpretative act is based on an earlier understanding (or at least, a preunderstanding) became fundamental for later discussions about hermeneutics. In general, the hermeneutic tradition has constantly been of the center of the philosophical scene since about 1850. An important figure in this movement was Wilhelm Dilthey (1833-1911) who claimed that understanding had to be based on empathy (a special kind of "infeeling" with the reflected objects) and on the objectivization of the spirit of culture. This last moment, similarly as historicism, linked Dilthey with the Hegelian tradition. The twofold activity of empathy became a prototype of the operation of *Verstehen.* very extensively studied in the methodology of humanities.

To a great extent, Schopenhauer's philosophy was a response to Kant. However, Schopenhauer went in a different direction than transcendental idealists. While Fichte, Schelling and Hegel were rationalists, Schopenhauer based his philosophy on irrationalism. The title of his main work *The World as Will and Representation* (1818, sec. extended ed. 1844) is very instructive. Schopenhauer's main philosophical view is voluntaristic: the world appears as our representation which is rooted in our will. The analysis of knowledge led him to the view that products of our imagination are the ultimate data. On the other hand, our self-knowledge inevitably informs us that we, as subjects, are reducible to the will. Kant's doctrine about subjects and *Dinge an sich* was simplified to the following form: we as knowing subjects are also objects of knowledge, things in themselves in a sense. Since will is a constitutive element of everything, and it is irrational, the world and its knowledge are also irrational. This metaphysical and epistemological view was supplemented by Schopenhauer's radical pessimism as a consequence of voluntarism.

In 1865, Otto Liebman published a small book *Kant und die Epigonen.* At the end of each chapter of this work there is a phrase: *Es musst auf Kant zurückgegangen* (One must come back to Kant). Liebmann's book is directed against several post-Kantian philosophers, including not only, Fichte, Schelling, Hegel and Schopenhauer, but also Johannes Herbart, and Jacob Fries, who regarded themselves as Kant's successors. Liebmann argued that they rather essentially departed from Kant than continued his ideas. Thus, according to Liebmann, we should come back to Kant himself, because it is the path to renewing philosophy. Liebmann's work is usually regarded as the beginning of the Neo-Kantian movement. It is to some extent controversial how many Neo-Kantian schools should be distinguished. The present standard account consists in the bipartite division into the Marburg School and the Badenian (Southwest) School). *Ad casuum* of this essay, I will add the third branch, namely the Neo-Friesean School, although I do not insist that the standard picture of Neo-Kantianism should be necessarily revised. Including the Neo-Friesean School into Neo-Kantianism is rather dictaded by *ad hoc* reasons stemming from the plan of this survey. In general, all Neo-Kantian movements rejected things in themselves and concentrated on the faculties of the subject as conditions of knowledge. In a sense, Neo-Kantianism became more epistemological than its master himself was.

Hermann Cohen (1842-1918), Paul Natorp (1872-1924), and Ernst Cassirer (1874-1945) were the main representatives of the Marburg School. This school abandoned Kant's view that knowledge was a synthesis of empirical data *via* aprioristic forms inborn in the mind. Cohen replaced this picture of knowledge with a much more rationalistic account on which aprioristic forms were conceived as universal necessary conditions of knowledge to be realized by every subject. Of course, the Marburgians maintained that Kant himself represented this view, but he was not consistent and made too many concessions to empiricism and psychologism. Cohen and his successors were looking for a new pure logic of science. Its outline was exposed in Cohen's main work *The Logic of Pure Knowledge* (1902). Although the Marburgians worked on all fields of philosophy, including ethics, aesthetics and social philosophy, the philosophy of exact sciences became their main concern. Cassirer applied the principles of the Marburg School to a historical and systematic examination of natural science. He stressed the importance of the distinction between form and content of knowledge as well as the role of symbolic elements in all fields of human activity. Due to the role of symbolism, Cassirer considered the humanity as *animal symbolicum.* The results of Cassirer's historical studies are contained in his monumental work *The Problem of Knowledge in the Modem Science and Philosophy*, 3 vls. (1920); volume 4 was published in English in 1950.

The Badenian Neo-Kantian School started with Wilhelm Windelband (1848-1915) and achieved its maturity in the works of Heinrich Rickert (1863-1936). For Windelband and Rickert philosophy was essentially concerned with values. There are principally three kinds of values: logical, ethical and aesthetic. According to this division, truth, goodness and beauty are the species of validity *(Geltung)* of values. The source of validity is transcendental due to a special kind of consciousness generating universally valid and intersubjective norms which are displayed by culture. Thus, the former ontologically oriented ethics and aesthetic were replaced by an epistemologically oriented philosophy of values. However, the Badenians claimed that this change did not result in subjectivism. In his main epistemological work *The Object of Knowledge Introduction to Transcendental Philosophy* (1892), Rickert outlined the general epistemological theory. He criticized the traditional account of knowledge in which cognition consisted in representing objects in mental contents. For Rickert, knowledge is based on the epistemological oughtness which demands that some propositions must be recognized as coherent with epistemological norms. This oughtness manifests transcendentality of knowledge. Ontologically speaking, Rickert's epistemology leads to idealism, because reality appears as a counterpart of the Subject acting accordingly to transcendental epistemological rules. The Southwest School was much more interested in the foundations of humanities than the methodology of natural science. Windelband introduced a famous distinction of idiographic and nomothetic disciplines. Although the latter formulate universally valid laws, the former aim at a detailed description of their objects. Mathematical physics is a paradigm of nomothetic science, but history is an example of idiographic discipline. Rickert replaced this picture by a more general division into generalizing and individualizing concept formation. He did this in one of the most influential treatises in the whole history of the philosophy of the humanities, namely *The Limits of Concept Formation in Natural Science, A Logical Introduction to the Historical Sciences* (1902). For Rickert, we have *Kulturwissenschaften* (science of culture) and *Naturwissenschaften* (natural science). The individualizing concept formation is characteristic for history which is the base of science of culture. Since the transcendental character of values requires recognition of some cultural values, the individualizing concept formation recurring to values gives a possibility of recognizing the significance of a considered historical event in the wider historical context.

The Neo-Frisean School was a continuation of the philosophy of Jacob Friedrich Fries (1773-1843) who was already mentioned among Kant's *Epigonen* in Liebmann's understanding. Fries criticized Kant in his basic work *New Critique of Reason* (1807), revised and published in 1838 under the title *New Anthropological Critique of Reason.* Fries' main objective was to defend critical philosophy against the excesses of transcendental idealism, particularly that of Fichte. However, Fries claimed that this task required a reinterpretation of Kantian philosophy. His revision of Kant consisted in replacing transcendental deduction of *a priori* categories by their anthropological explanation. Fries is famous for his triemma which displays some traditional difficult problems of epistemology. The question is: How to justify our beliefs? If we claim that all our statements should be justified, we have to reject dogmatism. Now, we cannot justify everything by logical demonstration, because this leads to *regressus ad infinitum.* Hence, we must recur to still another method, namely the anthropological one. It was the source of the idea of the regressive method, a special device of justifying our basic beliefs by immediate elementary experience. Fries claimed that the regressive method should replace Kant's transcendental deduction. This proposal is usually qualified as pushing Kant's doctrine into psychologism, although Fries did not abandon the category of *a priori,* but insisted that it had to be legitimized by the experience of human intellectual devices.

The Neo-Friesian School was established by Leonard Nelson (1882-1927) who tried to refine Fries' ideas concerning the criticism of epistemology. He did this in his fundamental work *On the So Called Problem of Knowledge* (1904). According to Nelson, the traditional task of epistemology, that is, the demonstration of the validity of knowledge is hopeless. This task is meaningful if we consider results of our cognitive acts as at least temporarily problematic. However, it must also concern epistemology itself under the danger of *petitio principii.* Moreover, if we agree that epistemological statements are to be suspended as problematic, we have no chance to perform our initial task unless *regressus ad infinitum* or *circulus vitiosus* are to be tolerated. Hence, logically correct epistemology is impossible, because *petitio principii, regressus ad infinitum* and *circulus vitiosus* are elementary logical errors. There is a very nice wording of Nelson's argument in more recent terminology. The task of epistemology requires that we should suspend all synthetic propostions. Thus, we can only assume that analytic propositions are valid. Since analytic truths do not entail synthetic ones, we have no basis for inferences proving the universal validity of epistemological principles. It is important to see the real gist of Nelson's critique. His conclusions are completely different from sceptical ones. Nelson did not argue that knowledge was impossible. He intended to prove that epistemology in its traditional form was inevitably burdened by logical defects. In order to save the situation, he used the regressive method as a remedy, admitting that it could not lead to any certainty other than psychological.

The end of the Neo-Kantian movement came after World War I. Today Neo-Kantianism is often interpreted as a typical academic philosophy which was not able to survive the confrontation with the new science and its philosophy. Neo-Kantianism had serious problems with accommodating the new physics or logic, but, on the other hand, the influence of this movement on the development of philosophy was enormous. This qualification concerns particularly the development of philosophical discussion about the foundations of the humanities and social sciences. Without any exaggeration we can say that the main points of contemporary controversies concerning the nature of these fields were stated by the Neo-Kantiants and Dilthey. It is worth noticing that the Neo-Kantian camp was populated not only by pure philosophers, but also practitioners of particular disciplines. Max Weber (sociology), Karl Mannheim (sociology), Florian Znaniecki (sociology), Rudolf Stammier (law), Hans Kelsen (law), Ernst Troeltsch (history), Karl Vossler (linguistics), Heinrich Wölfflin (history of art) and Max Dvorák (history of art) provide sufficiently strong evidence of the significance of Neo-Kantianism. But also some natural scientists, notably Hermann Helmholtz, were influenced by Neo-Kantianism. It is not a casual circumstance that sociologists (particularly, Mannheim and Znaniecki) related to Neo-Kantianism belonged to the pioneers of the sociology of knowledge. This movement, especially the Badenian School, also contributed to the view that scientific hypotheses are symbolic creations of the human mind. Nelson, who was a philosophical hero of David Hilbert, was particularly important. In fact, Hilbert's finitism in the foundations of mathematics was an application of the regressive method. It is perhaps interesting that the Neo-Friesian School was the only Neo-Kantian School which properly estimated the philosophical significance of mathematical logic: Paul Bernays, a close collaborator of Hilbert and Kurt Grelling, a distinguished logician, belonged to the Nelson circle. Rudolf Carnap and Hans Reichenbach, the leaders of logical empiricism began as Neo-Kantians, and this fact also shows how strong Neo-Kantianism was.

(B) Positivism, materialism and psychologism

This variety of standpoints covers several views that tried to reduce epistemology (and philosophy in general) to special sciences. Positivism arose in France and continued the anti-irrationalistic and empiristic traditions of the French Enlightenment. The main principles of positivism were formulated by August Comte (1798-1857). According to him, knowledge should serve practical needs: we know in order to predict and we predict in order to be able to realize our tasks, mainly those, that are connected with improvement of social life. Strongly believing in the great potential of science for practical ends, Comte divided history into three long periods: religious (knowledge is based on myths), metaphysical (knowledge is based on speculation) and positive (knowledge is based on science). Science must be restricted to facts, because going beyond empirically accessible data leads to speculation. Comte understood facts naturalistically and physicalistically, and he rejected inner or psychological facts. He divided sciences into the abstract and concrete. There are six abstract sciences: mathematics, astronomy (celestial mechanics), physics, chemistry, biology and sociology. This sequence is not accidental, because it proceeds from more to less general sciences and, moreover, less abstract fields are based on more abstract ones. This picture justifies reductive programs in science, for example, chemistry to physics or sociology to biology. Concrete sciences are associated with related concrete disciplines, for instance zoology with biology or history with sociology. This last discipline was a novelty, because it did not occur in the classifications of sciences proposed before Comte. He divided sociology into social dynamics (theory of social changes) and social statics (theory of social structure). Although he did not do any empirical sociological research, he is commonly regarded as the inventor of sociology. It is important to stress that Comte's idea of social science was completely different from that proposed by Hegelians and Neo-Kantians. Since, for Comte, sociology as a science must obey the methodological criteria of general science, which are the same for natural and social disciplines, this claim gave the rise to the positivistic project in humanities and sociology. It is perhaps an important feature of Comte's classification of sciences that it had no room for psychology and philosophy. The first was rejected for physicalism, and the second became the theory of science. In a general perspective, Comte offered a foundationalist epistemology regarding scientific statements as indubitable. He criticized scepticism on the one hand and probabilism on the other. It is certainly a paradox that his radically scientific attitude produced a vision of social development, that was fairly speculative.

Positivism rapidly became a very popular philosophy. Its minimalism attracted many scientists, its optimism many social reformers. It was also in accord with the development of science in the first half of the 19th century. For example, Comte's view that sociology is located as the next abstract science after biology, was welcomed by philosophers like Herbert Spencer (1820-1903), who tried to build social science based on Darwin's theory of biological evolution. The positivist style of thinking was also influential outside philosophy, particularly in jurisprudence (legal positivism) and literature (literary positivism). In general, positivism succeeded Romanticism and its ideology. Positivism found many representatives among British philosophers. John Stuart Mill (1803-1873) was the most distinguished philosopher of British positivism. He shared the general positivistic principles (naturalism, empiricism, foundationalism) in Comte's version with one exception: Mill admitted psychology. Mill was a radical empiricist, both genetic and methodological. Expressing this in Kantian terms, we can say that, according to Mill, all propositions admitted as results of knowledge are synthetic *a posteriori*. He also contributed to the logic of science. In his famous and influential *System of Logic: Ratiocinative and Inductive* (1843), Mill tried to formulate general principles of empirical inductive research, the so called principles of eliminative induction. The positivistic thought of Comte, Mill and their allies is called the first positivism. Richard Avenarius (1843-1896) and Ernst Mach (1838-1916) developed the second positivism (empiriocriticism). This form of positivism followed Hume's ideas to some extent. According to empiriocriticism, knowledge concerns neutral elements which are neither objective nor subjective. Any question about the nature of elements is meaningless. Knowledge, which empiriocritics identified with science, is governed by the principle of economy of thought demanding the simplest conceptual framework. Thus, explanation of phenomena by looking for their causes rules out the tasks of science, which remains descriptive and predictive. Since Mach spent his last years in Vienna, his ideas influenced the Vienna Circle (see (G) below).

The borderline between positivism and materialism is often very rough. In general, materialism explicitly answers the question of the nature of the world as just entirely consisting of matter, but such a view is typically qualified as meaningless by the positivists. On the other hand, the development of materialism was caused by similar circumstances as those contributing to the rise of positivism, namely the successes of natural science, in particular chemistry (the atomic theory of John Dalton) and biology (the first synthesis of an organic compound by Friedrich Wöhler in 1828). The first form of materialism in the 19th century was elaborated rather by scientists than philosophers. It was the naturalistic materialism of Ludwig Büchner (1824-1899), Jacob Moleschott (1822-1893) and Karl Vogt (1817-1895). Their materialism is sometimes termed as vulgar for its radical project of the reduction of everything to naturalistically understood matter. It was foremost an ontological view, but with some fundamental epistemological consequences. The most important of them was this: knowledge is a material process (Vogt: the relation between thought and brain is similar to that of bile and liver or urine and kidneys) which can and should be investigated by physiological methods.

Karl Marx (1818-1883) and Friedrich Engels (1820-1895) developed another form of materialism, which became the general philosophical foundation of Marxism as an extensive intellectual system. Here, we are only interested in Marxist epistemology. However, there is a deep controversy about the interpretation of Marxist philosophy which also concerns epistemological matters. Standardly, Marxism is described as a whole basically consisting of dialectical materialism (epistemology and ontology) and historical materialism (social philosophy). If epistemology is regarded as a part of dialectical materialism, it is usually interpreted as the theory of reflection of objective reality by the subjective mind. This way of looking at Marxist epistemology was proposed by Engels and later endorsed by Vladimir I. Lenin (1870-1924). However, there is also another approach to the matter, namely interpreting Marxist epistemology through the glasses of historical materialism, especially the theory of class consciousness and its impact on the individual consciousness. This approach is rooted in the early writings of Marx and Engels (to be more precise, Marx is much more relevant in this respect). Then, it was continued by György Lukâcs in Hungary and the Frankfurt School, but almost absent (maybe, except Poland) in the Soviet block. This second interpretation is much more Hegelian than the first, and its main thesis is that class consciousness always displays current economic situation. If economic life is based on the private possession of productive means and resources, this fact results in alienation of consciousness, which also has an individual dimension. Thus, knowledge is associated with a more or less determinate perspective which displays itself in ideology, law, morality, religion or philosophy. The total elimination of the alienation of consciousness is possible only by a change of the economic base into the system without private industrial property and without class conflicts. It resembles Hegel's doctrine of the absolute spirit as the final stage of dialectic development in which all contradictions are resolved. Independently of historical controversies about the interpretation of Marxism, it is important to point out that the second interpretation of Marxism strongly influenced sociology of knowledge. Thus, the contemporary sociology of knowledge is a child of Neo-Kantianism and Marxism to a great extent.

I already mentioned psychologism on the occasion of Fries and Mill. In fact. Fries is usually regarded as one of the founders of psychologism; the other is Friedrich Eduard Beneke (1798-1854). In the beginning, psychologism served means of reinterpreting Kant's philosophy or as a foundation for explaining the nature of logic. Later, due to the successes of experimental psychology (the first psychological laboratory was established by Wilhelm Wundt in Leipzig in 1879), psychologism became very influential in various fields, not only in philosophy, but also in jurisprudence, history of art and linguistics. It was regarded as a proper account of the ontological status of various objects investigated by philosophers, lawyers, historians of art, linguists, etc. Numbers, paintings, sculptures, language, law, morality, literary works, reasonings, values, etc. were conceived as psychical objects existing in the human mind. Hence, disciplines investigating such objects were considered as parts of psychology (perhaps except mathematics: psychologism was popular in the philosophy of mathematics, not in mathematics itself) This tendency also concerned epistemology, because looking at knowledge as a psychological phenomenon seemed fairly natural. Epistemological psychologism culminated in the book *Psychophysiological Theory of Knowledge* published by Theodor Ziehen in 1898.

Criticism of psychologism by Gottlob Frege and Edmund Husserl was effective to a great extent, but not fully, because this position was always attractive to philosophy understood as an empirical science. 19th century psychologism was based on introspective psychology. New prospects for psychologism were linked with behaviourism (recently Willard van Orman Quine) which arose as an attempt to transform psychology into science focusing only on external behavioural facts. Another influence of psychology on epistemology came from *Gestalt* theory. This psychological theory was developed by several psychologists (i.a. Christian von Ehrenfels, Vittorio Benussi, Max Wertheimer, Kurt Koffka and Wolfgang Köhler) at the end of the 19th. According to *Gestaltists,* human perception is directed not to ultimate elements of things and events, but wholes organized as *Gestalten.* Thus, identification of objects is made in the context of overall *Gestalten.* This view particularly influenced the theory of perception. In a sense, the recent proposal of reduction of epistemology to cognitive science echoes the program of psychologism. All currents reported in this section, perhaps with the exception of Marxism interpreted as Hegelianism, offered programs of naturalized epistemology.

(C) Logical objectivism

This position can be attributed to Bernard Bolzano (1781-1848), Hermann Lotze (1817-1881) and Gottlob Frege (1848-1925) as its founding fathers, at least in the post-Kantian philosophy. However, one should also remember that several points characteristic for logical objectivism were also present in Neo-Kantianism. Bolzano's *Theory of Science* (1837) is an extensive treatise on logic and methodology of science. Although it does not contain any systematic exposition of epistemology, several Bolzano's considerations are of the utmost importance for the analysis of knowledge. Bolzano introduced the concept of proposition in itself, an objective entity which was independent of particular human acts. This category enabled Bolzano to fight against scepticism and relativism. For example, he defended the absoluteness of truth and, against Kant, realism concerning the real world. Bolzano also defined several important concepts, like analyticity, logical truth and logical consequence. These definions were given with the help of semantic terms. Unfortunately, Bolzano's work was not properly appreciated during his lifetime or later. He was discovered too late and did not influence the development of philosophy in the way he deserved.

Lotze revived Platonism but with a new interpretation: he justified ontological objectivity *via* epistemological validity, that is, conversely to Plato. Lotze introduced the idea of *Geltung* (validity) which became so important for Neo-Kantians; let me note that the term *Geltung* was earlier used by Bolzano, but in a less general meaning. For Lotze, the world of Forms was primarily the realm of objective mental contents which existed in another mode than that of spatial things. This world of objective contents is the fundament of epistemological objectivity which appears as the truth of propositions, independently of mental acts. There we have a close affinity of Lotze to Bolzano's account of propositions in themselves. Universal epistemological validity is responsible for truths *a priori.* However, they are not grasped by pure deduction, but discovered as preconditions of rational thinking. Although this sounds Kantian, Lotze did not regard *a priori* truths as innate ingredients of the mind, but rather as elements of the objective realm of contents. Lotze's influence was much stronger than Bolzano's. In particular, Gottlob Frege participated in Lotze's course in the philosophy of religion. It is probable that Frege's philosophical horizon was Lotzean at least to some degree.

Frege's main historical merit consists in discovering mathematical logic and formulating logicism as one of the main positions in the foundations of mathematics. Until the 1950s, Frege was not perceived as a philosopher, but mainly as a logician. He certainly influenced philosophy of the first decades of the 20th century at least in three respects, namely by (a) his criticism of psychologism *(a priori truths* cannot be discovered empirically, but psychology is empirical; (b) the distinction between sense and reference; (c) the analysis of *a priori* and analytic truths as reducible to logic and definitions. However, he also formulated several other ideas relevant for epistemology, i. a., that truth is indefinable, that the predicate 'is true' is redundant, that the correspondence theory of truth is untenable, that truth is absolute, and that propositions (thoughts) in the logical sense are independent of particular mental acts. It is certain that Frege did not read Bolzano, but it is very much debated how far Fregean ideas were dependent of Lotze. The lack of further historical evidence going beyond Frege's participation in Lotze's course makes it impossible to answer this intriguing historical question.

Independently of the actual historical influences, the similarity of views advanced by Bolzano, Lotze and Frege is striking. It is a remarkable fact that all three were inspired by different circumstances: Bolzano by traditional logic, Lotze by philosophy, and Frege by mathematical logic. Nevertheless they offered similar views about knowledge and its foundations. In particular, they proposed the theory of the objective, atemporal realm which is the subject of genuine knowledge. All three were realists and anticipated the Popperian idea of the third world as something between the world of psyche and the world of things. It is interesting that Frege used the term *die dritte Reich* which is a strict counterpart of the name "the third world" in Popper's sense.

(D) Brentanism and phenomenology

In his *Psychology from an Empirical Standpoint* (1874) Franz Brentano (1838-1917) developed a new conception of the psyche and psychology. For him, the psyche is not a collection of sensations or other psychic atoms, but it consists of mental acts. The traditional task of psychology, that is, the analysis of the is of the psychic life and the formation of psychic items according to the laws of association must be supplemented by a careful description of mental acts. Thus, Brentano contrasted genetic with descriptive psycholgy, although he did not wholly abandon the former, but claimed that this approach was not able to give an adequate account of psyche. Even more, according to Brentano, it was descriptive psychology which captured the distinctive feature of psychic phenomena, namely intentionality. Brentano introduced the concept of intentionality in the following way:

> "Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction toward an object (which is not to be understood here as a meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgment something is affirmed or denied, in love loved, in hate hated, in desire desired and so on." (E Brentano, *Psychology from a Empirical Standpoint,* tr. by A. C. Rancurello, D. B. Terell, and L. L. McAlister, Routledge, London 1995, p. 88).

Doubtless, this is one of the most important passages in the history of modern philosophy. It explains the historical sources of the concept of intentionality, but, then, it tries to explain this phenomenon. For Brentano, intentionality is a primitive and fundamental property of mental acts so perfectly and evidently known that the quoted description should be taken just as an intuitive explanation, not as a definition. The intentional connection is completely different from the causal nexus.

It is the main reason why genetic psychology working with the concept of causality was not able to give an adequate description of psyche. Since Brentano characterized intentionality as reference to content or direction toward an object, it is clear that every mental act, according to his view, has some content. Thus, we have the distinction between acts and their contents. The content of a given act is known by the evident direct inner experience. It was Brentano's starting point for various epistemological theses. In particular, he strongly criticized the correspondence theory of truth which required a comparison of contents with things. However, such a comparison is simply impossible; I omit Brentano's other, more specific objections against the concept of correspondence. Brentano did not reject Aristotle's theory of truth, but proposed to interpret it as a kind of the evidence theory. In his later period, Brentano became a reist and argued for the existence of concreta as the only objects. He rejected all *irrealia* like contents, abstract objects, etc. and claimed that our knowledge concerned only concreta. This new position did not change his general epistemological views: realism, antirelativism and antiscepticism.

Brentano was an eminent teacher. His numerous students inherited his main general epistemological views indicated above. However, there was a point in Brentano's approach to intentionality that was unclear; Brentano fairly admitted that it actually was. Let me once more recall that Brentano spoke about reference to some content or direction toward an object. Are the phrases 'reference to some content' and 'direction toward an object' synonymous? Can we, for instance, say without a change of meaning that acts refer to objects and are directed toward contents? Briefly: are contents and objects of acts the same items? Everybody agreed that there was a great advantage of intentionality-talk, namely it allowed for an explanation what was going on when one spoke about fictions, because intentionality was independent of the real existence of what acts were directed to. Yet the alleged identification of objects and contents was felt to be mysterious. This problem bothered Alexius Meinong (1853-1920) and Kazimierz Twardowski (1866-1938). Twardowski resolved the problem by introducing the distinction between contents and objects of mental acts, particularly presentations. Meinong took Twardowski's claim that every act had its own object seriously, and developed the general theory of objects. Meinong was particularly interested in acts of judging. In order to give a general account of such acts, he introduced the concept of objective as the object of judging. Roughly speaking, objectives are items determined by the content of judgements, independent of the real state of the world. Carl Stumpf (1848-1936), another Brentanist, introduced the term *Sachverhalt* (state of affairs) in order to describe objects of judgments. Later, Stumpf and Twardowski tried to explain objective contents of subjective acts by the distinction of action and their products. Moreover, Brentano's students (Anton Marty (1847-1914) should be mentioned here) analyzed several epistemological concepts and problems. Perhaps the defense of the absoluteness of truth (Marty, Meinong, Stumpf, Twardowski) deserves a special attention, because it was independent of the evidence theory defended by Brentano himself. The philosophy of probability (Marty, Meinong, but also Brentano himself) is another direction of the research of this school worth to be mentioned.

Brentano was a psychologist of a sort, although different than philosophers described in section (C). The considerations of Meinong, Stumpf and Twardowski pushed Brentanism in the direction of logical objectivism, but not radically. It was Edmund Husserl (1859-1938), also Brentano's student, who entirely abandoned psychologism, although his early views belonged to this position. The epoch-making book *Logical Investigations* (1900-1901) offered a critique of psychologism in logic and mathematics. The criticism was similar to that of Frege and inspired by him. In fact, Frege wrote a critical review of Husserl's early book *Philosophy of Arithmetic* (1891), and the criticism mainly concerned the psychologism strongly present in this work. Husserl, like Frege, argued that since logic and mathematics were *a priori,* but psychology *a posteriori,* the former could not be reduced to the latter. Husserl extended his criticism to the whole of epistemology arguing that psychologism inevitably led to relativism and scepticism. Although Husserl himself admitted that Brentano was decisive for his philosophical development, Husserl's phenomenology became a new quality in philosophy, not reducible to Brentanism. However, Brentanian themes became constant in Husserl's thought through its essential evolution. Roughly speaking, throughout his philosophical career Husserl basically investigated one problem: the relation between the subject and the object. Clearly, the concept of intentionality played the central role in this business.

The Husserlian criticism of psychologism implied that epistemology could be naturalistic, in particular psychological. Unfortunately, none of Husserl's works gives a relatively complete account of his epistemology, and we must reconstruct its tenets from various sources. It is obvious that Husserl was a foundationalist. He considered this view as a necessary weapon against scepticism and relativism. Hence, epistemology cannot be superstructured on empirical sciences, like psychology or physiology. Anyway, Husserl did not reject empiricism, but he extended this view. His first version of phenomenology, outlined in *Logical Investigations* claimed that we should describe phenomena without any assumptions taken from science or the ordinary world-view. Thus, phenomenology and, *a fortiori.* epistemology, was conceived as presupossitionless. "To describe things as they present themselves to consciousness" – says Husserl's first principle, termed by him as the principle of all principles. It is highly probable that the idea of presuppositionless epistemology was Husserl's response to Nelson's argument for the impossibility of any theory of knowledge. Thus, Husserl defended the idea of autonomous epistemology, independent of any other discipline, also of ontology. He hoped that this perspective would allow a coherent fusion of empiricism (in his understanding) and apriorism. In particular, he admitted synthetic *a priori* propositions based on the phenomenological experience directed to essences. At first Husserl's phenomenology was realistic. Later, however, he changed this position and came to the conclusion that the problem of intentionality could not be satisfactorily solved in the framework of realism. Thus, he passed to transcendental phenomenology on which he worked to the end of his life. Under this new idea, the world is constructed by the pure transcendental consciousness (transcendental idealism). Husserl made several efforts to explain various details of this construction, especially the problem of intersubjectivity, but he was constantly dissatisfied with the results of his investigations. The last version of his philosophy was based on the concept of the life-world *(Lebenswelt,* the world which is priorly given to people (note, however, that this concept is differently understood by particular interpreters of Husserl's philosophy) and provides a horizon for their consciousness. This concept became the source of still another approach to the sociology of knowledge, proposed by Max Scheler (1874-1928) and Alfred Schütz (1899-1959). Not all phenomenologists accepted Husserl's passing to transcendental phenomenology and idealism. Roman Ingarden (1893-1970) was one of the most important defenders of realistic phenomenology. The influence of phenomenology in contemporary philosophy is enormous. First, phenomenology itself is an extensive movement with many camps, including the thought of Martin Heidegger (1889-1976). Second, phenomenology inspired several other philosophical orientations from existentialism to Marxism and postmodernism. Third, phenomenology raised several fundamental questions. To mention just one: intentionality in the context of cognitive science. It is not by accident that we witness topics like "Husserl and Computer Science".

(E) Relativism, pragmatism, intuitionism and conventionalism

Friedrich Nietzsche (1844-1900) was *l'enfant terrible* of philosophy of the second half of the 19th century. His main interest was in criticism of traditional morality and pedagogy. He proposed a new morality and new ideals based on individualism, inequality and voluntarism; in this last point he was inspired by Schopenhauer. Nietzche's relativism, moral as well as epistemological, was inspired by biology, particularly by Darwin's theory of evolution. Thus, knowledge for Nietzsche is a biological fact, conditioned, like every form of human behaviour, by practical tasks. Due to that, knowledge must be analyzed from the biological point of view and with respect to its practical functions. For Nietzsche, this perspective justified relativism. Knowledge taken from the biological and practical point of view is relative. Absolute and objective truth is a myth, because we never fully conceive reality. On the contrary, we apprehend it falsely for various simplifications made in the process of cognition. In particular, every generalization leads to an inadequacy, relatively to the degree of generality. However, we regard this simplified and false account of the world as true, because it is our creation and the object of our belief. Since we live with myths, we consider them as truths, finally forgetting that they are false. Nietzsche shocked his contemporaries for his radical moral claims which were at odds with standard morality. Thus his general epistemological views were in the shadow of other ingredients of his philosophy.

Pragmatism was the first philosophical novelty to come from the New World. Charles Sanders Peirce (1839-1914) and William James (1842-1910) are the founding fathers of pragmatism. However, though James claimed to be only a popularizer of Peirce, their views were not identical. Peirce questioned the Cartesian view that we could achieve absolutely certain beliefs. However, this situation does not justify scepticism. Thus, methods of fixation of beliefs, even fallible, are of the utmost importance. Peirce wanted to generalize the method of science as the only reliable warrant of beliefs. He found the key idea in the so called pragmatic maxim: in order to make concepts clear, one must know how to apply them. The maxim primarily regards meanings of concepts and propositions, but it also has a methodological dimension: application of the pragmatic maxim in inquiry requires criticism and empirical research. According to his general charge against Descartes, Peirce maintained that empirical science provided only fallible results. It uses abduction which looks for logical antecedents on the base of the given statements assumed to be consequents of something (in more recent terminology, abduction is a kind of induction in the wide sense), and never ends with certain statements. On the other hand, Peirce was an epistemological optimist and believed that any good question could be effectively solved, if our investigation have been performed long enough and according to correct procedures. Peirce defended epistemological realism. His theory of truth is fairly complex. Basically, truth consists in the convergence of opinions. At the first sight, it suggests the consensus theory of truth. However, if we remember that every question has 'its correct answer, Peirce's philosophy admits the correspondence theory of truth in the ideal limit of inquiry.

Peirce did not use the term 'pragmatism' very often. This label became central for James. He also referred to the pragmatic maxim, even to Peirce in this respect, but he changed its meaning. For James, the pragmatic maxim meant that we should look for practical consequences of our actions performed on the basis of beliefs. This point also determined James' theory of truth: true beliefs are those which lead to profitable actions (the utilitarian theory of truth), and this account of truth was relativistic. These points are relevant for James' radical empiricism that resulted with antirealism. Now it is clear that James' pragmatism was different than that of Peirce; as a matter of fact, Peirce himself realized differences and proposed the term 'pragmaticism' for his philosophy. The main differences include: (a) Peirce was a realist, James defended antirealism: (b) James was a radical empiricist, Peirce admitted some *a priori* elements, for example the methodological presuppositions of science; (c) Peirce's pragmatism was operational, so to speak, James developed its utilitarian version; (d) James entirely rejected the correspondence theory of truth, Peirce accepted it as valid in the ideal limit of inquiry; (e) James argued for unlimited relativism, Peirce's position was much more moderate in this respect. Pragmatism became a constant element of the American philosophical scene. John Dewey (1859-1952) joined the founding fathers and is commonly recognized as the third main representative of pragmatism. The ideas of pragmatism are evident in the operationism of William P. Brigdman (1882-1961), a philosopher of physics. Then, the neopragmatists appeared, with Clarence I. Lewis (1883-1964) and Quine (born 1908) as leading figures who, following Peirce, applied exact logical tools in philosophical research. Pragmatism influenced not only philosophy, but also other fields, particularly jurisprudence, becoming the philosophical context for American legal realism. Although this movement was characteristic of the US philosophy, it also appeared in Europe where Ferdinand C. S. Schiller (1864-1937) was its main defender.

Henri Bergson (1859-1941) was the most famous French philosopher at the turn of the 19th and 20th centuries. He developed an intuitionistic epistemology, more precisely, irrational intuitionism. He began with a criticism of intellect. According to Bergson, the competence of intellect is limited to science where analysis is a proper method. However, intellect fails in the case of the immediate stream of consciousness data. It cannot be analyzed, because every analysis simplifies its object, makes it static, stratifies it into parts, puts it into quantitative and mechanical categories, always considers things relatively to a perspective, and is indirect, being mediated by a symbolism. Another method, namely one appealing to intuition, must be executed in order to apprehend the actual data. Intuition works in a manner entirely opposite to analysis. In particular, intuition catches its object momentarily and in its integrity, allows to perception of its dynamics, is independent of specific perspectives, is qualitative, direct and essence oriented. Bergson did not deny that science produced a reliable picture of the world, relative to its inherent limitations and practical tasks of people. However, if we aspire to have true knowledge of reality, we need to go beyond the scientific perspective and appeal to intuitive knowledge. Bergson's popularity at the beginning of the 20th century was not confirmed by the later influence of his philosophy. Bergson being awarded with the Nobel prize and being active in the League of Nations did not initiate Bergsonism as a movement.

Henri Poincaré (1854-1912), Pierre Duhem (1861-1916) and Eduard Le Roy (1870-1954) pointed out that assertions of propositions presupposed conventions. Hence, this kind of philosophy is called conventionalism. Poincaré's conventionalism was motivated by his work in mathematics and mathematical physics. He asked for the grounds of axioms of mathematical and physical theories, for example, the parallel postulate in Euclidean geometry or the second law of Newtonian classical mechanics (force is equal to mass times acceleration). On the traditional account, they were theoretical descriptions of facts. Poincaré challenged this view and argued for the dependence of theoretical axioms on already accepted stipulations. Such conventions are adopted for various reasons, for example simplicity, economy, elegance, etc., but not because they are realistically true. Hence, axioms are implicit definitions of the concepts involved. Duhem derived similar views from the history of science. Moreover, he developed a holistic account of physical theories, which was important for a new look at testing in science. Traditionally, empirical confirmation concerned laws, hypotheses and assertions about singular facts as possibly separate units of testing. For Duhem, testing is never performed on isolated elements of science, but it is always directed to whole theories. Duhem concluded that no theory could be conclusively tested, positively or negatively. This had crucial consequences for the problem of the so called *experimentum crucis* (crucial experiment, that is, deciding between competing hypotheses). Before Duhem, it was commonly recognized that a positive *experimentum crucis,* i. e., an experiment allowing to select a true theory from a variety of mutually contradictory rivals, was impossible. However, negative *experimenta crucis,* that is, procedures rejecting a false theory, was regarded as possible. Duhem's famous thesis (nowadays called the Duhem-Quine thesis, because Quine renewed holistic epistemology) says that also a negative *experimentum crucis* is impossible. Le Roy, who studied with Bergson, developed still another form of conventionalism. He argued for the necessity of conventions appealing to Bergson's view that intellect was insufficient for cognitive views. Since rational procedures must be supplemented by acts of faith, conventions are unavoidable. A very radical version of conventionalism was proposed by Kazimierz Ajdukiewicz (1890-1963), in the 30's. Ajdukiewicz stressed the role of language in accepting propositions. He extended Poincaré's conventionalism to observation sentences. According to Ajdukiewicz, we always assert propositions in a definite language. Hence, we can change a language instead of rejecting propositions threatened by facts. This situation concerns not only theoretical principles as in Poincare's case, but also empirical reports.

It might be surprising that this section collects together very different views. However, relativism, pragmatism, Bergsonism and conventionalism destroyed certain well-established views, above all the traditional account of facts. All the views reported in this section stress that there are no brute facts, independent of our cognitive faculties, expectations, needs, theoretical frameworks, language, symbolism, etc. Perhaps Ajdukiewicz expressed these ideas in the strongest way. According to him, the set of concepts of a given language (I omit some additional clauses imposed by Ajdukiewicz on languages adequate for science) forms the so called conceptual apparatus. Ajdukiewicz's main thesis is: the world-picture essentially depends on an adopted conceptual apparatus and is not uniquely determined by experience. There are several problems connected with such views. Do they inevitably lead to relativism or antirealism? Is the classical or correspondence theory of truth consistent with conventionalism, etc? These questions belong to the heart of recent epistemology and we should remember that they go back to views summarized in the present section, similarly as the fallibilism explicit in Peirce.

(F) Early British analytic philosophy

Analytic philosophy arose in Great Britain as a protest against British Neo-Hegelianism. Bertrand Russell (1872-1970) and George E. Moore (1873-1958) were the architects of this movement. Then, Ludwig Wittgenstein (1889-1951) and Frank P. Ramsey (1903-1930) became prominent representatives of analytic school. However, Russell and Moore had different inspirations: Russell found it in mathematical logic (similarly, Wittgenstein and Ramsey), Moore in common sense. Russell changed his views several times and, in spite of his numerous writings, it is difficult to find a systematic exposition of his philosophy, in particular, epistemology. To be sure, he prepared an extensive manuscript *The Theory of Knowledge* (1913), but, due to Wittgenstein's criticism, the book was not published until 1983. It is not sure whether Russell was inclined to hold all the views expressed in this work. Russell's first epistemological interests concerned the theory of truth. He opposed very strongly the views of Bradley and James. For Russell, any correct theory of truth must satisfy three general conditions: it must (a) also be the theory of falsehood; (b) consider the truth of a judgement relationally, that is, define truth *via* a relation of judgements to something else; (c) distinguish truth and its criteria. These conditions forced the rejection of coherentism and pragmatism and motivated the correspondence theory. Roughly speaking, the Russellian theory defined truth as the correspondence of a proposition to a fact, and the correspondence relation was conceived as a structural similarity between propositions and facts. Russell, also inspired by Wittgenstein, developed a view called logical atomism. Basically it was an ontological theory on which simple particular objects formed the ultimate furniture of the universe, and the rest of the ontological inventory could be logically constructed over the simples. This ontological theory was supplemented by epistemology based on the distinction between knowledge by acquaintance and knowledge by description. The simples are accessible by acquaintance, that is, direct sensory experience. However, the constructed items are knowable by description. Besides Russell claimed that knowledge by description is always reducible to knowledge by acquaintance. This view was parallel to Russell's famous theory of descriptions. It illustrates well the connection of logic and philosophy in Russell. These two views are perhaps the most important in Russell's epistemology. He also discussed several other problems, but his ideas did not influence the further course of philosophy like his theory of truth and his distinction of two kinds of knowledge. In order to complete this report on Russell, I only mention that he defended a more or less radical empiricism and, in the last period of his philosophical development, considered knowledge as a biological process.

Moore's philosophical method was different than that of Russell, less based on formal logical constructions, and more directed to a very careful analysis of philosophical concepts, problems and theories. Moreover, as I already noted, he defended common sense as a source of principally correct insights. He argued that idealism confused perceiving things and the content of perception, and his proof of the existence of the external world directed against Berkeley was based on this observation. Moore also criticized Kant for deriving *a priori* from the properties of human mind and Bradley for confusing identity and difference. Thus, Moore rejected all forms of idealism. He offered a very detailed analysis of perception, a topic rather neglected in the 19th century. He introduced the convenient term 'sense-datum' and investigated the relation of sense-data to qualities of things. Moore offered the standard distinction of positions in the philosophy of perception; (a) direct realism; (b) indirect realism, and (c) phenomenalism; (a) was his favourite view, but he was not sure about its correctness. In general, Moore defended realism, the correspondence theory of truth, a moderate empiricism, and, like Russell, the classical account of knowledge as true justified belief. His ethical theory was very influential. Moore regarded goodness as a simple non-natural property, recognized by intuition. According to Moore, every attempt of defining goodness by natural properties has to fail, because it falls into the naturalistic fallacy. This view became decisive for subsequent discussions about epistemological aspects of axiology.

Tractatus logico-philosophicus (1922), Wittgenstein's *opus magnum* is basically an ontological treatise. However, it also contains several epistemological views. Since the limits of the language are, according to Wittgenstein, also the limits of the world, knowledge is closely related to language: there is no knowledge outside the language. Wittgenstein considered propositions as pictures of facts and the correspondence theory of truth became a simple consequence of this view. Moreover, propositions display or show their sense, but it is impossible to communicate it. We encounter here the problem of metaphysics. For Wittgenstein, metaphysics, that is, any attempt to answer questions about the relation between language and the world is nonsensical and unspeakably. He was consequent and identified most of his own propositions as meaningless. *Tractatus* is a difficult and cryptic book and has many conflicting interpretations, for instance, a Kantian one, in which language plays in Wittgenstein a similar role as mind in Kant. Independently of a correct, if possible, interpretation of Wittgenstein's early views, their influence was enormous, particularly in the rise and development of logical empiricism.

Ramsey's life was short, but his achievements are remarkable. He successfully worked in mathematical logic, pure mathematics, economics and philosophy. In epistemology, he advanced two important ideas. First, he elaborated the redundancy theory of truth. This theory consists in the view that the predicate 'is true' is redundant. Ramsey justified this view by pointing out that the equivalence '*A* is true if and only if *A*' motivates the redundancy of 'is true'. It adds nothing to the content of *A*. In fact, if one asserts a proposition *A*, that is, seriously uses it, he or she is inclined to express this attitude in some circumstances by saying '*A* is true', but saying *A* is quite sufficient. More complicated cases are analyzable with the help of quantifiers, for example, the context 'all propositions in that book are true' is analyzed as 'for any *A*, if *A* is a proposition occurring in that book, then *A*'. This simple conception became fairly influential. Ramsey's second important epistemological view concerned probability and belief. He developed the subjective theory of probability and the idea of degrees of belief as measured by actions of a sort.

(G) Logical empiricism and related views

This movement grouped several philosophers, in general, positivistically oriented (the 3rd positivism). The centre was located in Vienna (hence the name 'the Vienna Circle'), another branch was active in Berlin. Important ideas came from Poland (the Lvov-Warsaw school). Moreover, several single philosophers were more or less related to logical empiricism, for example Eino Kaila (1890-1958) in Finland, Karl Raimund Popper (1910-1989) in Vienna, later in London, and Alfred Ayer (1902-1994). The Vienna Circle was established by Moritz Schlick (1882-1936), other prominent members include Rudolf Carnap (1891-1970) and Otto Neurath (1882-1945). Hans Reichenbach (1891-1953) was the main figure in Berlin. The Lvov-Warsaw School was established by Twardowski at the beginning of the 20th century. Ajdukiewicz, already mentioned as a radical conventionalist, and Alfred Tarski (1901-1983) were particularly important for epistemology in Poland; let me note that I indicate only those philosophers whose views will be mentioned in this section. I would like to stress, even very strongly, that the title of this section "Logical empiricism and related views" should be taken literally. I do not suggest that Polish analytic philosophy or Popper belonged to logical empiricism in its classical version. What I do in this section consists rather in grouping together a variety of views that are interrelated by a proximity of ideas and mutual influences. In general, logical empiricists proper and their philosophical relatives were strongly influenced by logic and modern science, particularly physics. Hence, they intended to create a scientific philosophy.

The Vienna Circle was radically anti-metaphysical. Logical empiricists, influenced by Wittgenstein and physical discoveries in relativity theory and quantum mechanics, which suggested that concepts like simultaneity or location were meaningless unless effectively measurable. Thus, the Vienna Circle defined metaphysics as consisting of pseudo-sentences, and tried to show that the lack of meaning can always be demonstrably shown by analysis. A pseudo-sentence is a sentence in the grammatical sense, but it is neither analytic nor verifiable by empirical means (I neglect here various refinements of this principle, called the principle of verifiability). Metaphysical proclamations are meaningless and, thereby, devoid of sense. Logical empiricists argued that this view was directly derived from the logical analysis of language. They reduced philosophy to the logic of science which was identified with logical syntax at the beginning of the development of this movement. Hence, traditional problems of philosophy were declared meaningless. The same concerned epistemology of which only a few problems, for instance, the genesis of knowledge, were admitted as legitimate. However, a closer analysis shows that the logical guise of philosophy taken by logical empiricism was misleading. In fact, this movement elaborated a definite collection of epistemological views. Logical empiricism evolved from a radical position to a much more modest one. In general, it was caused by admitting not only syntactic, but also semantic tools of philosophical analysis.

Let me list the principal epistemological views of logical empiricism, and their evolution. Genetic empiricism was a commonly accepted view among logical empiricists. However, it was rather Humean than Millian. It was even more evident in relation to the problem of the debate about apriorism and aposteriorism. Logical empiricists identified analycity, aprioricity and necessity of sentences. This proposal is known as the linguistic theory of the *a priori,* and it is very close to Hume. Thus, logical empiricism offered a version of moderate aposteriorism: logic and mathematics were regarded as analytic, *a priori* and necessary, and the *a priori* knowledge was restricted to them. Some logical empiricists, like Reichenbach and the early Carnap, included certain Neo-Kantian features into their philosophy. Schlick and Carnap were foundationalists. They believed that we had a priviledged access to elementary (protocol) sentences that ascribed empirically knowable properties to concreta. On the other hand, Neurath defended anti-foundationalism; he introduced the metaphor of boat mentioned in the Introduction above. However, foundationalists and anti-foundationalists grouped in the Vienna Circle accepted physicalism, namely the view that the language of physics was proper for science, including psychology and the humanities. Thus, they revived naturalism in the philosophy of psychology and humanities. Genetic epistemology also resulted with phenomenalism in the philosophy of perception. This view was defended particularly by Ayer.

In his *Theory of Knowledge* (1918), Schlick defended the correspondence theory of truth. However, he distinguished correspondence as similarity or sameness and correspondence as correlation, accepting the latter understanding of the correspondence relation. For Neurath, the correspondence theory of truth was completely meaningless, because it led to metaphysics. Neurath himself developed a version of the coherence theory of truth. Carnap tried to eliminate the concept of truth in favour of syntactic notions in his famous *Logical Syntax of Language* (1934). In 1933, Tarski published his famous treatise on the concept of truth in formalized language which became a turning point in the development of contemporary analytic philosophy. Tarski's semantic theory of truth was accepted by Carnap who abandoned his earlier syntactic point of view. Only Neurath preserved his coherentist view. Since Tarski's theory is discussed by Marian David in this volume, I will not enter into further discussion about the semantic definition of truth.

As far as the matter concerns the issue of realism, the early logical empiricism regarded it as a pseudo-problem. The situation changed, largely due to Tarski's influence. In fact, the semantic approach to truth suggested realism. This view was accepted by Kaila and Popper. The latter extended realism to critical rationalism and logical objectivism (the concept of the third world). Carnap chose a compromise between traditional realism and principles of logical empiricism. He rejected realism as a view asserting transcendence of reality, but admitted that we had reasons to ask for existence matters, relative to a given linguistic framework. This was a sort of internal realism, using more recent terminology. Realism was also defended by Ajdukiewicz who proposed semantic epistemology and tried to demonstrate that semantics provided good arguments against idealism. In particular, Ajdukiewicz argued that epistemology should use a semantic language, because it was the proper language for analyzing the relation between cognitive acts and their objects. According to Ajdukiewicz, idealists employed a language which was similar to syntactic speech. Since, due to results of formal semantics, semantical properties of rich languages cannot be defined inside syntax, idealism is defective from the beginning.

(H) Later analytic philosophy

Wittgenstein radically changed his earlier views. In his *Philosophical Investigations,* published posthumously in 1953, he rejected the idea of a perfect language governed by strict logical rules. Instead, Wittgenstein recommended ordinary talk and developed a new approach to meaning based on the idea of language games. Language consists of various, mutually irreducible language games to which meanings must be relativized. In general, meaning of expressions is displayed by their uses in real situations. Hence, philosophers can do their best by referring to concrete applications of words, including epistemologically important contexts, like 'to know', 'to believe', 'to see', etc. Rejecting philosophical reconstructions *via* formal logical tools, Wittgenstein agreed people's behaviour was related to rules. He accepted so called rule-following, but he did not understand rules as abstract patterns serving as guides to evaluate actions as correct or not. Rules, according to Wittgenstein, are individual events, conventional and learnable, particularly useful in linguistic communications. This also concerns alleged rules related to epistemic activities. Since Wittgenstein in his second period did not intend to create a philosophical system, it is difficult to rectify concrete epistemological views from his fragmentary remarks. Perhaps the most important is his argument against the possibility of private language, because such a language would make communication impossible. This argument is sometimes used against idealism. Wittgenstein was also influenced by *Gestalt* psychology in his remarks about perception.

Wittgenstein's way of doing philosophy favoured descriptive methods, similar to that praticized by Moore. This attitude, directed against formalism of logical empiricism, attracted many philosophers, partly due to Moore's and Wittgenstein's influence, and partly to a fairly general disappointment with positivism after World War II. Gilbert Ryle (1900-1976), another practitioner of informal analysis, was intended to explain the difference between theory and praxis. Ryle was influential in Oxford, but it was John L. Austin who (1911-1960) established 'ordinary language philosophy', also called Oxonian philosophy. Austin practiced philosophical linguistics or linguistic phenomenology, which completely trusted ordinary language. According to Austin, all important philosophical concepts and distinctions were present in ordinary language, which is basically correct. Hence, the main philosophical task consists in careful analysis aiming at exhibiting the conceptual machinery of everyday speech. In epistemology, Austin defended direct realism and the correspondence theory of truth.

5. FINAL REMARKS

I would like to repeat once more that my report about the development of epistemology is incomplete. Although I did not restrict myself to analytic tradition, I am fully conscious that I neglected several topics and persons important for other styles of philosophizing. Sören Kierkegaard, new French rationalists, like Emilé Meyerson, Lèon Brunschvicg or Ferdinard Gonseth, Karl Jaspers, Jean-Paul Sartre, Jean Piaget, Martin Buber, Emmanuel Lévinas, Hans-Georg Gadamer and many others certainly deserve the attention of historians of epistemology. One can also complain that Martin Heidegger was mentioned only once. Of course, this list could be much longer. However, I hope that my survey will help readers of this book to better understand of epistemology itself. We can see that several problems and solutions are recurrent through the development of our subject. It is, as it always was, difficult to predict the further development of epistemology. Will it be organized around traditional views, like rationalism, empiricism, etc. or become a loose collection of concrete questions? Will it be consumed by cognitive science or preserve its philosophical character and independence of special fields? Who knows?

Jan Woleński
The Jagiellonian University

REFERENCES

Arndt, E.: 1908 *Das Verhältnis der Verstanderkenntnis zur sinnlichen in der vorsokratischen Philosophie*, Max Niemeyer, Halle; repr., Georg Olms, 1975.

Beare, J. L.: 1906, *Greek Theories of Elementary Cognition from Alemaeon to Aristotle*, Clarendon Press, Oxford; repr. Thoemmes Press, Bristol, 1992.

Bennett, J.: 1972 *Locke, Berkeley, Hume – Central Themes*, Clarendon Press, Oxford.

Denyer, N.: 1991, *Language, Truth and Falsehood in Ancient Greek Philosophy*. Routledge, London.

Dürr, K.: 1923, *Wesen und Geschichte der Erkenntnistheorie*, Orell Füsch, Zürich 1923.

Everson, S. (ed.): 1990, *Epistemology Companions to Ancient Thought 1*, Cambridge University Press, Cambridge.

Fleischer, M.: 1984, *Wahrheit und Wahrheitsgrund, Zur Wahrheitsproblem und zu seiner Geschichte*, De Gruyter, Berlin.

Freytag, Willy: 1905, *Die Entwicklung des griechischen Epistemologie bis Aristoteles*, Verlag von Max Niemayer, Halle.

Goedeckemeyer, Albert: 1905, *Die Geschichte der griechischen Skeptizismus*, Dieterich'sche Verlagbuchhandlung, Leipzig.

Groarke, L.: 1990, *Greek Scepticism*, McGill University Press, Montreal.

Hankinson, R. J.: 1995, *The Sceptics*, Routledge, London.

Herbertz, R.: *Das Wahrheitsproblem in der griechschen Philosophie*, Reimer, Berlin.

Hoven, A.: 1989, *Wegen zur Wahrheit: Eine typologische Studie über Wahrheitstheorien*, Peter Lang, Bern.

Hönigswald, R.: 1933, *Geschichte der Erkenntnistheorie*, Junker und Dünhaupt Verlag, Berlin.

Jansen, B.: *Die Geschichte der Erkenntnislehre in der neueren Philosophie bis Kant*, Schönings, Paderborn

Kramer, S.: 1982, *Berechenbare Vernüft. Kalkül und Rationalismus im 17. Jahrhundert*, De Gruyter, Berlin.

Kynast, R.: 1930, *Logik und Erkenntnistheorie der Gegenwart*, Junker und Dunnhaupt Verlag, Bedriin.

Law, J. D.: 1993, *The Rhetoric of Empiricism Language and Perception from Locke to I. A. Richards*, Cornell University Press, Ithaca.

Levet, J.-P.: 1976, *Le vrai et le faux dans la pensée grecque archaïque Étude de vocabulaire I Présentation générale Le vrai et le faux dans les épopées homériques*, Le Belles Lettres, Paris.

Natorp, P.: 1884, *Forschungen zur Geschichte des Erkenntnisproblems im Altertum Protagoras, Demokrit, Epikur und die Skepsis*, Wilhelm Hertz, Berlin.

Pasnau, R.: 1997, *Theories of Cognition in the Later Middle Ages*, Cambridge University Press, Cambridge.

Perler, D.: 1992, *Der propositionale Wahrheitsbegriff im 14. Jahrhundert*, De Gruyter, Berlin.

Rath, M.: 1994, *Der Psychologismusstreit in der deutschen Philosophie*, Karl Alber, Freiburg.

Richter, R.: 1908, *Der Skepticismus in der Philosophie und seine Überwindung*, 2 vis., Verlag der Durr'schcn Buchhandlung, Leipzig.

Riehl, A.: 1924-25, *Der Philosophische Kritizismus*, 3 vls., Kroner, Leipzig.

Stekeler-Weithofer, P.: *Sinnkriterien Die logischen Grundlagen kritischer Philosophie von Plato bis Wittgenstein*, Ferdinand Schoningh, Paderborn.

von Aster, E.: 1921, *Geschichte der neueren Erkenntnistheorie (Von Descartes bis Hegel)*, Walter de Gruyter, Berlin und Leipzig.

Walker, R. C. S.: 1989, *The Coherence Theory of Truth Realism, Anti-Realism, Idealism*, Routledge, London.

Yolton, J. W.: 1984, *Perceptual Acquaintance from Descartes to Reid*, University of Minnesota Press, Minneapolis.

Yolton, J. W.: 1996, *Perception and Reality. A History from Decartes to Kant*, Cornell University Press, Ithaca.

PART I: SOURCES OF KNOWLEDGE AND BELIEF

ROBERT AUDI

PERCEPTION AND CONSCIOUSNESS

In very general terms, perception is a response to the world. The paradigm cases of it are responses by the five senses: we see, hear, touch, taste, taste, and smell. But we also have an awareness of states of our own body, such as the position and movement of our limbs, and that awareness is at once similar in character to perception yet not dependent on the five senses. There is a third kind of awareness, one that is distinct, at least conceptually, from our awareness of our bodily condition and movements; its object is our own mental states. The first – ordinary perception – has been called exteroception ("outer perception"), the second interoception ("inner perception") or, in a special case, proprioception, though taking this term generically in the sense of 'self-perception' we might conveniently use it to designate the third case, in which the object of awareness is mental. All three are important for this study, particularly the first and third. Under the more general rubrics of perception and introspection (or self-consciousness), these are perennially basic topics in epistemology, construed as the theory of knowledge and justification.

Perception is also important in the philosophy of mind, and what follows will often explore it from that point of view. In the main, however, my task is to clarify the nature of perception, outer and inner, and detail its role in grounding knowledge and justification. This requires connecting perception with such psychological concepts as those of sensation and belief, as well as explaining how it depends on causal connections to the external world if it is to yield knowledge thereof. Part I concerns perception through the senses. Part II addresses self-perception: roughly, perception of oneself as it occurs in introspection.

I PERCEPTION OF THE EXTERNAL WORLD

The five senses may be viewed as corresponding to *modes* of perception. Seeing is perceiving in the visual mode, hearing is perceiving in the auditory mode, and so forth. A major question for both epistemology and the philosophy of mind is whether perception is always accompanied by some kind of *cognitive uptake*, paradigmatically the formation of some belief about the object perceived. It may appear that it could not otherwise ground knowledge, since it seems clear that perception is a source of knowledge and justification mainly by virtue of yielding beliefs that constitute knowledge or are justified. This conclusion would be at best premature. Even a good foundation need not have anything built on it. Let us first consider in some detail what perception is and then proceed to explore its relation to belief and its epistemological role.

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 57-108.
© 2004 *Kluwer Academic Publishers.*

1.1 The Elements and Basic Kinds of Perception

There are apparently at least four elements in perception, all evident in a simple case like seeing a green field in front of me: (1) the perceiver, me; (2) the object, the field; (3) the sensory experience, my visual experience of colors and shapes; and (4) the relation between the object and the subject, commonly considered a causal relation by which the object produces the sensory experience in the perceiver. To see the field is apparently at least this: to have a certain sensory experience as a result of the impact of the field on one's organs of vision.

Some accounts of perception add to the four items on this list; others subtract. We must consider both kinds of account and how these elements are to be conceived in relation to one another. First, however, we should explore some examples of perception and several perceptual locutions.

There are three quite different ways to speak of perception. Each corresponds to a different way of perceptually responding to experience. We often speak simply of what people perceive, for instance of what they see. We also speak of what they perceive the object in question to be, and we commonly talk of what they perceive in or about it. Let us start with visual perception. I see, hence perceive, a green field. Secondly, speaking in a less familiar way, I also see it *to be* rectangular. Thus, I might say that from the air one can see it to be perfectly rectangular. Thirdly, I see *that* it is rectangular. Perception – in this case seeing – is common to all three cases.

The first case is *simple perception,* perception taken by itself (here, visual perception). I simply see the field, and this experience is the visual parallel of hearing a bird (an auditory experience), touching a glass (a tactual experience), etc. If the first case is simply a *perceiving of* some object, the second is a case of *perceiving to be,* since it is seeing something to be so: I don't just see the field, as where I fly overhead at high speed; I see it to be rectangular. The third case is one of *perceiving that,* since it is seeing that a particular thing is so, namely, that the field is rectangular. These cases represent three kinds (or at least cases) of perception. Perception of the simplest kind, such as seeing, occurs in all three cases; but, especially because of their relation to knowledge and justified belief, they differ significantly. We can best understand these three kinds of perception if we first focus on their relation to belief.

Perceptual belief

The latter two cases – perceiving that, and perceiving to be – differ from the first – perceiving of – in implying corresponding kinds of beliefs: seeing that the field is rectangular implies believing that it is, and seeing it to be a green field implies believing it to be a green one. If we consider how both kinds of beliefs – beliefs *that* and beliefs *of* (or *about*) are related to perception, we can begin to understand how perception occurs in all three cases. In my second and third examples of perception, my visual perception issues in beliefs that are then grounded in it and can thereby constitute visual knowledge.

In the first example, that of simple perception, my just seeing the field provides a basis for both kinds of beliefs. It does this even if, because my mind is entirely occupied with what is on the radio as I glance over the field, no belief about the field actually arises in me. The visual experience is in this instance like a foundation that has nothing built on it but is ready to support a structure. If, for instance, someone were to ask if the field has shrubbery in it, then given the lilacs prominent in one place, I might immediately form the belief that it does and say so. This belief is visually grounded; it comes *from* my seeing the field, though it did not initially come *with* it.

When beliefs do arise from visual experiences, as is usual, how are they specifically perceptual? Many of my beliefs arising through perception correspond to perception that, say to seeing that something is so. I believe, for instance, that the field is lighter green toward its borders where it gets less sun. But one might also have various beliefs of the second kind: they correspond to perceiving to be, for instance to seeing something to be a certain color. Thus, one might believe the field to be green, to be symmetrical, to be rectangular, and so on. The difference between these two kinds of belief is significant. It corresponds both to two distinct ways we are related to the objects we perceive and, secondly, to two different ways of assessing the truth of what, on the basis of our perceptions, we believe.

The first kind of belief just described is *propositional,* since it is believing a proposition – say, *that* the field is rectangular. The belief is thus true or false depending on the truth value of that proposition. In holding the belief, moreover, in some way I think of what I see as a *field* which is rectangular; I conceive what I take to be rectangular *as* a field. The second kind of belief might be called *objectual:* it is a belief regarding an object, say the field, with which the belief is actually connected. This is an object *of* (or about) which I believe something, say that it is rectangular.

If I believe *the field* to be rectangular, there really is such an object, and I have a certain relation to it. A special feature of this relation is that I can stand in it without there being any particular proposition I must believe about the field. To see that there is no particular proposition, notice that in holding this objectual belief I need not think of what I see *as* a field; for I might mistakenly take it to be a huge canvass or a grasslike artificial turf, yet still believe it to be rectangular. I might think of it just in terms of what I believe it to be and not in terms of what it obviously is. Thus, although there is *some* property I must take it to have – corresponding to what I believe it to be – there is no other particular way I must think of it. Thus, my perceptual experience need supply no particular notion that must yield the subject of any proposition I believe: I do not have to believe that the *field* is green, that the *grass* is, or any such thing. Perception leaves us vast latitude as to what we learn from it. People differ markedly in the beliefs they form about the very same things they each clearly see.

A related way to see the difference between objectual and propositional beliefs is this. If I believe something to have a property (to be such-and-such), say I believe a British Airways plane to be a Boeing 747, this same belief can be ascribed to me using any correct description of that plane, say as the most travelled plane in their fleet: to say I believe their most travelled plane to be a 747 is to ascribe the same belief to me. This holds even if I do not believe the plane meets that description – and it can hold even where I cannot understand the description, as a child who believes a tachistoscope to be making noise cannot understand 'tachistoscope'. By contrast, if I have a propositional belief, say that the United Airlines plane on the runway is the most travelled in its fleet, this ascription cannot be truly made using just any correct description of that plane, say the plane on which a baby was delivered on Christmas Day, 1995. I may have no inkling of that surprising fact. In a different terminology, the position of ‘*x*’ in locutions of the form of ‘*S* believes *x* to be *F*’, where ‘*S*’ ranges over persons and ‘*F*’ over perceptible properties, is *transparent with respect to substitution* (as it is with respect to quantification); whereas its position in locutions of the form of ‘*S* believes that *x* is *F*’ is *opaque with respect to substitution* (and also quantification, since believing that *x* is *F* does not entail that there *is* anything one takes to be *F*). A rough way to put part of my point is to say that propositional beliefs about things are about them under a description or name, and objectual beliefs about things are not (even if the believer could describe them in terms of a property they are believed to have, such as being noisy). It is in part because we need not conceptualize things – as by thinking of them under a description – in order to have objectual beliefs about them that those beliefs are apparently more basic than propositional ones.

The concept of objectual perception, then, is very permissive about what propositions one believes about the object perceived. This is one reason why it leaves so much space for imagination and learning – a space often filled by the formation of propositional beliefs, each capturing a different aspect of what is perceived, say that the field is richly green and that it ends at a line of trees. Take a different example. Suppose I see a distant flare but do not take it to be what it is; after coming to believe, of this thing that looks blurry and far away, that it glowed, I might ask, “What on earth was it that glowed?” Before I can believe the proposition that a flare glowed, I may have to think about where I am, the movement and fading of the glow, and so forth. The objectual belief provides a guide by which I may arrive at propositional beliefs and propositional knowledge.

Perception, conception, and belief

The same kind of example can be used to illustrate how belief depends on our conceptual resources in a way (simple) perception does not. Suppose I had grown up in the desert and somehow failed to acquire the concept of a field. I could certainly still see the green field, and the intrinsic character of my visual experience would presumably be the same as it is now; from a purely visual point of view the field might look to me just as it does now. I could also still believe, regarding the field I see ▯ and perhaps conceive as sand artificially covered with something green – that it is rectangular. But I could not believe that *the field* is rectangular. This propositional belief as it were *portrays* what I see *as* a field in a way that requires my having a concept of one. If I believe that the field is rectangular, I should be able to say that it is and to know what I am talking about. But if I had no concept of a field, then in saying this I would not know what I am talking about.[1]

Similarly, a two-year-old, say Susie, who has no notion of a tachistoscope, can, upon seeing one and hearing it work, believe it to be making noise; but she cannot believe specifically that the tachistoscope is making noise. Her propositional belief, if any, would be, say, that the thing on the table is making noise. Since, this is true, what she believes is true and she may know this truth; but she need not know much about the object this truth concerns: in a way, she does not know what it is she has this true belief *about*. Still, her sensory experience could be qualitatively the same as that of an adult who has the relevant concept. This possibility bears on how perception figures in language learning and in translating from one language to another. Translation would plainly be at best more difficult if we could not assume similarities among people in perceptual experience.[2]

The general lesson here is important. A basic way we learn about objects is to find out truths about them in this elementary fashion: we get a handle on them through perception; we form objectual (and other) beliefs about them from different perspectives; and (often) we finally reach an adequate concept of what they are. From the properties I believe the distant flare to have (e.g., glowing and slowly falling), I finally figure out that it is a flare that has those properties. As this suggests, there is at least one respect in which our knowledge of (perceptible) properties is more basic than our knowledge of the substances that have them. It is in part because of this order of comprehensibility that phenomenalism (which is discussed in some detail below) is as appealing as it is.

Unlike propositional beliefs, objectual beliefs have a significant degree of indefiniteness and so are best not viewed as true without qualification; they are accurate or inaccurate, depending on whether what one believes of the object (such as that it is rectangular) is or is not *true of* it. Recall Susie. If she attributes noise-making to the tachistoscope, she truly believes, *of* it, that it is making noise. She is, then *right about it.* But if we say unqualifiedly that her belief about it is true, we invite the question 'What belief?' and the expectation that the answer will specify a particular proposition. We can be right about something without knowing or even having any notion of what kind of thing it is that we are right about. Knowledge is often partial in this way. Still, once we get such an epistemic handle on something we can usually use that to learn more about it.[3]

Corresponding to the two kinds of beliefs I have described are two ways of talking about perception. I see *that* the field is rectangular. This is (visual) *propositional perception:* perceiving that. I also see it *to be* rectangular. This is (visual) *objectual perception:* perceiving to be. The same distinction apparently applies to hearing and touch. Perhaps, for example, I can hear that a piano is out of tune by hearing its sour notes, as opposed to hearing the tuner say it needs tuning. As for taste and smell, we speak as if they yielded only simple perception: we talk of smelling mint in the iced tea, but not of smelling that it is minty or smelling it to be minty. Such talk is, however, quite intelligible on the model of seeing that something is so or seeing it to be so, and we may thus take the distinction between perceiving *that* and perceiving *to be* to apply in principle to all the senses.

In brief terms, propositional perception entails both conceptualization of the object perceived and of some property it is perceived to have, whereas objectual perception entails conceptualization only of the latter – except insofar as taking something to have a property *is* conceptualizing it as something having *that* property. There is a conceptual openness about objectual perception that is not present in the propositional case; that conceptually open space can be filled in indefinitely many ways.[4]

It is useful to think of perceptual beliefs as *embedded* in the corresponding propositional or objectual perception, roughly in the sense that they are integrally tied to perceiving of that kind and derive their character and authority from their perceptual grounding. Thus, my belief that the field is rectangular is embedded in my seeing that it is. This kind of perception might be called *cognitive*, since belief is a cognitive attitude: roughly the kind having a proposition (something true or false) as its object.[5]

Both propositional and objectual beliefs are grounded in simple perception: if I don't see a thing at all, I don't see *that* it has any particular property and I don't see it *to be* anything. Depending on whether perceptual beliefs are embedded in propositional or objectual perception, they may differ in the kind of knowledge they yield. Propositional perception yields knowledge both of *what* it is that we perceive, and of some *property* of it, for instance of the field's being rectangular. Objectual perception may, in special cases, give us knowledge only of a property of what we perceive, say that it is green, when we do not know what it is or have any belief as to what it is. In objectual perception, we are, to be sure, in a good position to come to know *something* or other about the object, say that it is a green expanse. Objectual perception may thus give us information not only about objects of which we have a definite conception, such as familiar things in a home, but also about utterly unfamiliar, unconceptualized objects or about objects of which we have only a very general conception, say "that noisy thing". This is important. We could not learn as readily from perception if it gave us information only about objects we conceive in the specific ways in which we conceive most of the familiar things we see, hear, touch, taste, and smell.[6]

I.2 SEEING AND BELIEVING

Both propositional and objectual perceptual beliefs are quite commonly grounded in perception in a way that apparently connects us with the real, outside world and assures their truth. For instance, my visual belief that the field is rectangular is so grounded in my seeing the field that I truly see that it is rectangular. Admittedly, I might visually (or tactually) believe that something is rectangular under conditions poor for judging it, as where I view a straight stick half submerged in water (it would look bent whether it is or not). My visually grounded belief might then be mistaken. But such a mistaken belief is not *embedded* in the propositional perception that the stick is bent – something one does not see is so, since it is false. The belief is merely produced by some element in the simple perception of the stick: I see the stick in the water, and the operation of reflected light causes the illusion of a bent stick. I thus do not see that the stick is bent: my genuine perception is of it, but not of its curvature.

As this suggests, there is something special about both perceiving *that* and perceiving *to be.* They are *veridical experiences,* i.e., they imply truth. Thus, when I simply see the rectangularity of the field, if I acquire the corresponding embedded perceptual beliefs – if I believe that it is rectangular when I see that it is, or believe it to be rectangular when I see it to be – then I am correct in so believing. If perceiving *that* and perceiving *to be* imply (truly) believing something about the object perceived, does simple perception – perception *of* something – which is required for either of these more complex kinds of perception, also imply true belief? Very commonly, simple perception does imply truly believing something about the object perceived. But could I not hear a car go by yet be so occupied with my reading that I form no belief about it? Let us explore this.

Perception as a source of dispositions to believe

As is suggested by the case of perception overshadowed by preoccupation with reading, there is reason to doubt that simple perceiving *must* produce belief. This may seem to fly in the face of the adage that seeing is believing. But properly understood, that may apply just to propositional or objectual seeing. There perception does produce beliefs. Seeing that golfball-size hail is falling is believing this.[7]

One may still wonder how I could in fact see the field and believe nothing regarding it. Must I not see it to be something or other, say, green? And if so, would I not believe, of it, *something* that is true of it, even if only that it is green? Consider a different example. Imagine that we are talking excitedly and a bird flies quickly across my path. Could I not see it, yet form no beliefs about it? There may be no decisive answer. For one thing, while there is much we *can* confidently say about seeing and believing, 'seeing' and 'believing' are, like most philosophically interesting terms, not precise.

A negative response might be supported as follows. Suppose I merely see the bird but pay no attention to it because I am utterly intent on what we are discussing. Why must I form any belief about the bird? Granted, if someone later asks if I saw a blue bird, I may assent, thereby indicating a belief that the bird *was* blue. But this belief is not perceptual: it is a *belief about a perceptible* and indeed has visual (roughly, visualizable) content, but it is not grounded in seeing or any other mode of perception. Moreover, it may have been formed only when I recalled my visual experience of the bird. Recalling that experience in such a context may produce a belief even if my original experience did not. For plainly a recollected sensory experience can produce beliefs about the object that caused it, especially when I have reason to provide information about that object.

It might be objected that genuinely seeing an object must produce beliefs. How else, one might ask, can perception guide behavior, as it does where, on seeing a log in our path, we step over it? One answer is that not everything we see, including the bird which flies by as I concentrate on something else, demands a cognitive response, even if it produces some other kind of response. If I am cataloging local birds, the situation is different. But where an unobtrusive object I see – as opposed to one blocking my path – has no particular relation to what I am doing, perhaps my visual impressions of it are simply a *basis* for forming beliefs about it should the situation call for it, and need not produce any belief if my concerns and the direction of my attention give the object no significance.

There may be an evolutionary explanation for the point that perception does not entail the formation of all the beliefs it warrants, if indeed of any at all. Certainly, it is in accord with what seems an economy of nature that beliefs not be formed unnecessarily. A single perceptual image, for instance, can contain, in readily usable form, all the information one needs to navigate an obstacle course. It may yield beliefs, say about how thin the ice is, the moment a relevant question arises; it may simply guide one's walk without yielding beliefs; and there are other ways in which, independently of producing beliefs, it may contribute to our survivability. Whether the brain is spared needless activity in these ways or not is an empirical question we need not pursue. The point is that the analysis of perception provides no good reason to posit all the perceptual beliefs some philosophers ascribe to perceivers and that there are preferable ways to explain the data, as will shortly be evident.[8]

Despite the complexity of the relation between seeing and believing, clearly we may affirm what is epistemologically most important here. If I can see a bird without believing anything about (or of) it, I still *can* see it to be something or other, and given my perceptual circumstances I might readily both come to believe something about it *and* see and know that to be true of it. Imagine that I am alone and see a dark bird in the distance for just a second, mistakenly taking it to be a speck of ash. If there is not too much color distortion, I may still both know and justifiedly believe it to be dark. Granted, I would misdescribe it, and I might falsely believe that it is a speck of ash. But I could still know something about it, and I might point the bird out under the misleading but true description, 'that dark thing out there'. It *is* that thing I point at; and I can see, know, and justifiedly believe that there is a dark thing there. My perception of the bird gives me a ready basis for this much knowledge and justification, even if the perception occurs in a way that does not cause me to believe (say) that there is a *bird* before me. Seeing *is* virtual believing, or at least potential believing. It is similar with the other senses.[9]

The perceptual hierarchy

Our discussion seems to show that simple perceiving need not produce belief, and objectual perceiving need not yield propositional perceiving. Still, the third kind of perception is not possible without the first and, I think, the second: I cannot see *that* the bird is anything, for example dark, if I do not see it at all, and apparently I must also see it to *be* something, say a speck of blue. Thus, simple perceiving is fundamental: it is required for objectual and propositional perceiving, though it does not clearly entail either. And since objectual perceiving seems possible without propositional perceiving, but not conversely, the former seems more nearly fundamental than the latter.

We have, then, a perceptual hierarchy: propositional perceiving depends on objectual perceiving, which in turn depends on simple perceiving. Simple perceiving is basic, and it commonly yields, even if it need not always yield, objectual perceiving, which, in turn, commonly yields, even if it need not always yield, propositional perceiving. Simple perceiving, such as just seeing a green field, may apparently occur without either of the other two kinds, but seeing something *to be* anything at all, such as rectangular, requires seeing it, and seeing *that* it is something in particular, say green, requires both seeing it to be something and, of course, seeing it.

If simple perception does not always produce at least one true belief, it characteristically puts us in a position to form any number of true beliefs. It gives us *access* to perceptual information, perhaps even *records* that information in some sense, whether or not we conceptually register the information by forming perceptual beliefs of either kind. As this suggests, perception by its very nature is informational; it might even be understood as equivalent to a kind of receipt of information about the object perceived.[10] The point here is that not all perceptually given information is *propositional* or even conceptualized. This is why we do not receive or store all of it in the contents of our beliefs. Some of the information is imagistic. Indeed, if we think of all the senses as capable of images or their non-visual counterparts for the other senses – *percepts* – it is in these sensory impressions that the bulk of perceptual information apparently resides. Hence the idea that a picture is worth a thousand words.

It is in part because perception is so richly informative that it normally gives us not only imagistic information but also what may be called *situational justification*: even if I could be so lost in conversation that I do not form any belief about the passing bird, I am, as I see it pass, normally *justified in believing* something about it, concerning its perceptible properties, for instance that it glides.[11] There may perhaps be nothing highly specific that I am justified in believing about it, say that it is a cardinal. But if I really see it, as opposed to its merely causing in me a visual impression too indistinct to qualify me as seeing it, then there is something or other that I may justifiably believe about it.

When we have a clear perception of something, it is even easier to have perceptual justification for believing a proposition about it without actually believing it. Just by taking stock of the size and texture of the field in clear view before me, I am justified in believing that it has more than 289 blades of grass; but I do not ordinarily believe any such thing about grassy fields I see. It was only when I sought a philosophical example about perception and belief, and then arbitrarily chose the proposition that the field has more that 289 blades of grass, that I came to believe this.

Seeing and Seeing As

What is it that explains why seeing the bird or the field justifies me in believing something about what I see, that is, gives me situational justification for such a belief? And does the same thing explain why seeing something enables one to know various facts about it? One possible answer is that if I see something at all, say a bird, I *see it as* something, for instance black, and I am justified in believing it to be what I see it as being. The idea is that all seeing and perhaps all perceiving is *aspectual perception.* We see things by seeing their properties or aspects, for instance their colors or their front sides, and we are justified in taking them to have the properties or aspects we see them as having.[12]

Let us not go too fast. First, might not the sort of distinction we have observed between situational and belief justification (doxastic justification) apply to seeing itself? Specifically, might not my seeing the bird only imply that I am in a *position* to see it *as* something, and not that I *do*? After all, just because, when I do see something, I see it *by* seeing some property or aspect of it, we may not conclude that I see it *as having* this property or aspect. I might think of the property as belonging to something else, as I might see a person by observing her movements under her umbrella but take them to be those of her sister. Second, supposing that seeing the bird does imply seeing it *as* something, clearly this *need* not be something one is justified in believing it to be (and perhaps it need not be something one *does* believe it to be). Charles might erroneously see a plainly black bird as blue, simply because he so loves birds of blue color and so dislikes black birds that (as he himself knows) his vision plays tricks on him when he is bird-watching. He might then not be justified in believing that the bird is blue.

Suppose for the sake of argument that seeing implies seeing *as* and that typically seeing *as* implies at least objectually believing something or other about the thing seen. Still, seeing an object as having a certain property – say, a stick in the water as bent – does not entail that it has the property. Nor does it always justify one's believing it to have that property.

Seeing As and Perceptual Grounds of Justification

Whether or not seeing always implies seeing *as,* it is clear that seeing something normally puts one in a position to form at least one justified belief about it. Suppose I see the bird so briefly and distractedly that I do not see it as anything in particular; still, my visual impression of it has some feature or other by which I am justified in believing something of the bird, if only that it is a moving thing. Even Charles would be justified in believing something like this. Suppose, however, that for hours Charles had been hallucinating all manner of unreal things, and he knows this. Then he might not be justified in taking the bird he sees to be *anything* real, even though it is real. For as a rational person in this position he should see that if his belief is true, it may well be grounded only in the unreliable way a lucky guess is. Thus, the best conclusion here is – and this is an important justification principle concerning perception – that *normally,* seeing an object gives one situational justification for believing something or other about it.

More broadly, the *evidence of the senses* – including above all the sensory experiences characteristic of perception – normally provides justification for beliefs with content appropriate to that evidence. If your (visual) experience is of a green expanse, you are justified in believing there is something green before you; if your (tactual) experience is of something cool in your hand, you are justified in believing there is something cool in your hand; and so on. The suggested principle is vulnerable to skeptical worries, but I suspect that the most plausible of them can be accommodated by granting the following double-barrelled point: the analysis of perception and perceptual justification given here allows both that normal conditions can be mistakenly thought to obtain and that the situational justification in question need not be strong. Even Hume did not carry his skeptical doubts so far as to prevent his averring that "none but a fool or madman will ever pretend to dispute the authority of experience ..." and elsewhere he went so far as to say, "I know with certainty, that he [a man] is not to put his hand in the fire, and hold it there, till it be consumed ..."[13]

One might also say something slightly different, in a terminology that is from some points of view preferable: seeing an object gives one *prima facie* justification for believing something or other about it, where prima facie justification is roughly justification that prevails unless defeated, for instance by a strong justification for believing something to the contrary. If I see a green field, I have a justification for believing it to be green, but I may not be justified, overall, in believing this if credible friends give me compelling reason to believe that despite appearances the field is entirely covered by blue grass, or that I am merely hallucinating.[14]

If seeing is typical of perception in (normally) putting us in a position to form at least one justified belief about the object seen, then perception in general normally gives us at least situational justification (roughly, justification *for* forming a belief of the proposition in question). It does not follow, however, that every perceptual belief *is* justified. Some perceptual beliefs, like perceptual beliefs that are evidentially outweighed by similar beliefs grounded in hallucinations, are not. An army of hallucinated bird songs can evidentially outweigh the veridical sight of an empty sky above.

Nevertheless, there is a simple principle of justification – call it the *visual principle* – that remains plausible despite these complexities: when a visual belief arises in such a way that one believes something in virtue of either seeing *that* it is so or seeing it *to be* so, normally the belief is justified (and it is always prima facie justified). If I see that the field is rectangular and, in virtue of seeing that it is rectangular, believe that it is, then (normally) I justifiedly believe that it is. I say *normally* (and that the justification is prima facie) because even here one's justification can be *defeated.* Thus, Charles might see that a bird is blue and believe on that basis that it is, yet realize that all morning he has been seeing black birds as dark blue and thus mistaking the black ones for the blue ones. Until he verifies his first impression (of a blue bird), then, he does not justifiedly believe that the bird is blue, even though it in fact is. (We could say that he has some justification for believing this, yet better justification for not believing it; but to simplify matters I am ignoring degrees of justification.)

Suppose, on the other hand, that Charles has no idea that he has been hallucinating. Then, even when he does hallucinate a blue bird he may be justified in believing that there is one before him. This suggests a related principle of justification, one applicable to visual experience whether it is a case of seeing or merely of visual hallucination: when, on the basis of an apparently normal visual experience, one believes something of the kind the experience seems to show (for instance that the bird is blue), normally this belief is justified. Call this the *visual experience principle,* since it applies to cases in which one has a belief based on visual experience even if not an experience of actually seeing. The visual principle takes us from seeing to justification; the visual experience principle takes us from visual experience – conceived as apparent seeing – to justification.

Similar principles can be formulated for all of the other senses. If, for example, you hear a note to be flat and on that basis believe that it is flat, normally your belief is justified. It is grounded in a veridical perception of the flatness of the note, a perception in which you have discriminated the flatness you believe the note has. And suppose, by contrast, that in what clearly seem to be everyday circumstances you have an utterly normal-seeming auditory hallucination of a flat note. If that experience makes it seem clear that you are hearing a flat note, then if you believe on the basis of the experience that this is a flat note, normally your belief would be justified. In your situation, you have no reason to suspect hallucination, and the justification of your belief that the note is flat piggybacks, as it were, on the principle that normally applies to veridical beliefs.[15]

Perception as a Ground of Knowledge

Some of what holds for the justification of perceptual beliefs also applies to perceptual knowledge. Seeing the green field, for instance, normally yields knowledge about the field as well as justified belief about it. This suggests another visual principle, which I shall call *epistemic* since it states a condition for the visual generation of *knowledge*: at least normally, if one sees that a thing has a property (say is rectangular), one (visually) knows that it has it. A parallel epistemic principle holds for objectual seeing: at least normally, if one sees something to have a property, one knows it to have the property.

There are, however, special circumstances that explain why these epistemic principles are restricted to "normal" cases. Perhaps I can see that something is so, believe on that basis that it is, and yet not know that it is. Charles's case *seems* to show this. For if, in similar circumstances, he often takes a black bird to be blue, then even if he sees that a certain blue bird is blue and, on that basis, believes it is blue, he apparently does not know that it is.[16] He is just lucky that this time his belief is true and he wasn't hallucinating. Since he has no reason to think he has been hallucinating, one cannot fault him for believing the bird is blue or regard the belief as inappropriate to his situation. Still, knowledge apparently needs better grounding than is provided by his blameless good fortune. This kind of case has led some philosophers to maintain that when we know that something, our being right is not *accidental*.

There is an important difference here between knowledge and justification. Take knowledge first. If Charles is making errors like this, then even if he has no idea that he is and no reason to suspect he is, he does not know that the bird he believes to be blue is blue. But even if he has no idea that he is making errors, or any reason to suspect he is, he may still justifiedly believe that the bird is blue. The main difference may be this: he can have a true belief which does not constitute knowledge because there is something wrong for which he is in no way criticizable; but he cannot have a true yet unjustified belief without being in some way criticizable. The standards for knowledge, one might say, permit fewer unsuspected weaknesses in discriminating the truth than those for justification, if the standards for knowledge permit any at all.

This difference between knowledge and justification must be reflected in the kinds of principles that indicate how justification, as opposed to knowledge, is generated. Justification principles need not imply that the relevant basis of a belief's justification assures its truth; but since a false belief cannot be knowledge, epistemic principles cannot capture elements that generate knowledge unless they rule out factors that may produce a false belief (or at least factors that have a significant chance of producing one). A ground of knowledge must, in *some* way, suffice for the truth of the proposition known and must be in that sense *externally successful*; a ground of justification must in some way *count toward* the truth of the proposition one is justified in believing, but need not rule out its falsehood. If this is so, one might say that it need only be *internally successful:* roughly, successful from the point of view of grounds for belief to which the subject has access by reflection or introspection.

On the basis of what we see, hear, feel, smell, and taste, we have a great many beliefs, propositional and objectual. It appears that these perceptual beliefs are commonly justified and, quite often, constitute knowledge. But to see that perception is a basis of justification and knowledge is to go only partway toward understanding what perception is. Until we have a more detailed understanding of what it is, we cannot see in detail how perception grounds belief, justification, and knowledge. I want to discuss (further) what perception is first and, later, to illustrate in new ways how it grounds what it does.

I.3 Some Commonsense Views of Perception

One natural thing to say about what it is for me to see the green field is appealingly brief: I simply see it, or at least its facing surface. It is squarely before me. I need no light to penetrate a haze or a telescope to magnify my view. I simply see the field, and it is as it appears. This sort of view thought to represent untutored common sense has been called *naive realism*: it says roughly that perception is simply a matter of the five senses telling us about real things.

The view is naive because it ignores problems of a kind to be described in a moment; it is a form of realism because it takes the objects of perception to be real things external to the perceiver, the sorts of things that are "out there" to be seen whether anyone sees them or not.

A more thoughtful commonsense view retains the realism without the naivety. It is quite commonsensical to say that I see the field *because* it is before my open eyes and stimulates my vision, thereby *appearing* to me as a green, rectangular shape. Stimulating my vision is a causal relation: the field, by reflecting light, causes me to have the visual experience that is part of my seeing that very field. Moreover, the field apparently must cause my visual experience if I am to *see* it. Since the more thoughtful commonsense view specifies that the object of perception must be a real external thing, we might call it a *perceptual realism.* Most theories of perception incorporate this kind of realism.

To see the need for a causal element, suppose I am looking at the field and, without my noticing, someone instantaneously drops a perfect picture of the field right in front of me. The scene might appear to me just as it did, yet I no longer *see* the field. Instead, I see a picture of it. (I do see the field *in* the picture, but that is *secondary seeing* and not the kind I am talking about.) The reason I do not now see the field is roughly that although it is indeed before me, it has no (causal) effect on my visual experience.

Perception as a causal relation and its four main elements

Examples like this suggest that *perception is a kind of causal relation* between the perceiver and whatever is perceived. This is an important point, though it does not tell us precisely what perception is. I call any theory of perception which incorporates the point *a causal theory of perception.* Most theories of perception are causal.[17]

We can now better understand the four elements I have described as among those crucial in perception: the perceiver, the object perceived, the sensory experience in which the object appears to the perceiver, and the causal relation between the object and the perceiver, by virtue of which the object produces that experience. Thus, if I see the field, there is a distinctive way, presumably through light transmission to my eyes, in which the field produces in me the visual sensory experience of a green, rectangular shape characteristic of my seeing it.

It is difficult, though fortunately not necessary for a general understanding of perception, to specify precisely what these ways – these causal paths from the object to the perceiver – are. Some of the details are the business of the psychology and neurophysiology of perception. Others are determinable by philosophical inquiry. Philosophical reflection shows, for instance, that not just any causal chain is the right sort for perception. Some of these chains are "wayward." Suppose the piano sounds cause a special machine to produce in me both temporary deafness and a faithful auditory hallucination of the piece. Then I do not *hear* it, though my sensory experience, the auditory experience I live through in my own consciousness, is just what it would be if I did hear it. Different theories of perception tend, as we shall see, to give strikingly different accounts of how these four elements (or some of them) figure in perception.

Illusion and hallucination

Suppose the book I am holding appears, from a certain angle, as if its cover were a parallelogram rather than a (right) rectangle, or feels warm only because my hand is cold. This is a *perceptual illusion*. Now imagine that the field I see burns up. I sorely miss its rich green, and on waking from a slumber in my chair I have a *hallucination* in which my visual experience is just as it would be if I were seeing the field as it originally was. Here the grass I seem to see is not there at all. The point here is not that *something* I see is not as it seems (as in the case of illusion), but that there seems to *be* something where there is nothing. With illusion, as illustrated by a partly submerged stick's looking bent, experience distorts what is there; with hallucination, something seems to be there that apparently is not there at all.

One way to deal with illusion and hallucination is to stress how they show the need to distinguish appearance from reality. In a visual illusion, one sees something, but it does not appear as it really is, say rectangular. In a hallucination, if anything appears to one, it is in reality even less what it appears to be than is the object of an illusion, or is not what it appears to be at all: instead of a blue spruce tree's appearing blue to me, for instance, perhaps the conical section of space where it stood appears "bespruced".

I.4 The Theory of Appearing

The sort of account of perception just sketched as an improvement over naive realism has been called *the theory of appearing:* it says roughly that perceiving an object, such as a book, is simply its appearing to one to have one or more properties, such as being rectangular. Thus, one perceives it – in this case, sees it – *as* rectangular. The theory can also provide the basis of an account of sensory experience, including not just the kind one has in actually perceiving something but also the sort one has in hallucination. That, too, the theory takes to be a case of something's appearing to one to have a set of properties; the object that appears is simply a different kind – hallucinatory.[18]

The theory of appearing is initially plausible. It includes the plausible view that if one sees something, then it appears to one in some way, say as a red barn or as a red spot in a field. The theory also does justice to the view that things are not always as they appear. Moreover, it can explain both illusion and, with some imaginative development, hallucination.

The theory of appearing says nothing, however, about the need for a causal relation between the object and its perceiver. If, consistently with its commonsense motivation, one stipulated that the crucial relation of appearing to the perceiver to have a property – say, to be rectangular – is or implies a causal relation, one would then have a different theory (of a kind to be discussed shortly).

In addition to the question of how the theory can do justice to the causal element in perception, it has difficulty accounting for hallucinations in which there apparently *is* no object to appear. I could hallucinate a green field when I see nothing physical, say because it is pitch dark. In such an *empty hallucination* – one that occurs despite my perceiving nothing – what is it that appears green to me? There is a plausible answer; but it is associated with a quite different theory of perception.

I.5 Sense-Datum Theories of Perception

Once we think seriously about illusion and hallucination, we begin to question not only naive realism but also any kind of *direct realism,* any view which, like the theory of appearing, says that we see (or otherwise perceive) external objects directly, rather than *through* seeing (or at least visually experiencing) something else. Hallucination illustrates most readily how such an intermediary may seem essential to perception. Imagine that when I vividly hallucinate the field, my visual experience – roughly, what I am aware of in my visual consciousness – is exactly like the experience I have when I see the field. Does it not then seem that the difference between ordinary seeing and visual hallucination is simply in what *causes* the visual experience, rather than in the visual experience itself or in what (if anything) I directly see? When I see the field, *it* causes my visual experience. When I hallucinate it, something else (such as my deep desire to have it back) causes my visual experience. But apparently what I directly see, that is, the immediate object of my visual experience, is the same in both cases. This point presumably explains why my visual experience – what occupies my visual consciousness – is the same whether I am hallucinating the field or really seeing it.

The argument from hallucination

We might develop these ideas by considering an argument from hallucination:
(1) A perfectly faithful (visual) hallucination of a field is intrinsically indistinguishable from an ordinary experience of seeing that field, that is, not distinguishable from it just in itself as a visual experience, as opposed to being distinguishable through verifying one's visual impression by touching the things around one.
Hence,
(2) What is *directly* seen, the immediate object of one's visual experience, is the same sort of (non-physical) thing in a perfect hallucination of a field as in an ordinary experience of seeing a field.
But clearly,
(3) What is directly seen in a hallucination of a field is not a field (or any other physical thing).
Indeed, no field is seen at all in an hallucinatory visual experience, so (3) seems plainly true. Hence, putting (1)-(3) together,
(4) What is directly seen in an ordinary experience of seeing a field is not a field.

The overall idea is that when we ordinarily see an everyday perceptible object such as a field, we see it through seeing something else *directly*. One may prefer (as some philosophers do) to say that we do not *see* such things but are only visually acquainted with them. To simplify, however, let us for most purposes use the more natural term 'see'.

Just what is directly seen when one sees a field, then, and how is the field *in*directly seen? Why not say that what is *directly* seen is a two-dimensional object (or perhaps even a three-dimensional item) consisting of the colors and shapes one sees in the hallucinatory experience? After all, nothing, not even (physical) light, intervenes between me and them. There is no "space" for intermediaries. Hence, no intermediaries can misrepresent these special objects. These objects are apparently internal to me. Yet I do see the field *by* seeing them; hence, I see it indirectly.

A sense-datum theory does not require giving up a causal theory of perception: the field causes the colors and shapes to arise in my visual consciousness in a way that fully accords with the view that perception is a causal relation between something external and the perceiver. Perception is simply a *mediated,* hence indirect, causal relation between external objects I perceive and me: the object produces the mediating colors and shapes that appear in my visual field, and, through seeing them, I see it. This *sense-datum theory of perception* (unlike the phenomenalist sense-datum view discussed below) is a realist view; but its realism, by contrast with that of naive realism and the theory of appearing, is indirect.[19]

Sense-datum theory as an indirect, representative realism

A sense-datum theory might be called a *representative realism* because it conceives perception as a relation in which sense-data represent perceived external (hence real) objects. On some conceptions of sense-data, they are copies of those objects: shape for shape, color for color, sound for sound. John Locke held a view of this kind,[20] though for him sense-data are copies ("resemblances") only of the *primary qualities*, solidity, extension (in space), shape, and mobility, not of the *secondary qualities*, above all colors, sounds smells, and tastes.

To appreciate the theory better, consider how it takes perception to be indirect. Sense-datum theorists might offer several reasons to explain why we do not ordinarily notice the indirectness of perception (I speak generally here, not solely of Locke's theory). Here are two important ones. First, normally what we directly see, say colors and shapes, roughly corresponds to the physical objects we indirectly see by means of what we see directly. It is only when there is an illusion or hallucination that we are forced to notice a discrepancy between what we directly see and the object in question, say a book. Second, our perceptual beliefs are spontaneously formed and not based on any process requiring us to consider sense-data. Above all, we do not normally *infer* what we believe about external objects from what we believe about the colors and shapes we directly see.[21] This is why it is easy to think we "just see" things, directly. Perceiving is not inferential, and for that reason (perhaps among others) it is not *epistemically indirect,* in the sense that knowledge of external objects or belief about them is based on knowledge of sense-data, or belief about them.[22] I know that the field is green through *having* green sense-data, not through *inference from* propositions about them. Perception is, however, causally and objectually indirect. The perceived object is presented to us via another object, though not by way of a *premise*. This is consistent with the idea, prominent in psychological literature, that one perception is often based on or at least affected by another perception and thereby indirect in a third sense.[23] Let me describe a bit differently how the sense-datum view conceives the indirectness of perception.

Perception is causally indirect because perceived physical objects cause sensory experience *by* causing the occurrence of sense-data, with which we are directly (and presumably non-causally) acquainted in perceptual experience. Perception is objectually indirect because we perceive external things, such as fields, *through* our acquaintance with other objects, namely, sense-data. Roughly, we perceive external things through perceptual acquaintance with internal things. By contrast, we normally do not use information about sense-data to arrive at perceptual beliefs inferentially, say by an inference from my directly seeing a grassy, green rectangular expanse to the conclusion that a green field is before me.

Appraisal of the sense-datum approach

Let us focus first of all on the argument from hallucination, whose conclusion suggests that what is directly seen in visual perception of external objects is a set of sense-data. Suppose I do have a hallucination intrinsically just like the normal experience of seeing a field. Does it follow that what is directly seen in the hallucination is the same sort of thing as what is directly seen in the normal experience? There are at least two problems that confront the sense-datum theory here.

First, why *must* anything be seen at all in a hallucination? Imagine that you see me hallucinate the burned-up field. I might get up, still half asleep, and cry out, 'It's back!', pointing to the area. You might conclude that I *think* I see the field again. My own initial reaction to realizing I had hallucinated the field might be that, hallucination or no, I *saw* it. But I might just as easily slump back in my chair and mumble that I wish I had seen it. We could compromise and agree that I saw the hallucinated field (vividly) *in my mind's eye.* But suppose I did see it in my mind's eye, and again suppose that the hallucination is intrinsically just like the ordinary seeing. Does it follow that what I directly see in the ordinary experience is the same as what I see in the hallucination, namely, something in my mind's eye? It does not. The notion of seeing in one's mind's eye is metaphorical, and such seeing need not imply that there is any real thing seen, in or outside the mind.

There is a second reason to resist the conclusion that something must be directly seen in hallucinations. Recall that my seeing a green field is apparently a causal relation between a sensory experience in me and the field that produces the experience. If so, why should the possibility that a hallucination can mimic my seeing the field tell us anything about what is directly seen when one sees that field? It is not as if we had to assume that only an *object* can produce the relevant sensory experience, and must then conclude that it is an internal perceptual object, since there is no other candidate. Many effects can have more than one cause, and the sense-datum theorist has no argument to show that only an internal perceptual *object*, as opposed, say, to an abnormality in the visual cortex, can cause the hallucinatory experience.

Moreover, from the similarity of the internal, experiential elements in the hallucination and the genuine perception, one might as well conclude that since the ordinary experience is one of seeing only an external rather than an internal object, the hallucinatory experience is different only in the absence of the external object. Rather than add to the components that seem needed to account for the ordinary experience, we subtract one that seems needed to account for the hallucination. This yields a more economical theory of perception. Consider an analogy. Two perfect ball bearings can be intrinsically indistinguishable, having the same diameter and constitution, yet still differ significantly, one being on my left and one on my right. Their intrinsic properties can thus be identical, while their *relations* (to me) differ: one is left of me, the other right of me; hence they *do* differ. Similarly, the hallucination of a field and the ordinary visual experience of a field can be intrinsically indistinguishable, yet differ in their relations to me or to other things.

One, the visual experience of a field, may be an element in a perceptual relation to the field; the experience we call hallucination, which is not based on perceiving the external object hallucinated, may not be an element in any perceptual relation to the field, but only an experience or a process I undergo.

To account for the difference between the two kinds of experience, we might say this. The visual experience, it seems, represents an external thing to me; the hallucinatory experience, though intrinsically just like the visual one, does not, but as it were only pretends to represent an external thing. Thus, for all the argument from hallucination shows, the ordinary experience of seeing might be a relation to an object such as a green field, namely the relation of directly seeing, while the hallucinatory experience of a green field is not a relation to that field, such as being an internal copy of it, nor even a relation to any other object, such as a perceiver.

The points just made about the argument from hallucination indicate that it is not sound. (1) does not entail (2). Nonetheless, the argument poses serious problems for alternative theories. What conception of hallucinations and illusions besides the sense-datum account might we adopt? Recall the book viewed from an angle. A sense-datum theory will say we directly see a parallelogrammic shape and indirectly see the book. The theory of appearing, however, can also explain this: it reminds us that things need not be what they appear to be and says simply that the book can appear parallelogrammic even if it is a right rectangle. One could also combine the causal element in the sense-datum approach with the direct realism of the theory of appearing and move to a third theory, one that says the book causes us to see it directly, rather than through producing sense-data in us, yet (because of our angle of vision), we see it as if it were parallelogrammic. To avoid suggesting that anything in one's experience need *be* parallelogrammic, one could take this to mean that the book visually appears parallelogrammically to us. Here the adverb 'parallelogrammically' describes a *way* in which we visually experience the book; it does not imply that there is an object that appears to us and *is* parallelogrammic.[24] Let us explore this idea.

I.6 Adverbial Theories of Perception

It should now be clear why we need not grant (what sense-datum theorists sometimes seem to assume about perception) that in order for an object to appear a given way to us there must *be* something we see that *is* that way, for instance a parallelogrammic sense-datum. Moreover, it is not only the theory of appearing that makes use of this point. Suppose that one says simply that the book appears parallelogrammically, using this adverb to designate the way it appears, or (speaking from the perceiver's point of view) *how* one visually experiences it: parallelogrammically. To say it appears parallelogrammically is roughly to say it appears in the way a parallelogram does, as opposed to the way a rectangle does. Similarly, if I say I have a fever, no one could plausibly insist that there is an object, a fever, which I have. 'I have a fever' is a way of saying I am feverish, i.e., my body is above a certain temperature. What our language seems to treat as a statement of a relation to an object, a fever, is really an ascription of a property.

Unlike the theory of appearing, which takes perception to be an unanalyzable relation, this *adverbial theory of perception* conceives perception as an analyzable way of experiencing things. In what may be its most plausible form, it says roughly that to perceive an object is for that object (in a certain way) to produce in one a sensory experience of it; more specifically, to cause one's experiencing it in a certain qualitative way, say to see a stick as straight (or, given the illusion induced by partial submersion, as bent). Both theories are, however, direct realist views. Other similarities (and some differences) between the two theories will soon be apparent.[25]

So far, so good, perhaps. But what about hallucinations? Here the adverbial theory again differs from the theory of appearing. Unlike the latter, it denies that all sensory experience is *of* some object. The importance of this denial is not immediately apparent, perhaps because we suppose that usually a person visually hallucinating does see *something*. Recall Shakespeare's Macbeth, who, distraught by his crime, hallucinated a dagger that seemed to him to hover in midair. Presumably he saw something, say the wall behind "the dagger" or at least a chunk of space where it hovers. An adverbial theorist might thus posit an object where the "dagger" seems located, which Macbeth experiences "daggerly." Somehow this object might be thought to play a role in causing him to have daggerish visual sensations, just as, for the theory of appearing, the space before him, despite being transparent, might somehow appear to him to be a dagger.

Supposing we accept this adverbialist account, what happens if it is pitch dark and Macbeth's hallucination is therefore *empty*, in the sense that there is nothing he sees, and hence no object distorted into an apparent dagger? Then, whereas the theory of appearing may have to posit something like a sense-datum to serve as what appears to be a dagger, the adverbial theory can deny that there is *any* kind of object appearing to him. It may posit some quite different account of his "bedaggered" visual experience, such as a psychological account appealing to the influence of drugs.

Is it really plausible to hold, with the adverbial theory, that Macbeth saw nothing at all? Can we explain how the normal and hallucinatory experiences are intrinsically alike without assuming they have the same direct objects? In the light of the special case of empty hallucination, the sense-datum theory may seem the most plausible of the three. It provides an object of Macbeth's visual experience in utter darkness, whereas the adverbial theory posits no objects at all to appear to one in empty hallucinations. Moreover, the sense-datum view postulates the same sort of direct object for ordinary perception, illusion, and hallucination, whereas the theory of appearing does not offer a uniform account of their direct objects and must explain why entities like sense-data do not occur in normal perception as well as in empty hallucination.

Perhaps, however, the hallucination problem seems more threatening than it should to the adverbial theory because hallucinations are felt to be *perceptual* experiences and hence expected to be *of* some object. But as we have seen, although hallucinatory experiences can be intrinsically indistinguishable from perceptual ones, all that can be assumed is that they are *sensory experiences*. Hallucinatory experiences, on the adverbial view, are simply not cases of perceiving, at least not in a sense requiring that any object appear to one. Thus, nothing at all need appear to one in hallucinations, though it may *appear to the subject* that there is something there.

I.7 Adverbial and Sense-Datum Theories of Sensory Experience

A perceptual experience is always sensory, and normally a sensory experience of the sort we have in perceiving is genuinely perceptual. But a kind of short-circuit can cause the sense-receptors to produce sensory experience that is not even part of a normal perceptual experience. It is important to consider the debate between adverbial and sense-datum theories in relation to sensory experience.

The most natural thing for adverbial theorists to say about hallucinatory experience is that it is not genuinely perceptual, but only sensory. They might, however, say instead that where a perceptual experience is hallucinatory, it is not a case of *seeing*. (One might perhaps consider it seeing "in the mind's eye", or perhaps in the sense that it is seeing colors and shapes conceived abstractly as properties and not as belonging to sense-datum objects, but this is not ordinary seeing.) The former description accords better with how seeing is normally understood.

The theory suggested by these responses to the hallucination problem might be called the *adverbial theory of sensory experience.* It says that having a sensory experience, such as a hallucination of a green field, is experiencing in a certain *way,* for example visually experiencing "green-fieldly". Our commonsense assumption is that hallucination is not normal and that most such vivid sensory experiences are genuinely perceptual. They are of, and thus caused by, the external object apparently perceived. But some sensory experiences are neither genuinely perceptual nor externally caused. People having them are in a vision-like state, and what is going on in their visual cortex may be the same sort of process that occurs when they see things; yet they are not seeing.

May we, then, regard sense-datum theories of perception as refuted by the points just made in criticism of the argument from hallucination and on behalf of the suggested adverbial theory and the theory of appearing? Certainly not. We have at most seen how one major argument for a sense-datum theory of perception fails and how alternative theories of perception can account for the apparently central elements in perception: the perceiver, the (ordinary) object perceived, the sensory experience, and the causal relation between the second and third.

Indeed, supposing that the argument from hallucination fails to show that sense-data are elements in normal everyday perception, sense-data might still be needed to account for non-perceptual sensory experience. This is sometimes loosely called "perceptual" experience because it is characteristic of that by virtue of subjective similarity to it. In this limited role, one might posit a *sense-datum theory of non-perceptual sensory experience:* such experience is simply direct acquaintance with sense-data. (I leave aside experiences of "inner sense", such as hunger, itching, and pain, which are treated below.) This view may seem preferable to an adverbial theory of sensory experience. For one thing, there is something prima facie unsatisfying about the idea that even in a visual hallucination so vivid that, if one did not suspect error, one would stake one's life on the presence of the hallucinated object, one sees nothing, except either metaphorically in one's mind's eye, or in a sense of 'see' which does not require that any object is seen. Still, perhaps there is such a sense of 'see', or perhaps one can experience colors and shapes in a visual way without seeing anything. Adverbial theorists will tend to argue that reflection on these possibilities should dispel the dissatisfaction.

Another aspect of the controversy concerns the metaphysics associated with adverbial and sense-datum theories, specifically, the sorts of things they require us to take as real. In this respect, the adverbial theories of perception and sensory experience have a definite advantage over the counterpart sense-datum theories: the former do not posit a *kind* of object we would not otherwise have to regard as real. From the adverbial perspective, the objects that perception and sensory experience involve are simply perceivers and what they perceive. These are quite familiar entities which we must recognize and deal with anyway.

Sense-data are quite different from ordinary (presumably physical) objects of perception. Sense-data are either mental or at least depend for their existence on the mind. Yet they are unlike some mental phenomena in that no good case can be made for their being really brain phenomena, since they have properties, for instance green color and rectangularity, not normally found in the brain.[26] Moreover, there are obstacles to fully understanding sense-data. Is there, for instance, a reasonable way of counting them? Suppose my image of the green field gradually gets greener. Is this a sense-datum changing or a new one replacing an old one? There seems to be no way to tell. If there is none, how can we ever be sure we learn more about a sense-datum than what initially appears to us in experiencing it: how can one distinguish learning something more about *it* from learning about something new?[27]

Problems like these also affect the theory of appearing insofar as it must posit sense-data or similar entities to account for hallucinations. To be sure, such problems can beset our understanding of ordinary objects as well. Can we always distinguish a mountain with two peaks from two mountains, or one snarled barberry bush from two? But apparently these problems are less serious, if only because there is no question that there are *some* things of the physical kind in question. The corresponding problems may in the end be soluble for sense-data, but they at least give us some reason to prefer a theory that does not force us to regard sense-data as the only objects, or as even among the objects, we are directly aware of when we see, hear, touch, taste, and smell.

I.8 Phenomenalism

If some philosophers have thought that perception can be understood without appeal to sense-data, others have conceived it as understandable in terms of sense-data alone as its objects. This view has the advantage of being, in at least one way, simpler than the adverbial and sense-datum theories. But the view is motivated by other considerations as well.

A sense-datum version of phenomenalism

The book you see is a perceptible object. Suppose we may conceive a real perceptible object as a perceptible object that is as it is independently of what we think it to be. Still, real perceptible objects, such as tables and chairs, are also plausibly conceived to be, by their very nature, *knowable.* Indeed, it is doubtful that real objects of this sort *could* be unknowable, or even unknowable through the senses if lighting and other perceptual conditions are good. Now suppose we add to these ideas the assumption that our only genuine, certain knowledge of perceptibles is restricted to what directly appears to us and would be as it is even if we should be hallucinating. And what more does appear to us besides the colors and shapes and other observable properties of perceptible objects? Further, how do we know that this book, for example, could even exist without someone's perceiving its color and other sensory properties? Certainly we cannot *observe* the book existing unperceived. If you observe it, you perceive it.

Moreover, if you imagine subtracting the book's sensory properties one by one – its color, shape, weight, and so on – what is left? This is like stripping layer after layer from an onion until nothing remains. Might we not conclude, then, that the book is not only *known by* its properties, as the other theories of perception also hold, but simply *is* a stable collection of sensory properties, a collection of visual, tactual, and other sense-data which in some sense recur in our experience, say confronting us each time we have the sense-data corresponding to a certain bookcase in our home?

George Berkeley argued from a variety of angles that this is indeed what a perceptible object is.[28] The view is a version of *phenomenalism,* so called because it constructs external objects out of phenomena, which, in this use of the term, are equivalent to sense-data. The view is also considered a kind of *idealism,* since it construes physical objects as ideal, in the sense of being composed of ideas rather than material stuff that would exist even if there were no minds and no ideas.[29] On either construction, the view is among the clearest cases in which a metaphysics seems to be derived from an epistemology, here the ontology of perceptible objects from the mental properties by which they are taken to be known.

Adverbial phenomenalism

Phenomenalism as just described is focused on the nature of perceptible objects but implies a related view of perception. In the sense-datum version of phenomenalism, the associated account of perception retains a sense-datum theory of sensory experience, but not a sense-datum theory of perception. The latter view posits external objects as causes of the sense-data experienced in ordinary perception, whereas sense-datum phenomenalism says physical objects *are* collections of sense-data.

Using the adverbial theory of sensory experience, one might also formulate an *adverbial phenomenalism,* which constructs physical objects out of sensory experience alone and says that to see (for instance) a green field is to experience "green-fieldly" in a certain vivid and stable way. To see such a thing is to have a visual experience that predictably occurs under certain conditions, say when one has the related experiences of looking outward from the porch. Thus, perception can occur without even sense-data; it requires only perceivers and their properties. Sense-datum versions of phenomenalism, however, have been more often discussed by philosophers, and I will concentrate on them.

Whereas the sense-datum theory is an indirect realism, phenomenalism is a *direct irrealism*: it says that perceptual objects are directly perceived, but denies that they are real in the sense that they are *mind-independent* and can exist apart from perceivers. This is not to say they are not perceptually real – real items in sensory experience. The point is that they are not metaphysically real: "out there" and such that they would exist even if there were no perceivers.

Phenomenalism does not, then, deny that physical objects exist in the sense that they are both stable elements of our experience and governed by causal laws, such as those of physics. Nor does it deny that there can be hallucinations, as where certain sense-data, like those constituting Macbeth's hallucinatory dagger, are too unstable to compose a physical object, or are perceivable only in one mode, such as vision, when they should have tactile elements as well, such as a cool smooth surface. What phenomenalism denies is that physical objects are real in the classical sense implying mind-independence.

One naturally wonders why things would not go in and out of existence depending on whether they are experienced and why, when they do exist, they obey the laws of physics, which certainly do not seem to depend on our minds. Berkeley did not neglect to consider what happens to things when we cease to perceive them. External objects are sustained by constant divine perception. A phenomenalist need not be a theist, however, to offer an account of the stability of external objects and their lawful behavior. John Stuart Mill, without any appeal to God, called external objects "permanent possibilities of sensation". To say that the book is in the room when no one is in there to perceive it is to say that there is a certain enduring possibility of the sensations one would have if one perceived such a book. If one enters the room and looks in the appropriate direction, that possibility should be realized. By contrast, if one had merely hallucinated a book in the room, there would be no reason to expect this.

Appraisal of phenomenalism

Unlike the sense-datum theory of perception, phenomenalism is only occasionally defended by contemporary philosophers. But it has had major influence. Moreover, compared with the sense-datum theory, it is more economical and in that way simpler. Instead of perceivers, sense-data, and external objects, it posits, as the things figuring in perception and sensory experience, just perceivers and sense-data. Indeed, adverbial phenomenalism does not even posit sense-data, though it does appeal to a special kind of property, that of experiencing in a certain way.

There is, however, an apparently decisive objection to phenomenalism, one that shows something important about the relation between sense experience and external objects. The theory says that a book, for instance, is – or at least that its presence implies – one's having or potentially having a suitably stable collection of sensory experiences (e.g. being visually acquainted with the relevant sense-data). If the implied analysis of what it is to see a book is correct, then there is a combination of sensory items like colors and shapes in one's visual field such that if, under appropriate conditions, these elements occur in me, then it follows that I see a book. But surely there is no such combination of sense-data. No matter how vividly and stably I (or anyone) may experience the colors and shapes appropriate to a book, it does not follow that anyone sees one. For it is still possible that I am just hallucinating one or seeing something else *as* a book.[30]

This kind of hallucination remains possible even if I have supporting tactual experiences, such as the smooth feel of paper. For even the sense of touch can be stimulated in this way without one's touching a book. Thus, seeing a book is not *just* having appropriate booklike experiences, even if it *is partly* this, and even though, as phenomenalists hold, there is no experienceable difference between a sufficiently stable combination of bookish sensory experiences and an independently real book. Still, if seeing a book is not equivalent to any such sensory experiences, phenomenalism fails as an account of the perception of ordinary objects. If there are objects for which it holds, they are not the kind we have in mind in seeking an account of perception.

I.9 Perception and the Senses

I want to conclude this treatment of ordinary perception by exploring some remaining problems. I have suggested that adverbial theories, sense-datum theories, and the theory of appearing provide plausible accounts of perception, though I consider some version of the first kind prima facie best and I leave open that some theory different from all of them may be better than any. Among the further kinds of problems we should explore are one concerning observation and another concerning the relation of perception to the five senses.

Indirect seeing and delayed perception

Observing something in a mirror can count as seeing it. Indeed, it illustrates the sort of thing ordinarily considered seeing something indirectly, as opposed to seeing it by seeing sense-data. We can also see through telescopes and other instruments of observation, again indirectly. But what if the object is microscopic and colorless, yet appears to us through our lens as gray? Perhaps we see it, but not quite as it is.

If we see a microscopic object at all, however, there must be some respect in which what we see it by is faithful to it or at least represents it by some relation of causal dependence – sometimes called *functional dependency*, since perceptual experience seems to vary as a function of certain changes in the object, as where a bird's moving leftward is reflected in a movement of the image. But what we see a thing *by*, such as color and shape, need not be faithful in all respects. A green field can look black at night; we are nonetheless seeing it, and we can see something move in the field even if its color *and* shape are distorted.

How much correspondence between an object and our sensory impressions representing it to us is required in order for us to see it (or hear it, touch it, and so on)? There may be no answer that is both precise and highly general. The cases vary greatly. Observation of faraway objects poses further problems. Consider seeing the nearest star. It is commonly taken to be about four light years away. Presumably we see it (if at all) only *as it was*. For the sense-datum theory, we have a sense-datum produced by it as it was; on the adverbial view, we are sensing "starly" in the way we would have if we had received the relevant visual stimuli at the time the star produced them. If, however, we see it only as it was, do we literally see it at all or just its traces?

Suppose that unbeknownst to us the star exploded two years ago. Is it not odd to say we now see it at all, as opposed to seeing traces of it (as it was)? The latter view is preferable, on the ground that if we unqualifiedly see something now, it exists now. This point is compatible with the view that even though we may see a thing that exists now only *as* it was, we still literally see it now.

Similar points hold for ordinary seeing, since there is still some temporal gap, and for hearing. But if I can see the field only as it was a fraction of a second ago, can I still know that it is now green? I think so, provided there is no reason to believe its color has suddenly changed. The same is not clear for the star: may we know by sight alone that it exists now, when it would take years to realize that the light that was being emitted is no more? This seems doubtful, but it may depend on how likely it is that a star of the kind in question might have burned out during the period in question. If we knew that such stars normally last billions of years and that this one is only a few million years old, we might plausibly think we know it still exists. It is plain, however, that understanding perception and perceptual knowledge in these sorts of cases is not easy.

Sight and light

We normally regard seeing as intimately connected with light. But must seeing involve light? Suppose you could step into a pitch-dark room and have the experiences you would have if it were fully lighted. The room would thus *look* to you just as it would if fully lighted, and you could find any unobscured object by looking around for it. Wouldn't this show that you can see in the dark? If so, then the presence of light is not strictly necessary for seeing.

The case does not, however, establish quite this much. For seeing is a causal relation, and for all I have said you are just vividly hallucinating precisely the right things rather than seeing them. But suppose you are not hallucinating. Indeed, if someone puts a coin in a box or covers your eyes, you no longer feel that you see the coin. Here, the coin somehow affects your eyes through a mechanism other than light transmission, yet requiring an unobstructed path between the object seen and your eyes. *Now* it begins to seem that you are seeing. You are responding visually to stimuli that causally affect your eyes. Yet their doing so does not depend on the presence of light.

Vision and the eyes

It would not ordinarily occur to one to question whether there is any way (literally) to see without eyes. But suppose that after an accident in which Emma has lost her eyes, a camera is connected to her brain in essentially the way her eyes were. When she points it in a given direction in good light, she has just the visual sensations, say of color and shape, that she would have had by looking with her eyes. Might this not be seeing? Indeed, do we not think of the camera as *functioning* like eyes? If, under the right causal conditions, she gets the right sorts of sensations through her eyes *or* a functional equivalent of them, she is seeing.

But are even "eyes" (or organs functioning like eyes) necessary for seeing? What if I lack "eyes" but can get visual sensations matching the objects in the room by strange radiations they emit? Suppose, for instance, that the sensations are stopped by enclosing the coin in cardboard, and that moving it away from me results in my visual impression's representing a decrease in its size. If no part of the body (other than the brain) is required for the visual impression of the coin, there is no organ plausibly considered a functional equivalent of eyes, but might I not be seeing?

If what is crucial for seeing an object is its producing visual sensations suitably corresponding to it, presumably I am seeing. If seeing requires the use of an eye or equivalent organ, then I am not – unless the brain itself is a visual organ. It is clear enough that I would have *knowledge* of what we might call visual properties, above all colors and shapes. One might call that visual knowledge. But visual knowledge of this kind could be held not to be grounded in seeing, nor acquired through use of any sense organs. For these reasons, it may seem somewhat doubtful whether it must be a kind of *perceptual* knowledge. But a case can surely be made for the visual sensation conception of seeing, as against the organ-of-sight conception.

This case, however, may be challenged: can there be *blind sight*, something that in psychological literature is construed as seeing in the absence of visual sensations? People with blind sight can apparently navigate among obstacles as if they saw them, while they honestly report having no visual sensations. Could this be seeing? We automatically tend to understand such behavior in terms of seeing. The inclination to say that they are seeing is even stronger if light's reaching the eyes is necessary for their avoiding the obstacles. But if the subjects have no visual sensation, it is not clear that we must say this, and I doubt that it would be so. The most we must say is that they seem to *know* where the obstacles are. Knowing through some causal process by which objects produce true beliefs about them is not necessarily perception, and certainly need not be seeing.[31]

It may seem that blind sight is genuine seeing because it produces knowledge of what we conceive as visual propositions, say that a red chair is in one's path. This is what one will say if one holds a purely *epistemic analysis of perception*, roughly the view that to perceive is simply to have non-inferential knowledge of perceptible features of one's environment as a result of the relevant information's being causally conveyed, more or less directly, from them to the mind (e.g. by light rays as opposed to testimony).[32] But surely knowledge of visual properties is possible without vision, for instance by something like sonar. Moreover, even dependence on light does not establish that the process in question is visual: the light could somehow stimulate non-visual mechanisms that convey information about the objects emitting it. Similar questions arise for the importance of sensations to perception in the other sensory modes, for instance of auditory sensations in hearing. There, too, we find hard questions for which competing answers are plausible. But there is no good reason to countenance perception where the subject is not "sensible to" the object in the way we are when we see or hear it in the ordinary senses of these terms. There are cognitive responses to the world that have much in common with perceptual experience; but rather than attenuate our concept of perception it is preferable to note their similarities to it and to retain the idea that to perceive something entails its entering one's experience in some sensory mode.

I.10 Interoception and Extra-sensory Perception

Supposing it is true that perception requires having a sensory experience caused by the perceived object, how are we to account for interoception, as where we "perceive" the position of our limbs or the (directly) the beating of our hearts? And in what sense might there be extra-sensory perception?

Interoception

If I non-inferentially know my left arm is behind me, and I do not know it by sight or by any of the five senses, is my knowledge perceptual? I see no reason not to say so, provided the position of the limb causes me to have an experience that is sufficiently like an experience through one of the five senses to be considered sensory. And does the relevant experience differ any more from, say, one of color than an experience of color does from one of taste? There seems to be no necessity to think so. It seems clear, however, that the boundaries of sensory experience are quite vague and rather fluid.

Perhaps the concept of perception is so closely connected with a phenomenal state that we take to be "directly" caused by an object of which it can give us non-inferential knowledge that we are willing to construe almost any kind of phenomenal state as central for some kind of perception or other. This seems so, at least, for kinds of states that represent – in some appropriate way – types of objects that are themselves perceptible in the standard sense in which this implies perceivability through at least one of the five senses. There are doubtless limits here, but notice that we think of dry ice as "burning" the skin, which illustrates that there can be a considerable disparity between the typical sensations a cold thing produces and others by which it may be perceived. Is it necessary, however, that perception produce a phenomenal state of a distinctive sort? This question brings us to the possibility of extra-sensory perception.

Extra-sensory perception

There is no a priori reason why we cannot acquire non-inferential knowledge of anything capable of causing the required true belief in the right way, say with sufficient reliability or with some other appropriate epistemic ground.[33] But if the basis of the belief is truly extra-sensory, then I doubt whether we should consider the relevant grounding of knowledge a case of perception. I believe that telepathic knowledge, if such there be, is said to be based on extra-sensory perception mainly because the relevant phenomenal states are not "five-sensory", as we might say. But there is usually supposed to be (for instance) some voice or appearance that grounds the knowledge, and this is surely a phenomenal state. If all we have is the belief constituting knowledge and it is not grounded in the subject by *any* phenomenal state, then we should posit a causal mechanism that bypasses consciousness. I do not see that such a mechanism could not produce knowledge,[34] but there is insufficient reason to call any such knowledge perceptual.[35]

Let us focus on the important special case in which the knowledge is not of something perceptible in the ordinary sense. Religious experience of spiritual realities would be an example, and there are plausible arguments for construing some of this experience as perceptual.[36] In my judgment, if there is a distinctive kind of phenomenal state that, in the right causal way, is produced by the object – even if it is, say, another mind – and if the relevant state grounds non-inferential beliefs in roughly the way visual states (for instance) do, we may speak of perception. It is true that there is not the usual presumption that perception is the kind of experience that enables one to get around efficiently in the physical world. But this feature of ordinary perception may be insufficiently important to do more than require calling perceptions of non-observables a special case. In any event, it is not obvious that such perceptions could not have at least a potential role in helping us navigate the physical world. This would depend on their content.[37]

I.11 The Perception of Value

Given the frequency with which we quite properly use terms like 'see', 'perceive', 'notice', and 'feel' in reference to what appears to be a recognition of some value property, something should be said here about the possibility of perceiving such properties. I want to consider (briefly) just two cases, the moral and the aesthetic (and I assume here that there *are* moral and aesthetic properties, since otherwise perceptual talk in these domains is at best non-cognitive).

Plainly, one can see *that* an act is wrong. But 'see that' applies in abstract matters, such as realizing that a proposition is logically true, and its applicability in moral cases shows nothing significant about moral perception. It is more important to consider objectual uses of 'see'. Surely one can see the immorality of a deed or, in the aesthetic case, hear the beauty of a lyrical passage in Mozart. It is clear, however, that seeing the immorality of a deed depends on seeing the wrong-doing – say, the snatching of an old woman's purse – and that hearing beauty requires hearing the passage that exhibits it. In the light of this dependency, many are inclined to say that moral and aesthetic properties are not themselves perceptual, but can often be attributed so readily upon perceiving certain perceptual properties related to them that it is natural to use terms like 'see' and 'hear' in ascribing the former. We can "just see" the immorality of snatching the purse, though we see it *by* seeing the nasty deed. And, whether or not there is any inference, our grounds for the moral attribution seem as good as, and much like, our grounds for attributing the property of being a tree on the basis of seeing a trunk, branches, and leaves.

In the case of arboreal perception, however, the crucial property, *being a tree*, is physical, and the parts of a tree are all physical, whereas *being immoral* is not physical, and the perceptible grounds for its attribution are not parts of immoral deeds. Similar points can be made regarding the property of being beautiful (there is, to be sure, a better case for its being physical in music or painting, for instance, but it is surely not physical in poetry). Moreover, it seems that the relevant physical properties are seen directly, or at least are not seen by seeing properties of any other kind. There is, then, an important contrast with the moral and aesthetic cases.

If moral and aesthetic properties are not *causal*, this may be seen as a further reason for denying that they are perceptual. Is it the immorality of the purse-snatching or the beauty of the music that causes our moral or aesthetic perception, or the underlying properties (the "subvenient" ones in some terminologies) that do this? I am prepared to grant, at least for the sake of argument, that it is the latter that have causal power and that the value properties do not themselves have it. This may imply that value properties are not perceptual and indeed are not even "natural" properties. There is surely some reason to hold both views.[38] If they are true, then our conclusion should be that what is properly called the perception of value essentially requires perception, but is not strictly a case of it.

This conclusion would not imply that perceptions of value do not share important similarities with ordinary perception, such as grounding justification and knowledge in an apparently non-inferential way. One might, to be sure, infer the immorality of a deed from its being a theft from a helpless old woman, but one might also infer that an object is a tree from its having a trunk, branches and foliage. The possibility of such inferences does not imply that they are part of our perceptual route to knowledge of the propositions they are available to support. Perhaps the most plausible view here is that, in the common cases in which there appears to be perception of value, there is perception of its basis, or a sufficient part of its basis, in non-valuational perceptual properties.

I.12 Conclusion to Part I

It will now be apparent that it is difficult to provide an overall philosophical account of just what seeing, or perception in general, is; and although all the theories we have discussed can help in answering our questions about perception and perceptual knowledge, none does so in such a simple and decisive way as to leave all its competitors without some plausibility. Still, in exploring those theories we have seen many important points about perception. It is a kind of causal relation. Even its least complex and apparently most basic form, simple perceiving, it requires, in addition to the perceiver, both an object of perception and a sensory experience that in some way corresponds to that object and records, if only imagistically, an indefinite and possibly quite extensive amount of information about the object. Partly on the basis of this information, perception tends to produce beliefs about the perceived object. It implies that the perceiver at least normally has justification for certain beliefs about the object, and it normally produces both justified beliefs about that object and knowledge of it.

Perception may be illusory, as where something appears to have a property it lacks. Perception – or, properly speaking, sensory experience that seems to the subject just like it – may also be hallucinatory, as in the case of Macbeth's vision of a dagger. When it is hallucinatory, the question arises whether there must be interior objects, sense-data, with which the subject is directly acquainted. But both illusions and hallucinations can apparently be accounted for without positing sense-data, and thus without adding a further kind of element to the four that seem central in perception – the perceiver, the object perceived, the sensory experience, and the causal relation between the object and perceiver in virtue of which that experience is produced – or reducing perceptual objects to sense-data. Illusion and hallucination can also be accounted for, I think, without denying that perceptual experience – possibly including some cases of extra-sensory perception – can yield justified belief and knowledge about the world outside the perceiver. So far, we have seen no reason to doubt that perception is a rich and basic source of both knowledge and justification.

II CONSCIOUSNESS AND THE PERCEPTION OF THE INTERNAL WORLD

So far, I have talked mainly about beliefs regarding external things, such as the green field before me. But there is much that we believe about our own minds. I believe that I am *thinking* about self-knowledge, that I am *imaging* cool blue waters, and that I *believe* I am a conscientious citizen. Are some of these self-directed beliefs products of a kind of inner perception? This seems a natural view. If there is truth in it, then exploring the analogy between ordinary perception and self-consciousness might help to explain how such beliefs are justified or constitute knowledge. Let us start by describing the kinds of mental properties illustrated by thinking, imaging, and believing.

II.1 Two Basic Kinds of Mental Properties

Thinking is a kind of process and involves a sequence of events. Thinking is constituted by what it is natural to call *mental events*, such as considering a proposition. By contrast, simply *having* an image, in the minimal way one does when there is a static, changeless picture in the mind's eye, is (I assume) being in a certain (mental) *state*. Unlike something that changes, such a state does not absolutely require the occurrence of any events. *Imaging* can be a process of calling up a succession of images or, as when one of them is held changeless in the imagination, static. I could image something for a time without any change whatever in my imaging, and without the occurrence of any mental event constituting part of the imaging.

Believing could also be called a mental state; but this terminology can mislead in suggesting that having a belief is a state of mind, where that implies a global mental condition like worry. Unlike images and aroused emotions like jubilation, beliefs do not tend to crowd one another out. Beliefs differ from images in at least two further ways. First, beliefs need not be in consciousness. We have many which, unlike the belief that I am now reading, we cannot call to mind without some effort. Second, believing need not be "pictorial". Consider a belief present in consciousness, say that 7 + 5 = 12 is. This belief is not pictorial, and it is present because I have called it to my attention; I had it before doing so.

Even a belief present in consciousness in this way *and* about something as readily picturable as the Statue of Liberty need not involve anything pictorial in the way imaging must. Suppose I believe that the Statue of Liberty has a majestic beauty standing high in the Bay of New York. Without picturing anything, I can entertain this proposition, and in that way have this belief present in my consciousness. By contrast, imaging cool blue waters requires picturing a blue surface. To be sure, when we *call up* this belief about the statue, we tend to picture it. But I could later get the proposition in mind, as where I use it in constructing a logic-book exercise, without picturing anything.

It will help if we observe a standard distinction. Let's call mental properties like beliefs *dispositional* and mental processes like thinking *occurrent*. The latter are constituted by mental events and are occurrences: they take place in the way events do and may be said to happen or to go on. The former are not occurrences and may not be said to happen, take place, or go on. The basic contrast is this. To have a dispositional property or (perhaps not quite equivalently) to be in a dispositional state is to be disposed – roughly, to tend – to do or undergo something under certain conditions, but not necessarily to be actually doing or undergoing or experiencing something or changing in any way. Thus, my believing that I am a conscientious citizen is, in part, my being disposed to say that I am one, under conditions that elicit that sort of verbal manifestation of my belief, such as your asking whether I intend to vote. Yet I can have this belief without doing or undergoing anything connected with it, just as sugar can be soluble in water while it is still in an unaltered lump. I can have the belief even in dreamless sleep. By contrast, to have an occurrent property *is* to be doing, undergoing, or experiencing something, as sugar undergoes the process of dissolving. Thus, if you are thinking about self-consciousness you are doing something, even if you are in an armchair; and if you are imaging a Rose of Sharon, you are experiencing something, at least in the sense that your imaging the shrub is now in your consciousness as a feature of your experience.

Having a static image, however, as opposed to *calling up* an image, is not a process as, for example, silently talking to oneself is. Occurrent mental properties, then, must be subdivided. To differentiate them, we might call occurrent mental properties like thinking *experiential process properties* and occurrent mental properties like having a static image *experiential state properties*.[39] Both differ from dispositional properties. All three kinds of mental properties are important for understanding the epistemology of introspection.[40]

II.2 Introspection and Inward Vision

If we take a cue from the etymology of 'introspection' (from the Latin *introspicere,* 'to look within'), we might construe introspection as attending to one's own consciousness and, when one's mind is not blank, thereby achieving inner seeing. I might introspect my images, for instance, and conclude that my image of my spruce tree indicates that it is taller than the neighboring maple. If introspective consciousness does produce inner seeing and other sensuous imagery (such as, commonly, sound), we can try to understand it by drawing on what we know about perception. For instance, we can explore introspectional counterparts of some theories of perception and sensory experience. But one limitation of that procedure is apparent the moment we reflect on the dispositional mental properties, for instance believing, wanting, and having a fear of cancer. We do not see such properties in any sensory way, as we see (in our mind's eye) an image of cool blue waters. Wants are not seen, not even in our mind's eye.

The analogy to vision might, however, still hold for introspection regarding *occurrent* mental properties. If it does, it presumably applies only to the mental state properties, like imaging. For surely thinking is not seen. It need not even be heard in the mind's ear. I may hear my silent recitation of Shelley's "Ozymandias", but thinking *need* not occur in inner speech. Perhaps only pictorial mental properties, such as those that are objects of imaging, are seen through inner vision; and perhaps it is only a kind of *sensory* property (those that are objects of inner sense), such as inner recitations, tactual imagings (say, of the coldness of a glass), and the like that seem accessible to inner analogues of perception: hearing in the mind's ear, touching in the tactual imagination, and so on. It is doubtful, then, that we can go very far conceiving introspection as simply producing inward seeing. Still, it is worth exploring how the analogy to seeing holds up for pictorial properties.

II.3 Some Theories of Introspective Consciousness

Suppose that introspecting such things as images of cool blue waters does produce a kind of inner seeing. Are we to understand this seeing on realist lines, so that there must be some real object, such as a sense-datum, that is seen by the introspective eye?

Realism about the objects of introspection

One might think that the sense-datum view simply cannot be extended to introspection. This is at least a natural assumption about self-understanding. For on the introspectional counterpart of the sense-datum view, seeing an image of cool blue waters would require something like *another* image, one that represents the first image in the way sense-data represent a physical object seen by virtue of the perceiver's acquaintance with them. Call it a *second-order image,* since it is an image of an image.

What would second-order images be like? If I try to have an image of my image of cool blue waters, I either get that very image again, or an image of something else, or something that is not an image, such as a *thought* of my original image. But perhaps there could be second-order images that are less vivid than the originals they picture. A defender of an adverbial account of sensory experience, however, might argue that even when a perceptual imaging is later "copied" in retrospective imagination, there is really just *one* kind of imaging process and that it occurs more vividly in perception than in imagination. Thus, imaging blue waters is simply imaginationally, rather than perceptually, sensing in the way one does upon seeing blue waters – in short, sensing blue-waterly. Since the adverbial view conceives imaging as a way of experiencing rather than as a relation to an object, there *is* no image as an object to be copied.

An adverbial view of introspected objects

On this adverbial view, then, there is no need to posit second-order images to represent first-order (ordinary) mental images to us, and the less vivid imagings which might seem to represent mental images are best construed as less vivid occurrences of the original imaging process. This point does not show that there *cannot* be second-order images. But the adverbial view reduces the inclination to think there are any by suggesting a plausible alternative account of the facts that originally seemed to demand positing second-order images. Chief among these facts is that in recalling an image, one may have a less vivid image which apparently stands to the former as an imaginational image of a scene stands to the sensory image of that scene from which the imaginational image seems copied. The adverbial account of sensory (and other) experience might explain this by interpreting the recalled image, say of blue waters, as *recollectively* sensing blue-waterly, where this is like visually sensing blue-waterly, but less vivid.

Given these and other points, it seems doubtful whether any realist theory of the introspection of images – one that takes them to be objects existing in their own right – can justify a strong analogy between that kind of introspection and ordinary viewing. For it is by no means clear that there *is* any object introspected to serve as the counterpart of an object of ordinary vision. For the adverbial approach to experience, although realism about the (physical) objects of perception is highly plausible, realism about the objects of introspection is not. It is ontologically lavish: mental properties, such as imaging, can adequately represent physical objects in our mental life; inner objects should not be postulated for this task.

The anti-realism of this view should not be exaggerated. That mental images are not objects having their own properties, and in that sense are not real, does not entail that *imaging* is not real. Imaging processes are real properties of persons, even though they are apparently not relations between persons and objects of inner perception. This is not to deny that introspection has an object in the sense of something it is *of*, such as imaging trees. But on the adverbial view of introspection, this kind of object is *intentional* and is determined by the content of the introspection – what it is about – and is not a thing with properties such as colors and shapes.[41]

The analogy between introspection and ordinary perception

The adverbial view in question may seem unable to do justice to the apparently causal character of introspection. There is surely some causal explanation of our being acquainted with, say, imaging green fields rather than imaging the Statue of Liberty when we monitor a daydream of a rural summer holiday. Perhaps such introspective consciousness differs from seeing mainly in what causes the relevant imaging. How might this difference be explained? On the adverbial account of introspection, it may be like simple perception in two ways. First, introspective viewing may imply some kind of causal relation between what is introspected in it, say imaging, and the introspective consciousness of that state or process. Secondly, such viewing may imply a causal relation between the object of introspective knowledge – for instance one's imaging blue waters – and the beliefs constituting this knowledge.

In explaining the analogy between introspection and perception, I want to concentrate on introspective beliefs as compared with perceptual beliefs; we can then understand how introspection, and indeed consciousness in general, can ground justification and knowledge. A major question here is how we can tell whether, in introspecting something, as when we concentrate on our own imaging, the beliefs we thereby form about what we are concentrating on are produced by that very thing, or by some aspect of it, such as its imagined blue color. It is only to the extent that they are so produced that we should expect introspection to ground justification and knowledge in the broadly causal way perception does. Many considerations are relevant here, but let me cite just two sorts.

First of all, it is surely *because* I am imaging cool blue waters that, when I introspectively consider what I am conscious of, I believe that I am imaging them (and am conscious of my imaging them). It is reasonable to take this 'because' to express a causal relation. If the cause is not some inner object seen (as the sense-datum theory holds), it is presumably the state or process of imaging. This is, in any event, how the adverbial theory of sensory experience would view the causal relations. Similarly, if I introspectively believe that I am thinking about introspection, I believe this because I *am* thinking about it. In both cases the introspective beliefs are produced by inner processes, and indeed in a way that makes it plausible to consider them true.

A second point is this. Suppose my believing that I am imaging cool blue waters is not caused by my imaging them (and that I am not doing so). The belief is then *not* introspective at all. It is *about* what is introspectable, but it is not grounded in introspection, any more than a belief merely about a perceptible, such as the rich red in a painting in a faraway museum, is a perceptual belief. Here, then, is another important similarity between introspection and ordinary perception. The beliefs characteristically grounded by each are identified not by their subject matter, but by their causal basis.

Introspective beliefs, beliefs about introspectables, and fallibility

It may seem that the case described – believing one is imaging something, when in fact one is not – is impossible. But suppose I am asked to image cool blue waters, yet I hate the water and anyway have a lot on my mind. Still, if I want to be cooperative, then even though my mind is mainly on my problems, I may call up an image. Since I am not concentrating on calling up the image, however, the image that I actually get might be only of a blue surface, not of blue waters. I might now inattentively assume (and thereby come to believe) that I have called up the requested image of cool blue waters. This belief is produced by a combination of my calling up the wrong image, which I do not attentively introspect at all, and by non-imaginational factors such as my desire to cooperate. I might even retain the belief for at least some moments after I cease to image at all. In that case, it is neither true nor introspective.

This example suggests that even a true belief about one's conscious states or processes would not be introspective without being causally connected with them. It would be about these introspectable elements but not grounded in "seeing" them. Other examples support the same point. Imagine that my task is to think about introspection for an hour. I monitor myself and, on the basis of introspection, conclude from time to time that I am thinking about introspection. As I reflect on my topic, I continue to believe that I am thinking about introspection. Now when I truly believe this simply because I have repeatedly confirmed it and am confident of steady concentration, and *not* because I am still monitoring myself introspectively, my belief, though perfectly true, is not introspective.

The best explanation of this point seems, again, to be that my belief is not caused (in the right way, at least) by the thinking that should be its ground. It is a retained belief about my ongoing mental activity; it is not produced by that activity. The language appropriate to perception is appropriate here too: my belief that I am thinking about introspection is a propositional belief – a belief *that* I am presently doing so – but it is not an objectual belief, regarding my present thinking, to the effect that it is about introspection. It is not grounded in my *present* thinking, any more than my belief about the rich red in a painting in a distant museum is grounded in seeing it.

My conclusion here is that although there may be no objects such as sense-data or imaginational copies of them which we introspect, the process by which introspection leads to introspective beliefs, and thereby to knowledge and justified beliefs about one's own mind, is nevertheless causal. Like perception of the outside world, it produces something akin to a sensory impression and, often, beliefs about what seems to be revealed to one by that impression. The causes of introspective beliefs, however, are apparently processes and events in the mind, not objects that reside therein.

II.4 Consciousness and Privileged Access

In the light of what has been said, let us suppose that introspective consciousness is a causal process, though with limited similarities to seeing. Still, if it is a causal process, then we should raise some of the same epistemological questions about it that we raised about perception. For instance, is introspection subject to counterparts of illusion and hallucination?

Infallibility, omniscience, and privileged access

One might think that the inner domain, the subject of introspective beliefs, is a realm about which one cannot make mistakes. Indeed, Hume maintained that since the contents of the mind are known by "consciousness" (by which he meant something at least much like introspection), they must appear in every respect what they are, and be what they appear.[42] Hume's statement suggests two far-reaching claims. One claim – that the contents of the mind must be what they appear to one to be – expresses the idea that introspective consciousness can give us beliefs that cannot be mistaken. The other claim – that, to one who has them, these contents must appear to be what they are – expresses the idea that introspective consciousness is so richly aware of the (introspectable) contents of the mind that it guarantees us knowledge of them.

The first Humean claim suggests a thesis of *infallibility*: one cannot be mistaken in a belief to the effect that one is now in an occurrent mental state (e.g. imaging) or that one is undergoing a mental process (e.g. thinking) or that one is experiencing something (e.g. pain).The infallibility thesis rests largely on the idea that we are in such a strong position regarding occurrent mental phenomena that we cannot err in thinking they are going on inside us. The second Humean claim suggests a thesis of *omniscience* with respect to the current contents of consciousness: if one is in an occurrent mental state, undergoing a mental process, or experiencing something, one cannot fail to know that one is. The omniscience thesis rests largely on the idea that occurrent mental phenomena are so prominent in consciousness that one cannot help knowing of their occurrence.

Together, these two theses constitute the *strong doctrine of privileged access.* The first says that our access to what is (mentally) occurring in us is so good that our beliefs about its present make-up are infallible; there is no risk of error. The second says that our access to it is so good that we cannot fail to know what (mentally) occurs in us; there is no risk of ignorance. It is because no one else is in such a good position to know about our mental life, and because we ourselves are not in such a good position to know about the external world, that it is natural to speak of *privileged* access. The strong doctrine of privileged access is associated not only with Hume but, even more, with Descartes, who is widely taken to maintain it in the *Meditations* (1641), especially in Meditation Two.

Suppose that both the infallibility and omniscience theses are true. Would that rule out inward counterparts of illusion and hallucination? No. For once we distinguish between dispositional beliefs and dispositions to believe (as in Part I), we can see that having illusions and hallucinations does *not* imply having false beliefs or being ignorant in any relevant way. Looking from a sharp angle in a line from corner to corner, you can see a book as having the shape of a parallelogram, without believing that it has that shape or even being ignorant of its actual shape.

Suppose, on the other hand, that there *are* no inner objects, such as blue, watery images, to appear to us to have properties they do not possess, such as wavy surfaces. If not, then illusions of the kind we have in perception, in which an object appears to have properties it actually lacks, cannot occur. Nor can a hallucination of, say, an image of blue waters be *of* such an object and true or false *to* it. Suppose, however, that there are inner objects that we see when we image. What would be the difference between hallucinating an image of, say, a loved one, and just *having* that image? A sense-datum theorist might hold that the hallucinatory image would be less vivid or less stable than a real one. But it is still an image of the same thing and might also be just like a normal image in other respects. It would be wrong to say, then, that an hallucinatory image is simply a less vivid or unstable version of a normal image, and the difficulty of explaining the difference between hallucinatory and real images is an additional reason to avoid (as the adverbial view does) positing mental images as objects.[43]

Difficulties for the thesis of privileged access

Quite apart from illusion or hallucination, perhaps we can have false beliefs, or suffer some degree of ignorance, about our mental life. I think this is clear for *some* mental phenomena, such as dispositions like believing, wanting, and fearing. We can mistakenly believe that we do not have a certain ignoble desire (say, to make a fool of a pretentious friend), particularly if it is important to our self-image that we see ourselves as having only righteous desires. For the same reasons, we can fail to know that we *do* have the desire. One can also discover a fear which, previously, one quite honestly disavowed because it was at odds with one's sense of oneself as courageous.[44]

Dispositions, however, should not be conceived as *occurring* in us, and in any case it is occurrent mental phenomena to which philosophers have tended to think we have the kind of privileged access expressed in the theses of infallibility and omniscience. Can we be mistaken, or at least ignorant, about our occurrent mental states or processes?

Consider first the possibility of mistake. Could one believe one is thinking about the concept of introspection when one is only daydreaming about the images and feelings one might introspect? It would seem so, provided one does not attend closely to what is occurring within oneself. This would be a bit like thinking one is watching someone else's observing a game but getting preoccupied with the game itself and ceasing to pay attention to its observer. But suppose the infallibility thesis is restricted to beliefs based on *attentive* introspection, where this implies "looking" closely. Call this the *restricted infallibility view.*

If I carefully consider the proposition that I am thinking about the concept of introspection, and I believe it on the basis of attentive introspection (that is, on the basis of my carefully focusing on the relevant aspect of my consciousness), could this belief be mistaken? This seems doubtful. But is it impossible? Suppose I desperately want to believe that I am doing such thinking. Could this not lead me to take my daydreaming about imaging to be such thinking and even to have an attentive introspective belief that I am doing such thinking? It seems so. Similarly, I could believe, on the basis of attentive but imperfect introspection, that I am imaging an octagon and then, concentrating harder and counting sides, discover that the figure has only seven.

If it is possible to be mistaken in believing that one is now in a particular occurrent mental state (such as thinking), then the omniscience thesis of privileged access should also be abandoned along with the infallibility view. This holds even if the omniscience thesis, too, is restricted, as it should be, to cases of carefully attending to consciousness. The easiest way to see why fallibility cuts against omniscience is to note how omniscience would tend to guarantee *in*fallibility and so would be cast in doubt if the latter is. Let me explain. Given the extensive self-knowledge implied by omniscience, if I am daydreaming rather than thinking about the nature of introspection, then I must know that I am daydreaming. But then I will presumably not be so foolish as *also* to believe that I am thinking about introspection – something plainly incompatible. Since I would know as well that I am occupied with, say, a series of images that portray me as swimming in cool blue waters, it is even less likely that I will believe I am thinking about introspection. It appears, then, that if I am omniscient about my consciousness, then I presumably cannot believe any falsehood about it, and so am infallible about it.[45]

It is at best unlikely (though not impossible) that these two things – knowing every truth about one's consciousness and nonetheless believing some falsehood about it – coincide, leaving one omniscient regarding one's own consciousness, yet inconsistent and fallible about it. One would know every truth about it yet would also somehow believe falsehoods incompatible with those truths. This being at best improbable, if I am fallible I am at least very likely not omniscient. Now recall our daydreaming example. It casts doubt even on the restricted thesis of omniscience. In that case, while I am in fact daydreaming, I would presumably not know that I am. If I did know that I am daydreaming, I would believe this, and then it is very doubtful that I would *also* believe I am thinking about the concept of introspection.

These points suggest that, contrary to the thesis of omniscience, I can fail to know certain things about my consciousness even when I am attending to it; but they do not imply that the omniscience side of the privileged access view is wildly mistaken, in that I might be ignorant of *every* truth about my daydreaming. Far from it. Since I (objectually) believe it to be thinking about introspection, I presumably at least know my daydreaming to involve words or colors or shapes.

The possibility of scientific grounds for rejecting privileged access

It may help to point out that there could someday be a source of significant evidence against various doctrines of privileged access. For it could turn out that every occurrent mental phenomenon is uniquely correlated with some distinct brain process. Then someone could devise a "cerebroscope" for viewing the brain and could read off the contents of consciousness from the cerebroscopic data. What would guarantee that our introspective beliefs must match what the machine says about our mental lives? Imagine that we could discover cerebroscopically a unique neural pattern for, say, believing on the basis of attentive introspection that one is imaging cool blue waters, at the same time as we discover the pattern for imaging only a field of blue-green grass. It would be natural here to suppose the subject is mistaking the grassy image (or imaging process) for a watery one. Might we not regard the sophisticated equipment as more likely to be right than the subject?

There is a problem here. How could one *establish* the unique correlations except by relying on people's introspective beliefs? Wouldn't it be necessary to start by *asking* people what they are, say, imaging, to assume they are correct, and only *then* record the associated brain state? And if learning the correlations would depend on the accuracy of introspective reports, how could the correlations show such reports to be mistaken? A possible reply is this. First, let us grant for the sake of argument that learning the correlations would depend on the accuracy of introspective reports. Still, neuroscientists would not have had to rely on the accuracy of precisely the introspective belief being shown to be mistaken, and perhaps not even on the accuracy of highly similar beliefs. In any event, once they construct their instrument, they might no longer consult introspection. They might throw away the very ladder they have climbed up on.

Imagine, however, that they did have to rely on just the sorts of belief we are examining. Would this imply that the cerebroscope could not provide powerful evidence against introspective beliefs? Consider an analogy. We might use a mercury thermometer to construct a gas thermometer. We might calibrate a container of gas with a piston that rises and falls as the gas is heated and cooled. The new temperature readings might correlate perfectly with mercury readings in many instances: in measuring water temperature, wood temperature, and other cases. The gas thermometer might then do the same jobs as the mercury thermometer *and* might gauge temperatures that the mercury thermometer cannot measure, say because they are above the boiling point of mercury. Could we not use a gas thermometer to correct a mercury thermometer in some cases, or perhaps to correct all mercury thermometers in restricted ways? We could. This seems so even if we had originally taken the mercury thermometer to be infallible in measuring temperature, perhaps because we mistakenly thought of its readings as partly defining what temperature *is.* We can also rebuild the ladder we have climbed up on.

Similar points might hold for beliefs about what is now occurring in one. If the analogy does extend this far – if the gas thermometer is to the mercury thermometer rather as the cerebroscope is to sincere testimony about one's current mental life – then even the restricted omniscience view fares no better than the restricted infallibility view. For even when one is attentive to what is occurring internally, a cerebroscope could indicate that one does not believe (hence does not know) that a certain thing is occurring, such as a frightening image which one thinks one has put out of mind.

II.5 Introspective Consciousness as a Source of Justification and Knowledge

It is important not to overextend our criticism of various claims of privileged access. After all, even the restricted infallibility and omniscience views are very strong claims of privileged access. Giving them up is quite consistent with holding that our access to what is occurring in us is very privileged indeed. Let us explore the extent of this privilege.

The range of introspective knowledge and justification

Nothing I have said undermines a qualified epistemic principle: that our attentively formed introspective beliefs about what is now occurring in us are *normally* true and constitute knowledge.[46] The difficulty of finding grounds for thinking they even *could* be false provides some reason to consider them at least very likely correct. Similarly, when we are attentive to what is occurring in us, then if something (knowable) *is* occurring, such as a certain melody in the mind's ear, *normally* we know that it is occurring, or at least one is in a position to know this by attentively forming the belief that the melody is going through one's mind.

Granted, our "access" to our dispositional properties is not as good as our access to what is occurring in us. We need not be conscious of the former properties, whereas the very existence of one's imaging (or of an image if there are such objects) *consists in* its place in consciousness. Beliefs and other mental dispositions need not even enter consciousness, nor ever be a subject of thoughts or concerns. Some of them may indeed be "repressed", so that we normally cannot easily become aware of them.[47] Nevertheless, it is quite plausible to maintain – and here is a justification principle – that our beliefs to the effect that we are now in a dispositional mental state, for instance now want, fear, intend, or believe something, are normally justified. We might also say that such beliefs, though defeasibly justified, are prima facie justified, so that they are justified overall unless some defeating factor, such as an abnormal psychological interference, occurs. Moreover, it is also quite plausible to hold that normally, when we have a want (or fear, intention, belief, or similar disposition) we are in a position to know (and justifiedly believe) this. We can, then, usually know this if we need to. We very commonly do *not* know it, however; for such things may not enter consciousness at all, and there is often no reason to take any notice of them or form any beliefs about them.

There are a great many issues and details I have not mentioned; but if what I have said is correct, we can now generalize about introspection (roughly, consciousness turned toward one's own mind) in relation to belief, justification, and knowledge, and summarize our main epistemological conclusions regarding inner perception. Plainly, many beliefs arise from introspection, and the points that have emerged suggest an epistemic principle which, though much weaker than the infallibility thesis, is far-reaching: normally, beliefs grounded in attentive introspection (what we have been calling introspective beliefs) are true and constitute knowledge. A second epistemic principle, though far weaker than the omniscience thesis, is that normally, if I attentively focus introspectively on something going on in me, I know that it is going on, under at least some description: I may not know that I am humming the slow movement of Beethoven's *Pathetique* Sonata, but I do know I am humming a melodic piano piece. The corresponding justification principles suggested by our discussion seem at least equally plausible: normally, beliefs grounded in attentive introspection are justified; and normally, if I attentively focus on something going on in me, I am justified in believing that it is going on in me.

There are many possible principles regarding our justification and knowledge about ourselves, and there are many possible qualifications of the four just stated. But those four principles are sufficient to suggest the power of introspection as a source of justification and knowledge. The examples I used to argue that introspection is fallible do not show that the apparently false introspective beliefs were *unjustified* or that true ones are not knowledge. A false belief, particularly if it is of a kind usually justified, can still be justified; and a true belief of a kind that can sometimes be false may itself constitute knowledge.[48]

The defeasibility of introspective justification

These points about the degree of privileged access we apparently do have may create a danger of overestimating the strength of introspective justification. From our examples, it might be thought that attentive introspection, even if not absolutely infallible, generates a kind of justification that at least cannot be defeated.

How could I fail to be justified in believing that I am imaging cool blue waters, if my belief is grounded in attentive introspection? If the question seems rhetorical, this may be because one thinks that there simply is nothing else I should have done besides attending and hence that there can be no possible defeaters of my justification by appeal to the results of some other kind of ground for belief. Let us explore this.

Granting that I could not fail to be justified *unless* I could have good reason to believe I may be mistaken, still, perhaps I could have such reason, for instance repeated cerebroscopic results indicating that I have erred in many quite similar cases. It is far from obvious that I could not have sufficient evidence of this sort. It seems wisest, then, to conclude that although introspective justification tends to be very strong, it remains prima facie rather than absolute and can be defeated by counterevidence.

In any case, plainly beliefs grounded in attentive introspection, such as my belief that I am now imaging blue waters, are normally justified to a very high degree. Moreover – and here we have still another justification principle – normally, my simply being engaged in attentive introspection also yields situational justification for beliefs about what I am attending to, even where it does not in fact yield any such beliefs. If I somehow "notice" my imaging blue waters yet do not form the belief that I am doing so, I am nonetheless (prima facie) justified *in* believing that I am, just as, even if I take no special notice of a bird I see fly past, I am still justified in believing it flew past. The analogy to outer perception seems sound here, and that is one reason why introspection is considered a kind of inner observation and (unless it somehow yields no content, as where the mind is utterly blank) a kind of inner perception[49].

Consciousness as a basic source

If we now ask whether consciousness, including especially introspective consciousness, is like ordinary perception in being a *basic* source of belief, justification, and knowledge, the answer should be evident. It is. But it may well be that the degree of justification which consciousness (including introspection) generates is greater than the degree generated by ordinary perceptual experience, other things being equal. The special strength of justification on the part of beliefs about elements in consciousness has led some philosophers to think that these beliefs are a kind of foundation for knowledge and for the justification of all other beliefs (and Descartes is often thought to have so regarded introspectively grounded beliefs or knowledge).

There seems to be a further epistemologically significant difference between ordinary perception and consciousness, especially as manifested in introspection, as sources of knowledge and justification. We can by and large introspect *at will* – roughly, just by (sufficiently) wanting to – though we may also do it quite spontaneously; and there is no limit to how many things we can come to know by introspecting, if only because we can, without limit, call up images and construct thoughts. But we cannot perceive at will; and what we can know through perception is limited by what there is outside us to perceive and by external conditions of observation.[50]

Introspective consciousness, then, is unlike perception and memory in enabling us to acquire a considerable amount of knowledge and justification whether external circumstances cooperate or not. Whatever one can "observe" in one's own mind is a possible subject of study, and many of the beliefs we attentively form concerning our mental lives tend to constitute genuine knowledge. Very roughly, introspective consciousness is a substantially *active* faculty; perception is a largely *reactive* faculty. Granting that some content – like sensations of pain – comes into consciousness uninvited, we can very freely *call to mind* both propositional and imagistic content. But normally, sensory content, such as perceptual images, enters our mind only when our senses are *taken*, by our own observational efforts or by contingencies of experience, to it. In the inner world, by sharp contrast with the external world, there is far more at our beck and call. This is perhaps another reason why introspectively grounded beliefs have sometimes seemed to be such good material to serve as foundations for knowledge and justification.

There is a trade-off, however. Through perception, we acquire justified beliefs and knowledge about the external world; without these, we would be unlikely to survive. Through introspection, we acquire justified beliefs and knowledge only about the internal world; with only this, our knowledge and justification would be sadly limited to our own minds. This is not to underplay the importance of the internal world: without good access to it we would have little if any self-knowledge and, for that reason, probably at best shallow knowledge of others.

Self-knowledge is also important as a back-up when questions arise about one's justification or knowledge regarding external objects. Confronted with a strange object, one may carefully consider the stability, apparent normality, coherence, and predictable variations of one's perceptual experiences of it in order to rule out hallucination. Told that one merely imagined a car's passing, one may try to recall it and then scrutinize both the vividness of one's imagery and one's confidence that the belief comes from memory rather than merely from imagination. Without the kind of self-knowledge possible here, we would have less knowledge about the external world.

II.6 Some Broad Epistemological Implications

Three very broad closing points are appropriate here to bring out some of the wider epistemological implications of the proposed account of perception. One point concerns the classical idea that the sources of knowledge and justification are experience and reason. The second concerns the foundationalism-coherentism controversy. The third concerns the significance of our results for the issue of skepticism. These are large topics, and my aim is simply to locate this study with respect to them.

Those who think that knowledge and justification are grounded in experience and reason chiefly have in mind the a priori use of reason on the one side and, on the other, perception as a generative source of knowledge and justification, and memory as a preservative source of both. Our results are in broad accord with this conception, but they do not dictate that there *is* any purely a priori knowledge, nor, if there is, do they preclude simply distinguishing two kinds of experience, the ratiocinative and, on the other hand, the perceptual and the memorial. They also do not preclude the possibility of each kind of source constraining the other: the principle of non-contradiction, for example, is plausibly construed as an a priori constraint on what we can know through perception; and the existence of a priori knowledge and justification does not preclude the possibility that a priori (prima facie) justification for a proposition is defeated by empirical justification for believing a proposition incompatible with it.[51]

Second, our results are strictly speaking neutral with respect to the foundationalism-coherentism controversy, though they favor foundationalism insofar as perceptual states are non-cognitive and can justify *as such* rather than by virtue of producing coherence among, say, perceptual beliefs. If my experience of a grassy green expanse can justify my believing there is a green field before me, then coherence among my beliefs is not the ground of my justification, even if it can strengthen that justification. This is not to deny, however, that I could have beliefs that, because this perceptual belief is *incoherent* with them, defeat my perceptual justification, so that, overall, I should suspend judgment or even conclude I am hallucinating. The point that such justification may be defeated by incoherence among beliefs is a major one which coherentists insist on; but to say that perceptual justification is vulnerable to defeat by *in*coherence is not to say that it is grounded in coherence. The former point ascribes to perception a negative epistemic dependence on belief; the latter attributes to it a positive epistemic dependence on belief. Counterparts of these points about perception hold for introspection.[52]

The issue of defeasibility brings us to the problem of skepticism. I have certainly spoken as if I take us to have both perceptual knowledge and perceptually justified belief. None of my conceptual points depends on this, however: perception (inner as well as outer) can be the sort of thing I maintain it is whether there is any knowledge or justification or not. Still, in suggesting the plausibility of certain epistemic principles, I am taking the side of common-sense, particularly with respect to justification, which, as portrayed here, is internal in a way knowledge is not. If justification is internal in the suggested way, and especially if the kinds of justification principles I have suggested are a priori, then there would seem to be good philosophical grounds at least to construe skepticism about perceptual justification as not established and indeed as deserving some degree of doubt.[53] Whether there are such grounds or not, the account of perception presented here provides a number of points at which the issue of skepticism and many other epistemological problems can be usefully focused.[54]

Robert Audi
University of Nebraska, Lincoln

NOTES

[1] In terminology common in epistemology, objectual belief is *de re* – of the thing – whereas propositional belief is *de dicto* – of the proposition; and I similarly distinguish between objectual and propositional perception. The objectual cases, unlike the propositional ones, require no particular concept of the thing in question. To be sure, those who do have the concept of a field and know that I believe it to be rectangular may *say*, 'He believes the field is rectangular', meaning that I believe it to *be* rectangular. English idiom is often permissive in this way, and nothing need turn on the difference in everyday life. Moreover, some philosophers have held that a thing, such as a field, can be a constituent in a proposition, and this might provide a basis for saying that the two belief-ascriptions may be properly interchangeable. Here I ignore that controversial and uncommon conception of a proposition.

[2] Caution is needed here; both language learning and translation are possible despite *some* significant differences in sensory experience. I also ignore skepticism here; one skeptical hypothesis is that unbeknownst to us others do *not* have sensory experiences like ours.

[3] I leave open that Susie could, at least for a moment, believe of a tachistoscope that it is making noise, yet not believe any proposition about it: she *attributes* noise-making to it, yet does not conceptualize it in the way required for having a propositional belief about it, the kind of belief expressed in a complete declarative sentence such as 'The thing on the table is making noise'. She would then have no propositional belief about the instrument, the kind of belief that should be unqualifiedly called true (or false), such as that the tachistoscope is making noise. On this approach, what I am calling objectual belief is better called property attribution.

[4] Even propositional perception can have a kind of openness, even beyond its involving an attribution of a merely generic property, such as *being colored*. One might see that *x* is like *y* without conceptualizing the similarity even in regard to the type of similarity, e.g. facial characteristics.

[5] Specifically, this is a *doxastic* attitude. A fear can be propositional and thereby cognitive, but need not entail believing the proposition feared. Some might consider objectual awareness, say awareness of perfect symmetry, cognitive, at least when the person has the concept of relevant property. By contrast, desires, the paradigm *conative* attitudes, are not generally taken to have propositional objects (e.g. 'to swim', in 'my desire to swim', does not express a truth or falsehood). Perceptions that embody beliefs in the ways illustrated are also called *epistemic*, since the embedded belief is commonly considered to constitute knowledge. Their connection will knowledge is pursued below.

[6] The distinction between simple and propositional perceiving and others drawn in this chapter are not always observed. At one point W. V. Quine says, "think of '*x* perceives *y*' rather in the image of '*x* perceives that *p*'. We say 'Tom perceives the bowl' because in emphasizing Tom's situation we fancy ourselves volunteering the observation sentence 'Bowl' rather than 'Surface of a bowl', 'Front half of a bowl', 'Bowl and background'... When we ask 'What did he perceive?' we are content with an answer of the form 'He perceived that *p*'. See *Pursuit of Truth,* revised edn. (Cambridge; Harvard University Press, 1992), p. 65. Notice that since seeing that (say) there is a bowl before one obviously entails seeing a bowl, it is no surprise that we are content with a report of the propositional perception even if we wanted to know only what object was seen. It does not follow that simple seeing *is* or even entails propositional seeing. It is also worth noting that Quine is apparently thinking of seeing here; for the other four senses, there is less plausibility in maintaining what he does. For a case to show that perception is not necessarily epistemic in the way Quine implies, see Fred Dretske, *Seeing and Knowing* (Chicago: University of Chicago Press, 1969).

[7] The adage should not be applied to simple seeing, for what we simply see, say a glass or leaf or field, is not the sort of thing that can be believed (to be true or false). Seeing something, especially something as striking as gold-ball size hail, does produce a *disposition to believe* certain propositions, say that this is a dangerous storm. But there are many things we are disposed to believe but do not. I have defended these points in detail in "Dispositional Beliefs and Dispositions to Believe," *Nous* 28 (1994), 419-34.

[8] A far more detailed account of the relevant data is given in my "Dispositional Beliefs and Dispositions to Believe."

[9] In the light of what has been said so far we can accommodate much of what is plausible in the common view that, as D. M. Armstrong puts it, perception "is an acquiring of knowledge or belief about our physical environment (including our own body). It is a flow of information. In some cases it may be something less than the acquiring of knowledge or belief, as in the cases where perceptions are entirely discounted or where their content has been confidently anticipated." See *Belief, Truth and Knowledge* (Cambridge: Cambridge University Press, 1973), p. 22. First, I can agree that perception entails acquisition of information; the point is that *not all our information is possessed as the content of a belief.* Second, Armstrong himself notes an important way in which perception might fail to produce belief: it is "discounted," as, e.g., where one is sure one is hallucinating and so resolutely refuses to accept any of the relevant propositions.

[10] This is the kind of view developed in detail by Fred Dretske. See esp. *Knowledge and the Flow of Information* (Cambridge, Mass.: MIT Press, 1981).

[11] Two points are appropriate here. First, what I call situational justification (justification, provided by one's epistemic situation, for believing a proposition, but not entailing that one does believe it) is roughly equivalent to what Roderick Firth called "propositional justification" in "Are Epistemic Concepts Reducible to Ethical Concepts?" in Alvin Goldman and Jaegwon Kim, eds., *Values and Morals* (Dordrecht: D. Reidel, 1978), 215-29; this is a kind of *justification for forming a belief,* which does not entail that any belief is formed on the basis of the justificatory ground. Second, the notion of normality here is not statistical; it

implies that what is not normal is calls for explanation. In the world as we know it, exceptions to the normality generalizations I proposed seem at least quite rare. But the point is not that statistical one; it is to bring out that the very concepts in question, such as those of seeing and knowing, have a connection in virtue of which explanation is called for if what is normally the case does not occur.

[12] The psychological literature contains many examples of things that can be seen in different ways, or even as different things, depending on either perspective or subtle elements of what one might call *interpretation*. The duck-rabbit drawing in Ludwig Wittgenstein's *Philosophical Investigations* (Oxford: Blackwell, 1953) is one famous case: one may see either a duck or a rabbit and may shift from seeing the one to seeing the other. Similarly, the Necker Cube can be seen as oriented differently toward one depending on which side one sees as the front and which the back. An interesting question here is whether what one sees something *as* can affect the intrinsic quality of one's visual experience, as opposed to being a matter of one's beliefs or dispositions to believe, or at least one's dispositions to behave in certain ways toward the object. It could be that such "ambiguous" drawings (or other perceptual objects) cannot be seen *without* being seen as something or other; but even if they cannot, it does not follow that *everything* seen is seen as something or other in the relevant sense.

[13] David Hume, *An Enquiry Concerning Human Understanding*, Eric Steinberg, ed. (Indianapolis: Hackett, 1977), pp. 23 and 61.

[14] In speaking of justification that prevails, and of overall justification, I have in mind the kind appropriate to a rational person's believing the proposition in question, construed as roughly the kind such that, when we believe a true proposition with that kind of justification then (apart from the kinds of cases (sometimes called "Gettier cases") that show how justified true beliefs *need* not constitute knowledge), we know it.

[15] There are complexities I cannot go into, such as how one's competence figures. I am imagining here someone competent to tell whether a note is flat (hence not virtually tone deaf): in general, if one is not competent to tell whether a kind of thing has a property, an experience in which it seems to have it may not justify one in believing it does. There is also the question of *what* the belief is about when the "object" is hallucinatory, a problem discussed shortly. Still other problems raised by this justification principle are discussed in ch. 8 of my *Epistemology* (London and New York: Routledge, 1998) in connection with the controversy between internalism and externalism.

[16] If, as is arguable, seeing that it is blue entails knowing that it is, then he does *not* see that it is, though he sees its blue color. But this entailment claim is far from self-evident.

[17] Locke is a good case of a causal theorist, but H. P. Grice's article on the subject did much to make the term prominent and gave additional appeal to the kind of theory in question. See Grice's "The Causal Theory of Perception," *Proceedings of the Aristotelian Society* Supplementary Volume 35 (1961). As I use the term here, however, there is no suggestion that perception requires the occurrence of sense-data, whereas Locke, Grice, and many other causalists about perception favored sense-datum theories.

[18] The theory of appearing has not been widely defended, but a detailed sympathetic treatment is given in William P. Alston's "Back to the Theory of Appearing," *Philosophical Perspectives* 13 (1999), 181-203. Cf. Roderick M. Chisholm, "The Theory of Appearing," in Max Black, ed., *Philosophical Analysis* (Englewood Cliffs, NJ: Prentice-Hall, 1963).

[19] For a contemporary study and defense of a sense-datum theory see Howard Robinson, *Perception* (London and New York: Routledge, 1994).

[20] See John Locke, *An Essay Concerning Human Understanding* (1689), esp. bks II and IV).

[21] This point is compatible with a great deal of complexity in the processing that occurs between light rays' striking the retina and the subject's having a visual experience. It is, e.g., probably neutral with respect to David Marr's view that, in the psychology of vision, object shapes are derived from images in three stages: (1) primal sketches representing changes in intensity, critical features such as terminal points, and geometrical relations; (2) the "2½-D sketch", that gives a preliminary analysis of depth, surface features, and other visual properties, centered on the viewer; and (3) the 3-D model representation in an object-centered coordinate system, which enables us to see objects three-dimensionally despite their being presented from a single viewpoint. See "Representing and Computing Visual Information", in H. C. Longuet-Higgins and N. S. Sutherland, eds., *The Psychology of Vision* (London, 1980). It should be added that when psychologists speak of *indirect perception* they are not in general implying mediation by what philosophers commonly conceive as inference, though Helmholtz is often credited with holding an inferentialist view of perception. See H. von Helmholtz, *Treatise on Physiological Optics* (1867), trans. by J. P. C. Southall (New York: Dover Publications, 1962). For detailed recent psychological papers on this issue and many other aspects of perception see Irwin Rock, ed., *Indirect Perception* (Cambridge, MA: MIT Press, 1997).

[22] The view that ordinary perceptual belief is non-inferential is controversial and – for various senses of inferences – has been widely discussed by both philosophers and psychologists. Not *all* sense-datum views, moreover, take perceptual belief to be non-inferential. For a discussion of perception that brings to bear both psychological as well as philosophical literature see John Heil, *Perception and Cognition* (Berkeley and Los Angeles: University of California Press, 1983), esp. ch. 2.

[23] "[Irwin] Rock makes the argument that if such sequences of conscious perceptions exist, they constitute a form of *indirect perception* to be contrasted with Gibson's notion of direct perception. His reasoning is that because the latter, higher-level percept is mediated by the earlier one, rather than being mediated by higher-order retinal structure in the optical stimulus, it is not direct and unmediated as Gibson proposed". See Stephen E. Palmer's wide-ranging Foreword to Rock, op. cit., pp. xx-xxi

[24] Granted, the book does not appear to us *to be* parallelogrammic if we realize its shape cannot be judged from how it visually appears at an angle, but that is a different point. It concerns what shape we *take* it to have, not what shape visually appears in our consciousness antecedently to our taking it to be of any particular kind.

[25] For a detailed and influential discussion of the adverbial theory, with criticism of the sense-datum view, see R. M. Chisholm, *Perceiving* (Ithaca: Cornell University Press, 1957).

[26] This is a very important point. One major materialist theory of the mind-body relation – the identity theory – says that mental phenomena are identical with brain states or processes. But this theory fails if sense-data exist as mental entities and have properties, such as being green and rectangular, that no brain process has. Identity theorists thus generally oppose the sense-datum theory. See, e.g., J. J. C. Smart's influential "Sensations and Brain Processes", *Philosophical Review* 68 (1959), 141-56.

[27] These and other problems are brought against the sense-datum theory by Winston H. F. Barnes in "The Myth of Sense-Data", *Proceedings of the Aristotelian Society* 45 (1944-45).

[28] See Berkeley's *Treatise Concerning the Principles of Human Knowledge* (1710).

[29] For detailed Twentieth-Century defense of phenomenalism, see Book II of C. I. Lewis's *An Analysis of Knowledge and Valuation* (La Salle, Illinois: Open Court, 1946); cf. R. M. Chisholm's widely known criticism of this defense in "The Problem of Empiricism", *Journal of Philosophy* 45 (1948).

[30] Berkeley might hold that if *God* has bookish sense-data, it does follow that there really is a book. A case can be made for this, but one might also argue that as an all-powerful being

God could bring it about that there is a distinction between his creating a physical object and his having the corresponding sense-data.

[31] A subject who really *does* have visual impressions could also misreport, a possibility discussed below.

[32] D. M. Armstrong has defended a view of this sort. In *A Materialist Theory of the Mind* (London" Routledge and Kegan Paul, 1968), e.g., he construes perception as the "acquiring of true beliefs" (p. 109), and maintains that it is "a flow of information ... Perceptual *experience*, as opposed to mere perception, is simply this flow insofar as we are conscious of it" (p. 226).

[33] If reliable production or a true belief is not sufficient (under certain conditions) for knowledge, the case for the kind of knowledge described here is more difficult to make, but I believe one could still make it from a largely internalist point of view. For some of the issues and a case for an externalist conception of knowledge, see Armstrong, Dretske, Alston, and *Epistemology*, ch. 8.

[34] As argued in ch 8 of *Epistemology*.

[35] That perception requires experience in a sense that implies instantiating phenomenal properties is, however, not uncontroversial. For epistemic theories of perception – also called *cognitivist* – on which it is fundamentally a kind of acquisition of propositional information, see, e.g., James Gibson, *The Perception of the Visual World* (Boston: Houghton Mifflin, 1950) and D. M. Armstrong, *A Materialist Theory of Mind* (London: Routledge and Kegan Paul, 1968) and *Perception and the Physical World* (London: Penguin Books, 1961). Perhaps the most obvious problem with such views (at least in strong forms) is that they do not provide a good account of the difference between, e.g., seeing what is before one and not seeing it when one closes one's eyes – a striking difference that does not seem to entail any relevant change in beliefs, if any such change is entailed at all.

[36] See e.g., William P. Alston, *Perceiving God* (Ithaca and London, Cornell University Press, 1991).

[37] I leave aside here the issue of whether the mental has causal power; if there can be no mental causation, that would be ground for denying that certain kinds of religious experiences are perceptual. For discussion of the causal powers of the mental see John Heil and Alfred Mele, eds., *Mental Causation* (Oxford and New York: Oxford University Press, 1993).

[38] A number of reasons to hold this (and many relevant references) are given in my "Ethical Naturalism and the Explanatory Power of Moral Concepts", in my *Moral Knowledge*.

[39] To be sure, images can be possessed memorially, as is my image of the Statue of Liberty when I don't have it in mind; and 'imaging' can designate a process, as when I call up the series of images corresponding to looking at the Statue from the Brooklyn Heights Promenade and glancing northward to Lower Manhattan.

[40] Both kinds of properties are experiential, in that they represent features of experience. Both, then, might be considered *phenomenal*, but sometimes the term 'phenomenal property' is restricted to the sensory kind that characterizes either the five senses or "inner sense", by which pain and pleasurable sensations are felt.

[41] Such contentual objects are often called *intentional*, largely on the ground that, like lofty deeds we intend to perform but do not do, they need not exist.

[42] See David Hume, *A Treatise of Human Nature* (first published in 1739-40), Part IV, Section II), ed. by L. A. Selby-Bigge (Oxford: Oxford University Press, 1888).

[43] One might still distinguish between genuine and hallucinatory images by insisting that in order to be a genuine image *of* (say) a loved one, an image *must* be caused by, say, seeing that very person. This view has an odd consequence, however. Through hearing a detailed description I could have an accurate image of Maj that is in a sense *of* her, since it matches her sufficiently well, even if I have never seen her; but this would be a hallucinatory image,

on the causal conception just stated. There are certainly different kinds of images and various ways they can mislead, but the analogy between perception and introspective consciousness does not extend in any simple way to the possibility of inner illusions and hallucinations, and there is no need to pursue the matter in more detail here. For a detailed non-technical discussion of mental imagery see Alastair Hannay, *Mental Images: A Defence* (London: George Allen and Unwin, 1971) and my critical examination of this book in "The Ontological Status of Mental Images", *Inquiry* 21 (1978), 348-61.

[44] Some of these cases seem to occur in *self-deception*, a phenomenon that raises profound questions for both epistemology and the philosophy of mind. For a comprehensive collection of papers on it (including one offering my own account), see Brian P. McLaughlin and Amelia O. Rorty, *Perspectives on Self-Deception* (Berkeley and Los Angeles: University of California Press, 1988).

[45] The thesis of omniscience might be restricted to *introspectable* truths, as opposed to such truths as that there are 1,001 berries visible on the blackberry bush I am imaging, which I could know only on the basis of memory (and arithmetic) as well as introspection. The infallibility thesis might also be plausibly restricted in a similar way. This point bears on the connection between the two theses but should not affect the argumentation in the text.

[46] They can be attentively formed without there being anything in consciousness one is attending to; attentive formation is a matter of one's forming the belief(s) in question through turning one's attention in an appropriate "direction". One need not find anything in that direction.

[47] Repression need not be exactly the kind of thing Sigmund Freud described, requiring psychoanalysis or very special techniques to come to consciousness. There are various kinds and degrees of repression; the point here is simply that *having* a belief (or other dispositional state) is possible even if it is repressed. One might, e.g., still act in the way expected of a believer of the relevant proposition.

[48] Skeptics, of course, tend to deny this, in part on the ground that such a possibility of falsehood implies uncertainty, which is incompatible with knowledge.

[49] Even if the mind is blank, one could think of introspecting as like looking into the dark: this might be thought to be non-seeing, as opposed to seeing blackness, so the analogy between the ordinary and inner kind of perception would hold.

[50.] There is less disanalogy in the negative cases: we cannot always cease at will to concentrate introspectively on our mental life, as illustrated by preoccupying pains; and we cannot, at will, cease perceiving what we do without, e.g., closing our eyes or turning off a radio. This blocks the path of observation, just as an aspirin might block the path of pain.

[51] On the compatibility of the defeasibility of a priori justification with a moderate rationalism, see Laurence BonJour, *In Defense of Pure Reason* (Cambridge and New York, CUP, 1997) and my "Intuitionism, Pluralism, and the Foundations of Ethics", in my *Moral Knowledge and Ethical Character.*

[52] This point and many relevant to the epistemology of perception are defended in my "The Foundationalism Coherentism Controversy: Hardened Stereotypes and Overlapping Theories", in my *The Structure of Justification* (Cambridge and New York: Cambridge University Press, 1993).

[53] For defense of internalism about justification and references to relevant literature, see my *Epistemology*, esp. ch 8, and for discussion of skepticism see ch 10.

[54] For comments on an earlier draft I thank William P. Alston and Dan Crawford. This study draws heavily on chs. 1 and 3 of my *Epistemology*, and I am grateful to the publisher for permitting me to reuse the relevant material.

REFERENCES

Alston, W.: 1963, 'Back to the Theory of Appearing', *Philosophical Perspectives* **13** (1999), 181-203, Cf. R. M. Chisholm, in M. Black, (ed.), *Philosophical Analysis*, Prentice Hall, Englewood Cliffs, NJ.

Alston, W.: 1991, *Perceiving God*, Cornell University Press, Ithaca and London.

Armstrong, D. M.: 1961, *Perception and the Physical World*, Penguin Books, London.

Armstrong, D. M.: 1968, *A Materialist Theory of the Mind*, Routledge and Kegan Paul, London.

Armstrong, D. M.: 1973, *Belief, Truth and Knowledge*, Cambridge University Press, Cambridge.

Audi, R.: 1978, 'The Ontological Status of Mental Images', Inquiry **21**, 348-61.

Audi, R.(ed.): 1993, *The Structure of Justification*, Cambridge University Press, Cambridge and New York.

Audi, R.: 1993, 'The Foundationalism Coherentism Controversy: Hardened Stereotypes and Overlapping Theories', in Audi, *The Structure of Justification.*

Audi, R.: 1994, 'Dispositional Beliefs and Dispositions to Believe,' *Nous* **28**, 419-34.

Audi, R.: 1997, *Moral Knowledge and Ethical Character*, Oxford University Press, New York.

Audi, R.: 1997, 'Intuitionism, Pluralism and the Foundations of Ethics', in Audi, *Moral Knowledge and Ethical Character.*

Audi, R.:1998, *Epistemology*, Routledge, London and New York.

Barnes, W. H. F.: 1944-45, 'The Myth of Sense Data', *Proceedings of the Aristotelian Society* **45**.

BonJour, L.: 1997, *In Defence of Pure Reason*, Cambridge University Press, Cambridge and New York.

Chisholm, R. M.: 1948, 'The Problem of Empirism', *Journal of Philosophy* **45**.

Chisholm, R. M.: 1957, *Perceiving*, Cornell University Press, Ithaca.

Dretske, F.: 1969, *Seeing and Knowing*, University of Chicago Press, Chicago.

Dretske, F.: 1981, *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA.

Firth, R.: 1978, 'Are Epistemic Concepts Reducible to Ethical Concepts?', in A. Goldman and J. Kim (eds.): *Values and Morals*, pp. 215-29, D. Reidel, Dordrecht.

Gibson J.: 1950, *The Perception of the Visual World*, Houghton Mifflin, Boston.

Grice, H. P.: 1961, 'The Causal Theory of Perception', *Proceedings of the Aristotelian Society*, Supplementary Volume **35**.

Hannay, A.: 1971, *Mental Images: A Defence*, George Allen and Unwin, London.

Heil, J.: 1983, Perception and Cognition, University of California Press, Berkeley and Los Angeles.

Heil, J and Mele A. (eds.): 1993, *Mental Causation*, Oxford University Press, Oxford and New York.

Helmholtz, H., von: 1867, *Treatise on Physiological Optics*, transl. by J.P.C. Southall, Dover Publications
1962, New York.

Hume, D.: *An Enquiry Concerning Human Understanding*, in E. Steinberg (ed.) 1977, Hackett, Indianapolis.

Hume D. 1739-40: *A Treatise of Human Nature*, L.A. Selby-Bigge (ed.) 1888, Oxford University Press, Oxford.

Lewis C. I.: 1946, *An Analysis of Knowledge and Valuation,* Open Court, La Salle, Illinois.

Locke, J.: 1689, *An Essay Concerning Human Understanding.*

McLaughlin, B.P. and A.O.Rorty: 1988, *Perspectives on Self-Deception*, University of California Press, Berkeley and Los Angeles.

Marr, D.: 1980, 'Representing and Computing Visual Information', in H.C. Longuet-Higgins and N.S. Sutherland (eds.), *The Psychology of Vision*, London.

Quine, W. V.: 1992, *Pursuit of Truth*, Harvard University Press, Cambridge.

Robinson, H.: 1994, *Perception*, Routledge, London and New York.

Rock, I. (ed.) : 1997, *Indirect Perception*, MIT Press, Cambridge MA.

Smart, J. J. C.: 1959, 'Sensations and Brain Processes', *Philosophical Review* **68**, 141-56.

Wittgenstein, L.: 1953, *Philosophical Investigations*, Blackwell, Oxford.

ELIZABETH FRICKER

TESTIMONY: KNOWING THROUGH BEING TOLD

1. OUR SUBJECT: WHAT IS TESTIMONY?

The expression 'testimony' in everyday usage in English is confined to reports by witnesses or by experts given in a courtroom, or other formal setting. But in analytic philosophy the expression is used as a label for the process by which knowledge or belief is gained from understanding and believing the spoken or written reports of others generally, regardless of setting. In a modern society testimony thus broadly understood is one of the main sources of belief. Very many of an individual's beliefs are gained second-hand: from personal communication, from all sorts of purportedly factual books, from written records of many kinds, and from newspapers, television and the internet. Testimony enables the diffusion of current news, information (or misinformation), opinion and gossip throughout a community with a shared language. It also enables the preservation and passing on of our accumulated heritage of knowledge and belief: in history, geography, the sciences, technology, etc. We would be almost unimaginably epistemically impoverished, without the resources provided by testimony in its various forms.

What are the philosophical issues concerning testimony?

When testimony is trustingly accepted by an individual, she acquires beliefs through it. In a modern society, very many of an individual's beliefs are derived directly from testimony, or depend for their grounding on other beliefs so derived (see sect.8). Are these beliefs derived from testimony ever justified, and apt to be knowledge? The primary concern of philosophy regarding testimony is epistemological: to explain the status as potentially justified and knowledgeable of beliefs dependent on testimony. – Or, if the upshot is skeptical, to show why such beliefs are not apt to be justified and knowledgeable.

This primary concern involves, or overlaps with, others. First: Testimony as an epistemic kind needs to be more precisely delineated, and characterised (sect.2 below). Second: The acquisition of belief through testimony essentially involves understanding the content and force of a speech act made to one as audience (mutatis mutandis for written testimony) (see sect.2.). Thus in testimony we have a locus where epistemology interlocks with philosophy of language. Suppose we say that, strictly, the epistemology of testimony concerns the epistemic status of a hearer H's belief that P, acquired through H being told that P by a testifier T, and H trusting T. (Call beliefs derived from and still grounded in such a source *testimony-beliefs*.) Still, our account of the epistemic status of testimony-beliefs must mesh with our account of a closely related matter: how it is that H understands what she hears,

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 109-130.
© 2004 *Kluwer Academic Publishers.*

what is involved in this, and if – as seems plausible – it entails knowing that she has been told by T that P, how this epistemic feat is achieved. Thus an epistemology of testimony needs to be complemented by an epistemology of understanding. Nor can the latter be completed, without a philosophical account of the nature of meaning. An account of how meanings can be known must interlock with an account of the nature of the objects of this knowledge.[1] Third, an account of how beliefs derived from testimony can be justified, and knowledgeable, cannot be elaborated ad hoc. To convince, it must be the application to this case of a general conception of justified belief, and of the conditions for knowledge. Thus an epistemology of testimony must instantiate a preferred theory in general epistemology. We will see later that recognition of the ubiquitous dependence on past trusted testimony in our belief-system provides pressure towards a coherentist, not foundationalist, account of the justification of our empirical beliefs.

Returning to our primary project, this can be further specified and subdivided. Normative epistemology is one thing, and the plotting of the actual psychology of belief-acquisition through testimony, and the actual facts about the place of testimony-beliefs in our belief system, is another. Normative epistemology will tell us the conditions, if any, under which a belief acquired through testimony could and would be justified, and whether and how a belief system with extensive dependence on testimony can be so. Descriptive psychology will tell us what human belief acquisition through testimony is actually like, and what extent of dependence on testimony our belief systems actually exhibit. Given this distinction, we can divide our central issue about testimony along two dimensions, yielding four distinct questions to investigate, thus:

Descriptive Local Question: How do human hearers typically form belief in response to testimony? In particular, do they just trust their informant unthinkingly, blindly; or do they somehow (consciously, or sub-consciously) evaluate the informant for trustworthiness, and believe what they are told only if the evaluation is positive? (The process of testimony)

Normative Local Question: In what conditions, and with what controls, *should* a mature adult hearer believe what she is told, on some particular occasion? (Fresh instances of testimony , for an adult hearer.)

Descriptive Global Question: What is the actual place of testimony-beliefs overall, in a person's structure of empirical belief? What is the extent of dependence on testimony for grounding (*epistemic dependence*) of our beliefs? And what is the relation between testimony and our other sources of empirical belief: perception, memory, and deductive and inductive inference from empirical premisses?

Normative Global Question: how, if ever, can a system of beliefs with uneliminated epistemic dependence on testimony be justified?

For a philosopher who is ready to accept skeptical conclusions, where they arise from her initial suppositions, these descriptive and the normative issues are distinct. But for one, like myself, who regards it as a datum to which our theorising is answerable, that an epistemically responsible human believer's belief system is, broadly, justified, the philosophical task is to provide an epistemological account of testimony which explains this, rather than challenging it. We should accept the Attainability Constraint as such: take it as a fact that knowledge, and justified belief, can be and sometimes are gained through testimony.[2] The *Attainability Constraint* links the descriptive with the normative: our normative theorising is constrained by it to harmonise with the actual structure of dependence on testimony in our belief system, and with the actual process of testimony, including the psychology of human acquisition of belief through testimony. Thus, even if our driving interest is in normative matters, we had better pay close attention to these facts about the psychology of acquisition of belief through testimony.

The philosophical task in relation to testimony is beginning to look rather large, indeed daunting! In this essay I will not attempt to address in any depth, let alone answer, all the issues just raised. I will help myself to briefly outlined views about understanding, and in general epistemology, in order to set out some central features of the terrain in relation to testimony-beliefs. The sections which follow discuss each of the issues introduced above. I do not resolve all the issues regarding testimony, but sketch a map of them, identifying the key issues to be investigated.

2. DEFINITION: THE SCOPE OF TESTIMONY

How more precisely should our subject matter be delineated? We need not be slaves to ordinary language – we need a good epistemic kind to build our theory around, and we may need to construct it. But in fact the English-language concept of *telling* captures the core of our epistemic kind. It is unsurprising that we already have a concept for the core kind, since we have a folk epistemology, a theory of the various ways in which we can come to know things. Telling ranks in it as a way of gaining knowledge along with seeing, hearing and the other senses, "working it out" (i.e. inference), plus memory as a method of retaining knowledge.[3] It is part of everyday social epistemic life that we ask someone who makes a claim to knowledge: "How do you know that?", and we standardly expect and accept a range of answers: "I saw it", "I remember doing it", "I worked it out", and "He/She told me".

Rather than a sharply-delineated epistemic kind, we find with testimony a central paradigm, telling, and then cases which depart more or less from it, in epistemically relevant features. (In subsequent sections I will concentrate on spoken testimony – tellings and other assertions.) That there are epistemic kinds about which we can fruitfully theorise is a substantial methodological assumption of positive epistemology. Perception, memory, deduction and induction are standardly taken to be such kinds. The present suggestion is that telling, and the cases at its fringe, are another such epistemic kind. The methodological assumption in this case is that there are some illuminating *general* things to be said about how the process of telling yields justified belief and knowledge, when it does; and as a corollary, the circumstances in which beliefs acquired through this process, and still dependent on it for their grounding – what we called testimony-beliefs – are justified or are knowledge. (Compare this with the parallel assumption for perception.) But it is indeed a substantial methodological assumption that this is so – that we can discern a *common* mechanism, set of conditions, by which all tellings yield justified belief or knowledge, when they do so. This initial presumption is to be justified by its results. It guides us in how we delineate our kind (or our paradigm case): we want to define a kind such that the assumption holds for it.

Consider the following schematic predicate:

(U) " ... is a belief of H's such that: H comes to hold ... as a result of, and on the basis of, observing S make As on O." – where H holds the place for a hearer, S for a speaker, As for an assertion by S, and O for an occasion.

This schematic predicate does not pick out an epistemic kind, in our desired sense. There are no interesting generalisations be made about the topic: what one might be able to infer, given one's background knowledge, from observing someone make an assertion on an occasion. This could include: conclusions about the speaker's mental state, about her background, about where she has recently been, and about what has happened in the recent past. There seems no general limit to be set, to what one might be able to infer, and no distinctive general process to be discerned, at this level of inclusiveness. An epistemic kind associated with occasions on which a speaker asserts something to a hearer will be more restricted. A better epistemic kind to pick out is:

(T) All and only instances of someone's coming to believe that P, as a result of perceiving and understanding someone make an assertion that P.

In contrast with (U), (T) places restrictions first, on what the belief is, which is formed as a result of observing the assertion: only a belief in the content that has been asserted is a *testimony-belief*, a belief acquired *through* testimony, in our intended sense. (Though, to repeat: a hearer may acquire many other justified beliefs as a result of observing a piece of testimony – that the speaker is in a bad temper, that she comes from a certain region, and so forth.) Second, (T) restricts the process by which that belief is formed by the hearer: it must be via understanding the speech act she observes, in which it is asserted that P.[4] Our hope is, that there is something illuminating and general to be said, about how justified belief or knowledge can sometimes be acquired through the process of *understanding* what one is told, and *trusting* the teller – believing what she says, on her say-so.

In fact tellings are a sub-class of assertions, and believing what one is told is the central paradigm of the kind picked out by (T).[5] (T) itself picks out the central case of testimony, there being a cluster of fringe cases that depart from it more or less. Listening to the radio, and watching television and films, all furnish fringe instances of testimony, as does reading purportedly factual written material of all kinds. These differ from the central case (T) epistemically, since once the teller is not directly observable by the hearer, the latter's scope for evaluating her trustworthiness (her motives, and competence) is greatly reduced, or at least altered – no perceptual cues to this being available.[6] Amongst assertions, tellings are epistemically special, since evaluating the trustworthiness of an assertion must be via estimating the motives in making it of the speaker, and a teller will have certain motives – if she is sincere, the desire to inform her audience, for some purpose.

(T) does *not* make the following restrictions on what is to count as 'testimony': the fact that what is said is true; that the speaker's intent is not deceptive; that the content of her assertion is of any particular kind, either in itself, or in her relation to it. Thus (T) does not restrict 'testimony' to eye-witness reports, nor specify that the speaker is an expert or for other reasons authoritative about her subject. These would be inapt restrictions on our epistemic kind, since to include them in our definition of testimony would mask, by definitional stop, the epistemic problem that typically confronts the hearer wondering whether to believe what another person tells her. What a hearer will typically be able to perceive to be the case is this: that S has asserted that P to her (told her that P). We need to consider in what circumstances, and with what evaluations and checks, she may, from this initial endowment of perceptually-gained knowledge, justifiedly form belief in what she is told. To build into the definition of testimony that it is by an expert, or is true, would mask this epistemic problem faced by the hearer.[7] (Her problem would re-emerge as the question: was the assertion I observed a piece of *testimony*?)

3. UNDERSTANDING AND KNOWLEDGE OF WHAT HAS BEEN ASSERTED

A socio-linguistically[8] competent hearer, when she is told that P in a language she understands, will understand both the content and force of that utterance, and on that basis will know that she has been told that P. I take this to be a species of perceptual knowledge.[9] Explaining how such knowledge is achieved is a different task to that addressed in this essay. From the standpoint of the epistemology of testimony, knowledge that one has been told that P is the perceptual given, and we must explain how a hearer can get from there to knowledge or justified belief that P: characterise the legitimate epistemic route from the first belief to the second. However, knowledge of what one has been told is an input to this second epistemic problem, and there are some connections.

First, it is a presupposition of our account of testimony that knowledge that such-and-such has been asserted is generally available to a socio-linguistically competent hearer. A radical skepticism about the determinacy of meaning, and consequently about the possibility of knowledge of the content and force of particular speech acts, would preclude knowledge being acquired through testimony.

Second, the approach taken below to our task assumes that knowing that one has been told that P is one psychological and epistemic state, and believing that P on that basis is a further, independent one. If, per contra, an account of understanding were offered which connected it *internally* with forming belief in what one perceived to be asserted, this would place constraints on our epistemology of testimony. It is therefore an important question for our main task, whether the capacity to understand assertoric utterances made in a language is in itself distinct from any disposition to form belief in what one is told. If there were a necessary link here, this could be argued to support a non-inferentialist account of testimony (see sect. 6).

Third, understanding a speech act necessarily involves perceiving that act, and correctly apprehending both its content and its force. But does such understanding always involve, or cause, forming belief that the speech act has been made? There must surely be a disposition to form knowledgeable belief, if one were to attend to the question what speech act one has observed. But perhaps when one is listening, in trusting mode, to a fluent discourse, there may not be any formation of actual belief, as to what speech acts one is hearing: one may just take in what one is being told, via one's understanding of what is said, but without any beliefs about what is being said being formed. If this is the actual psychology of testimony, and we hold the Attainability Constraint, then our normative epistemology had better fit with this fact. [10]

4. CONTEXT: THE GENERAL ACCOUNT OF KNOWLEDGE, AND JUSTIFIEDNESS[11] OF BELIEF

We are seeking to give a general explanatory account of how testimony-beliefs may be justified, and knowledgeable: both belief in fresh instances of testimony, and the question of global epistemic dependence on testimony in our belief system. I suggested that such an account will be convincing only if it is the result of applying to the case of testimony a general conception of what it takes for a belief to be justified, or to be knowledge – a set of conditions established a priori as necessary and sufficient for this.[12] (Of course the general conception, and our account of how both testimony and other epistemic links yield knowledge and justified belief, may be developed simultaneously.)

We can see how the general account constrains, even if it does not fully determine, what we say about testimony in particular, with an example. Suppose one holds a Pure Reliabilist general conception of knowledge: knowledge is belief formed through a belief-forming method which is sufficiently reliable. Then, clearly, a hearer's belief gained through testimony will be knowledge just if the method she used was sufficiently reliable. The conception only constrains, rather than fully determining the account, because how reliable is sufficient, and how we individuate methods, is left to be determined; and the latter in particular is crucial, and open. For instance, if a hearer's method is individuated speaker-specifically, as a set of methods: 'believing speaker N', 'believing speaker M', etc, this will give the verdict that she gains knowledge whenever she learns from a speaker who is in fact trustworthy – even if she would equally believe an untrustworthy one. But truly describe this same, gullible hearer as forming beliefs through the method of 'believing anyone who tells her something', and her method is then revealed as unreliable. (Thus by individuating methods to suit our intuitions, we can get almost any result we like about particular cases, from a Pure Reliabilist general conception of knowledge.)

In the rest of this essay I shall focus mainly on the issues about justifiedness, local and global: when and with what checks a hearer is justified in believing what she is told; and whether and how a belief-system with global dependence on past trusted testimony can be justified. The notion of justifiedness of belief I pursue is identified as one which is potentially prescriptive: providing precepts which a hearer should seek to follow, and should normally be able to follow, in her formation of belief.[13] This is clearly an important topic, whether or not it is the only proper notion of justifiedness we may seek to characterise, and whether or not such justifiedness is a necessary condition for knowledge. In sect. 7 I argue that it *is* a necessary condition for knowledge. Considering testimony highlights the unsatisfactoriness of a purely reliabilist conception of knowledge. In testimony, as elsewhere, a creature capable of reflection needs a logos for her own belief if it is to be stable –

specifically, she needs to be able to tell a story about the genesis of her belief which reveals it as likely to be true. If she cannot, then reflection will undermine it.

Some philosophers do not adopt the unified top-down method I have proposed, but set out to give an account of when testimony-beliefs are knowledgeable, or justified, by bottom-up extraction of epistemic principles specific to testimony from intuitions about particular cases. These principles are not constrained to fit with any general conception of knowledge, or justified belief. But this approach is open to the danger of failing to see knowledge as one thing, albeit acquired from various sources This is so if, say, accounts are given of what it is for a testimony-belief, a perceptual belief, and a belief arrived at through induction, to be knowledge, which fail to instance a single general conception. The criticised philosopher may respond that our everyday concept of knowledge is not unified, but is a disjunctive or overlapping family resemblance concept. If so, we should ignore ordinary language, and develop a theory of a central, epistemically desirable property of beliefs. This is worthy to be called knowledge.

Space prevents further discussion of these issues in general epistemology, despite their close bearing on how we should approach testimony. For the rest of this essay I shall examine first our local issues, descriptive and normative – the psychology and epistemology of fresh acquisitions of belief through testimony (sects.5-7); and then turn to global considerations regarding the place of testimony in our system of empirical belief: the extent of dependence on testimony in the belief system of a member of a modern society, and if and how a belief-system which exhibits such dependence on testimony can be justified (sects.8.9).

5. THE PSYCHOLOGY OF BELIEF-ACQUISITION THROUGH TESTIMONY

Each of us (normal adult humans) has mastered the same basic world-view. It has three overlapping components: a commonsense theory of the nature of the material world, a theory of our own natures as embodied agents and thinkers, and a theory of our place in this world: how we are acted upon by it, and act upon it. This last includes a conception of the various ways we have of gaining knowledge about the world – the *epistemic links* of seeing, hearing, and other perceptual modalities. Our theory about ourselves is equally a theory about other human persons, a folk psychology. Overlapping with this is a folk conception of the nature of language, and their and our use of it: a commonsense theory of language as both semantic system and social institution. This includes a conception of the nature of various speech acts, including assertions. As part of this folk linguistics, we have the conception of a further epistemic link, testimony – one which, in favourable instances, gives us information about the material world via a route that goes through the psychology of another person, and the intentional linguistic acts made by them as a result of their mental state. This commonsense conception of the link of testimony shows us that, flukes apart, what one is told (what someone asserts) is true just if the teller is sincere (believes what she asserts), and her belief is true.

Adult hearers encounter fresh instances of testimony with all this in their cognitive background; this, plus quite a lot of further empirical knowledge of human nature, and the weaknesses and fallacies it is prone to – in particular, lying, and honest error. Even if one knows someone to be a sincere person, and they show no sign of deceitful intent, nor lack of confidence in what they tell one – even the best of us can be wrong sometimes, and many people form belief much too incautiously.

Given these facts, would one not expect a judicious person to approach others' testimony cautiously, if not skeptically? When told something, will she not estimate the chances, on the evidence available, of the teller being sincere and competent; and believe her only if she makes a positive estimate of these?

This is certainly true when we hear testimony in a relatively formal setting. We consciously attend to the fact of someone's testifying that P; and we then consider, as a separate matter, whether or not we should believe her: whether or not she is *trustworthy* – that is, both sincere, and likely to be right about the subject matter, P, of her assertion (call this her *competence wrt P*).[14] The defendant in a trial will surely be attended to in this manner by the jury. A person also shows readiness to evaluate informants critically when, looking for someone to ask the way of in a strange town, she does not ask the first person she sees, but looks out for someone whose aspect suggests he will know – someone who looks like a resident, and intelligent and alert.

But while in settings like these we consciously attempt judicious evaluation of the speaker, this is not universal. There are many other settings in which it is certainly not typical for a person to devote conscious attention to evaluating her informant's trustworthiness. My friend comes in from outside, exclaiming that the traffic is terrible today. I do not consider the question whether she is trustworthy regarding this utterance: I immediately form belief in what she has told me, accepting what she says automatically, unquestioningly, or so it seems.

However the fact that conscious attention is not devoted to evaluating the speaker in such settings does not mean that in them we just accept what we are told blindly, without any evaluation of her. It is an empirical question how the psychology of belief-acquisition through testimony typically goes, and my remarks here are no more than casually informed speculation. But we can see that non-conscious, or non-attentional mechanisms can be at work, governing a person's formation of belief in response to what she is told, though she is not consciously thinking about whether she should trust the speaker. Say that someone is a *blind truster* if she is characterised by this unqualified universal conditional: For all testifiers T, and for all propositions P, if H is told that P by T, then she will form the belief that P.[15] The fact that a hearer does not engage in conscious evaluation of a speaker does not entail that she is a blind truster. Unconscious, automatic monitoring of a speaker can operate in a hearer, and all kinds of sensitivities to aspects of the speaker, and to what precisely she asserts, can be present in her, to block the formation of belief, when they are cued.

There are two broad kinds of possible sub-attentional mechanism modifying a hearer's response away from blind trust. First, existing background beliefs – about the teller herself, or about all those in her circumstances, or with her subject matter – may operate like switches, modifying her response to being told, either towards acceptance, or to disbelief, or to suspended belief. Second, a hearer may, sub-attentionally, monitor the teller perceptually for current cues as her to trustworthiness. (Of course data from perception will interact with background beliefs, in this monitoring process.) This monitoring may be a full assessment, such that belief is formed if and only if it yields a positive verdict; or it may be that the teller is monitored only for defeating cues, signs of lack of sincerity or competence: we believe unless some such defeating cue is picked up – a hesitancy in the voice, an insincere-seeming smile, or just extreme prior improbability of what she tells us.[16]

A blind truster fails to have such monitoring mechanisms at work, in her doxastic response to others' utterances. Of course someone may be excessively gullible, while falling short of the extreme of blind trust. At the other extreme, someone may be unwilling to trust others to an irrational extent. No doubt all degrees along the scale are instantiated in the diversity of individual human psychology. This is an empirical issue, and I shall not speculate on its detail. Thomas Reid (1813) thought there was a natural human disposition to trust others, though he admitted that it was modified by experience of human folly and iniquity. Our present conclusions are, summarising: that we sometimes, but not always, attend to the question of whether a certain speaker is to be trusted on an occasion; but that even where attention is not devoted to this, there may be, and sometimes certainly is, sub-attentional sensitivity to defeating background beliefs, and monitoring of the speaker for signs of untrustworthiness. In the next section I turn to normative matters: conditions of justifiedness for beliefs acquired through testimony. The important conclusion from this section is that, though our formation of belief through receipt of testimony is often not via conscious deliberation about the trustworthiness of the speaker, and inference from this to belief in what she asserted, nonetheless it may be mediated by background beliefs and perceptual cues, in a fashion which *preserves the causal dependence of belief-formation on our possession of appropriate grounds*, which there is when belief is formed via conscious deliberation about trustworthiness of the speaker.

6. CONDITIONS FOR JUSTIFIED BELIEF IN WHAT ONE IS TOLD

At the start of the previous section we saw the basic nature of the epistemic link of testimony. Both seeing and testimony have conditions of veridical operation, conditions such that, flukes apart,[17] the deliverances of the link (for perception, the content of a perceptual experience; for testimony, the asserted content of a speech act) are veridical, match how things are, if and only if they obtain. Call these the *Veridicality-conditions* (V-conditions) of the link in question. The V-conditions of seeing are that the conditions of viewing are normal in various critical respects, and similarly for the state of the perceiver herself.[18] For testimony, the V-conditions are that the speaker is trustworthy – both sincere, and competent with respect to P. When these hold of a speaker, her testimony is necessarily veridical. That is: She asserted that P, and she is sincere, and competent with respect to P, logically necessitates P, in virtue of the content of the concepts of assertion, sincerity, and competence with respect to P.[19]

A hearer who has normal knowledge of commonsense linguistics thereby appreciates these V-conditions for testimony. If she were to form a belief in what she has been told, grounded in belief in the fact that she has been told, plus belief that the teller is trustworthy, this first belief would clearly be justified, so long as she was justified in believing these grounds for it. Justly apprehended entailment is an adequate grounding relation, if anything is.[20] But we observed in the previous section that hearers do not always devote conscious attention to the question whether a teller is trustworthy or not – although, we saw, this is consistent with their monitoring for this sub-attentionally. Our present topic is conditions of justifiedness for testimony beliefs, and central to this is the issue of what constitutes adequate grounds for testimony-beliefs.[21] In the case of visual perception, seeing, it is implausible that one is only justified in believing in what one seems to see (accepting as true the content of one's visual experience), if one has evidence that the V-conditions of perception obtain. A more plausible account of when a perceptual belief has adequate grounds is that one is justified in forming belief in what one seems to see – the content of one's experience – so long as this is not defeated by evidence in one's possession, or immediately available to one, that the V-conditions of perception are not fulfilled. That is to say, there is an a priori warrant to 'believe one's eyes'. Clearly this warrant is defeasible: justified belief, or mere grounds for suspicion, that the V-conditions of perception are not fulfilled defeats it.[22]

Given that our concern is with the question in what circumstances a hearer is justified in forming belief in what she is told, our central question about this 'local' issue in the epistemology of testimony is: Is or is not a hearer entitled to assume that the V-conditions of testimony are fulfilled, without needing evidence of this? Is there an a priori warrant to believe what one is told, simply on the ground that one has been told it? Again, this warrant, if it exists, will clearly be defeasible: as with seeing, justified belief that the V-conditions of the link do not obtain, or merely some ground for doubt that they do, must defeat it. (In the case of testimony, this means anything which calls in doubt the sincerity or competence of the speaker. Notice that evidence against the truth of what she asserts is one thing which does this. See note 16.) We may call the thesis that there is such a defeasible a priori warrant – a presumptive epistemic right – to believe what one is told as such the *Presumptive Right (PR) Thesis*. The PR Thesis amounts to the thesis that *a hearer is entitled to presume a speaker to be sincere and competent regarding her subject matter, unless she has grounds to doubt this*. To deny it is to insist that a hearer should not believe what she is told, unless she has empirical grounds for believing the speaker to be trustworthy. Space limits in this essay do not permit extended discussion of arguments for and against the PR Thesis. I shall mention some of the main arguments which might be advanced for and against. [23]

A mature hearer appreciates the nature of the link of testimony. In addition, she will have a grasp of folk psychology, and as part of this be aware of the various motives for deceit to which humans are prone, and equally the many ways in which they can fall into honest error. It seems that a rational hearer with this background of knowledge will not believe a speaker without assessing her trustworthiness. If she forms belief without such assessment, she is in effect assuming without evidence that the speaker is trustworthy; but this flies in the fact of her knowledge of folk psychology.[24] These facts make a prima facie case against the PR Thesis: given how easily testimony can fail to be true, why should the PR thesis hold? However the case is not decisive: the consideration just raised is near to decisive against the view that, for an adult hearer, the presumptive right to believe what she is told as such *still stands*. But it is consistent with this to hold that for a hearer wholly ignorant of folk psychology, the presumptive right obtains – thus that small children enjoy that right; but that by the time a person has reached maturity, the empirical knowledge of human nature she has gained defeats it, once and for all.[25] Nonetheless, given the facts about the nature of the link, and about human psychology, the burden of proof is with the PR theorist, to explain why, despite these facts, the PR thesis holds: as yet we have no argument for it.

One way of arguing for the PR Thesis, is from the Attainability Constraint. The form is a transcendental argument: there is knowledge and justified belief from testimony; this is possible if and only if the PR Thesis holds; therefore the PR Thesis holds. This form could be filled out by arguing in detail that it is impossible for a hearer to get adequate empirical confirmation of a speaker's trustworthiness, without assuming this very fact. So this cannot be needed, for justified belief through testimony. This attempted transcendental argument fails. It may indeed be impossible to get independent confirmation of the unrestricted universal generalisation: 'testimony is generally reliable' (indeed it must be, since this is false!). But there are many occasions on which the trustworthiness of a *particular* speaker, regarding a particular utterance of hers, can be empirically established without reliance on any testimony from her. One way is if she has a good track record about the topic of her assertion – her past pronouncements about this topic have all been subsequently confirmed through the hearer's own perception.[26]

A second argument invoking the Attainability Constraint, together with a thesis about the nature of understanding, might be tried. Suppose it were shown that it is internal to the nature of the psychological state of understanding an utterance and perceiving it as an assertion that P, that this state tends, albeit defeasibly, to produce belief that P. If so, this fact could be used as the basis of an argument for the PR thesis. ("One *can't help* tending to just believe what one is told; therefore this is epistemically permissible.") I think that an argument of this kind can be made to defend the analogous PR thesis for perception. The very nature of a perceptual experience with objective content is such that, in the absence of defeat, it produces belief in what one seems to see. But I think the parallel argument for testimony, from the nature of understanding as a psychological state, fails. This is one point where our preferred account of understanding interlocks crucially with our epistemology of testimony.[27]

I suggested in sect.4 that our epistemology of testimony should be developed together with a general conception of knowledge and justified belief which it instantiates. An alternative method, we saw, is not to aspire to such systematicity, but simply to extract certain principles about testimony from our everyday practices and intuitions. A case for the PR Thesis might be mounted, employing this method. But this method is theoretically unsatisfying. We should seek to give an illuminating general account showing *why* forming belief in accordance with the PR thesis is doing so justifiedly. Thus deciding this issue requires developing a general conception of when conditions that must hold for an inference to be truth-preserving can be assumed without evidence, or consideration, to hold by a believer, and when they should be, as it were, in the forefront of her space of reasons, included in her grounds for belief, and themselves adequately grounded.

7. KNOWLEDGE OF WHAT ONE IS TOLD

We have been considering under what conditions a hearer is justified in believing what she is told. Given our internalist conception of justifiedness, this is equivalent to the issue what are adequate grounds for a testimony-belief. But what about knowledge gained through testimony? To what extent are issues about knowledge, and about adequate grounds for belief, distinct?

Clearly, knowledge and justifiedness can come apart in one direction: justified belief in what one is told can fail to be knowledge, because the belief is false, or because though true itself, it rests on a false, though justified belief.[28] But should we take the subject's possession of an adequate ground, a *logos*, for her belief, to be a necessary condition for it to be knowledge – in testimonial beliefs, and elsewhere? I think considering the case of testimony can reinforce a general point about how to conceive of knowledge, that centrally desirable epistemic condition.

What is the best general conception of knowledge?[29] Perhaps a Pure Reliabilist notion of knowledge as belief formed by a reliable method has a useful application to unreflective creatures – higher animals, and *very* small children (very small – children start to be capable of reflection about the pedigree, the source, of their beliefs, very soon). But we are considering when an adult human gains knowledge through testimony. We are reflective creatures, with the capacity to ask ourselves questions about the status of our own beliefs. In particular, we can frame the questions: How do I know this? What reason do I have for thinking this is true? – What evidence do I have for this? Now, for a creature capable of such reflection, a belief cannot be stable under such reflection unless she can provide at least a basic answer to these questions. But the answers will be, or will advert to, a logos for her belief, at least a basic form of grounding support for it that she can supply. Thus we have arrived at a minimum 'internalist' condition: that a person has available to her the means to provide at least a basic answer to the question: How do you know that? If she cannot do this, then her own doxastic state is absurd to her, she cannot make sense of it. (Any belief is a belief that the world is a certain way: How can I coherently hold it, if I have no idea how my belief connects with the world being that way? Knowledge of lack of any such connection rationally undermines the belief.)[30] In creatures like our adult selves, capable of this sort of reflection, this minimal internalist condition is surely a requirement on knowledge. If we do not require it, we allow that a belief of mine can be knowledge, although I have no idea how I came to hold it, and whether that connects with its truth; in which case it cannot be stable under reflection. This seems very wrong. In practice, in commonsense epistemology, we have concepts of a range of epistemic links: the various modes of perception, inference, memory and testimony; and we cite their operation to answer the question: "How do you know that?" By doing so, we provide, both for others and *ourselves*, an explanation of how we have come to know that thing, which one of the familiar types of epistemic access to that fact we have enjoyed. – I saw it; I remember doing it; I worked it out; Someone told me. [31]

Let us turn to testimony armed with these general thoughts about knowledge. There is a prima facie rather attractive idea that in testimony knowledge just 'rubs off' on one person from another.[32] The conjecture is that it is a correct epistemic principle about testimony that:

> **Pure Transmission principle (PTP)**: If A knows that P, and tells B that P, and B understands what she is told and thereby comes to believe it, then B knows that P.

–If the teller knows, and the hearer understands her, this is sufficient for the teller's knowledge to 'rub off' onto the hearer. Talk of knowledge being 'transmitted' through testimony is nonsense, if it posits a kind of stuff that is transferred from one person to another, like electrical current or a virus. (The danger of falling into such nonsensical thinking is a good reason to avoid the metaphor of transmission.) It is intelligible, if it is no more than an expression of PTP. But PTP, though prima facie attractive, is quickly revealed to be implausible. First consider that belief which satisfies PTP is not even ensured to be reliable, unless we cheat on our individuation of methods to get that result: 'Believing a teller who knows what she asserts' is of course a reliable method. But it can be that 'Believing whatever A tells one' is not, and yet sometimes A's tellings are expressions of knowledge: someone who is a habitual liar, or who frequently jumps to false conclusions on inadequate grounds, will sometimes know, and tell what she knows to others. There is, so far as I can see, no general conception of conditions for knowledge found in the literature, which is consistent with PTP. If we were content with proposing particular epistemic principles ad hoc, we would not mind this. But I have suggested that we should not be so content.

Reliabilism suggests modifying PTP to:

> **Reliable Transmission Principle**: If A knows that P, and is generally disposed to be knowledgeable on that sort of topic, and A tells B that P, and B understands what she is told and thereby comes to believe it, then B knows that P.

RTP is less implausible than PTP. If someone is generally knowledgeable about some topic, then just believing what they say on it will lead to mainly true beliefs. But RTP is not a proper formulation of a reliabilist account of testimonial knowledge either: we need to formulate the method which the hearer B is using. For all that is said in RTP, she may be a blind truster. But then her method is not reliable, even when she in fact learns from a reliable source. [33]

In any case, we have raised doubts about the sufficiency of reliability as a condition for knowledge. Suppose a hearer has learned that P from a reliable informant. How does she make sense of her own belief, provide an account of how she knows that thing? If she is normally conceptually equipped, she has ready to hand the answer: I was told it. (To recap: this is one of our standard everyday explanations of how we have had epistemic access to some fact, testimony being a known epistemic link.) Reliabilism does not build in any requirement that a subject can make sense of the pedigree of her belief in this way.[34] But we have suggested that, in reflective creatures, this is a requirement for knowledge, since it is a requirement for stability of the belief under reflection. So it seems that, to gain knowledge through testimony, a hearer must be able to form the knowledgeable belief: I have been told that P by T, and to cite this to explain how she knows that P.

Can we stop here? If someone has the concept of telling, then she appreciates its V-conditions. But then she knows how easily a telling can fail to be true, if the speaker is lying or mistaken. But then does not making sense of her belief equally include having a reason to believe the speaker to be trustworthy? We are back with the issue of the PR Thesis. We have seen that the Pure Transmission Principle, while it has a certain initial appeal, does not stand up to scrutiny. To know something which one has been told, one must be able to make sense of the pedigree of one's belief, see why it is likely to be true. This means being capable of defending one's claim to know by citing the fact that one has been told. But once this is acknowledged, it is hard to resist the further conclusion that one must have grounds to take one's informant to be trustworthy. In any case, we have seen that having an internal logos, an adequate ground, for one's belief, is a requirement on testimonial knowledge: one cannot gain knowledge through testimony unless one has a ground in virtue of which one is justified in accepting as true what has been told to one. What exactly is required for such a ground, for a testimony-belief, is the issue of the PR Thesis discussed in the previous section. Since justifiedness is a necessary condition for knowledge, whatever is necessary for the first is so for the second.

8. OUR EPISTEMIC DEPENDENCE ON PAST TRUSTED TESTIMONY

In this section I shall examine the extent of dependence on testimony within an adult person's system of empirical belief. We will see that it is extensive. We observed that, in a modern society, there are extensive resources for acquiring belief from testimony, in an extended sense which includes books, the media, the internet, and so forth. A great deal of what we believe, we have come to believe at second hand, through testimony – sometimes through a long chain of testimonial links. Many other beliefs of ours have been formed through inference which included testimony-beliefs amongst its premisses. However, to say that one has originally acquired a belief wholly or in part from testimony is one thing; that it still depends, for its grounding, on that source in testimony, is another matter. Moreover, even if it does still depend on testimony, the individual's acceptance of the original testimony may or may not have been mediated via empirically justified belief in the teller's trustworthiness. I shall look first in more detail at the extent of causal dependence on testimony as a source of beliefs, in an individual's developmental cognitive history, and then return to these questions about current dependence for grounding.

We start acquiring beliefs at second-hand, through testimony, as soon as we are able to understand language at all. Many particular factual beliefs are acquired through testimony, from our parents, and then a broader range of teachers. More fundamentally, testimony is heavily involved in laying down the conceptual framework into which particular beliefs, including those which we acquire at first-hand from our own perception, are fitted. This is true of the geographical, political and historical framework into which particular facts of these kinds are slotted: general beliefs about geography, politics and history organise the data we individually perceive. When I judge that Bologna is humid, I perceive the humidity myself, but my concept of Bologna as a north Italian city, and my knowledge that I am currently in it, rests on testimony. Such learning of general organising beliefs is not sharply distinguished from the gradual process by which we come to be master of the concepts expressed in our first language. There is, for instance, no sharp distinction between learning facts about chairs – that one can sit on them, that they have backs, that some are made of wood – and coming to understand the word 'chair'.

In a child's very early days, her attitude to her teachers is necessarily one of *simple trust*: he or she is disposed to react acceptingly to what she is taught (form belief, once she is intellectually developed enough for that notion to be applicable), and lacks the conceptual resources to raise the question whether she should trust what she is told, or schooled in. The early phases of language-learning involve such accepting reactions to ostensive teaching of word meanings. Quibbling over whether this is precisely 'testimony' does not do away with the basic fact that an infant's learning of language involves unquestioning practical acceptance as true, of what her teachers say to her.

If there can be no thought without language, and if – as is almost certainly a law of human psychology and neurobiology, even if not logically necessary – a human being cannot acquire language except through being taught it, then it is at least a psychological necessity that all humans have extensive historical dependence on testimony in their development of a system of empirical belief. But, as remarked, historical dependence on testimony in the process of acquisition of concepts, language and beliefs is not the same thing as current dependence for grounding. Perhaps, having ascended via the ladder, one can then take hold elsewhere, and kick it away. This possibility deserves exploration. Certainly, there are some beliefs which a person first acquires through believing what she is told, but then later acquires independent confirmation of: you tell me that it is raining; shortly I go out and see for myself that you spoke truly.[35] But could this later independent confirmation work for our testimony-beliefs generally, so that the initial dependence for grounding on testimony is wholly removed? Once this possibility is raised, we see that the idea of wholly independent empirical confirmation of all of our testimony beliefs, as it were simultaneously, is absurd. It is absurd since, as remarked, testimony beliefs play a major role in laying down the framework of commonsense worldly knowledge which organises our formation and confirmation of individual beliefs. That there are any beliefs of ours at all which are wholly independent of testimony is debatable, given this role of testimony in concept-formation and laying down of organising background beliefs.

Given the extensive causal dependence on testimony for its structure and contents which our belief-system exhibits, is it just an unbacked article of faith we have no choice but to live by, that this testimony has inducted us into a world-view which is broadly correct? – That most of what we accepted as true, in the process by which 'light dawned' for us on a specific world-view, and we became believers and agents, really was true? No, the situation is not so bleak epistemically. Although wholly independent confirmation is not possible, we can have a non-reductive, internal confirmation of the broad correctness of the world-view we have bootstrapped our way into in part through trusting testimony, from the fact of its extensive internal coherence. It is a contingent fact that an individual has an internally coherent system of empirical beliefs; not one guaranteed by any general law. It is contingent that the system of empirical beliefs we arrive at through trusting testimony, along with the deliverances of perception, memory and inference, is broadly coherent. Specifically, it is contingent that what we are told coheres with what we see and remember for ourselves. What an individual is currently told by her peers may conflict with what she perceives for herself; and a person may reject what she has been taught as a child, when her own perception and independent thought subsequently reveal its falsity.

Now there is a good abductive argument to be made, from the contingent fact of coherence in our system of beliefs, to the likely truth of the epistemic sources which have led to its formation. Broadly speaking: if what the natives tell you fits with what you seem to see for yourself, and to remember, chances are that all these three are true. Explanations of fortuitous coincidence in a series of misleading sources would be more elaborate, and less plausible. Thus, while there is no chance of confirming all we came to believe through testimony while abrogating all reliance on it, still we can have a powerful abductive argument from coherence in our beliefs, to likely truth of our sources of belief, including testimony.

Is testimony then of equal status with our other empirical sources of belief? – perception, and memory – which does not originate belief, but preserves it through time, hence is a source of beliefs at a time. In some ways they are on a par: each is inextricably involved as a causal source of beliefs, and of grounding, in our belief system, and each can, in some cases, trump the evidence of the other. In particular, there are circumstances in which another's testimony concerning some matter may be better evidence for me than my own memory, or my own apparent perception, regarding it. But testimony always depends on perception, because the receipt of testimony depends on perception of the written or spoken act in which it is made. Neither perception nor memory depend similarly on testimony, or on each other. Thus testimony is not a causally autonomous source of belief, as perception is, in simple cases. This dependence means that testimony can be no more reliable than perception is. When I trust what another reports to me that she has seen, there is a double reliance on perception – both the teller's original perception of what she reports, and my perception of her speech act; as well as a reliance on her memory. Consequently there is no possible world in which testimony is reliable, although perception is massively unreliable: the latter will infect the former, since it is involved in the process of telling itself.

9. GLOBAL NORMATIVE ISSUES

We have seen that any human language-user's belief system exhibits a diffused general or 'global' dependence on testimony, in its empirical grounding. The most we can have by way of reassurance of its likely truth-in-the-main is the abductive argument just examined. Can a belief system with such global dependence on past testimony be justified?

We must introduce a refinement to our discussion of this issue. When a hearer believes what she is told in a fresh instance of testimony, but her belief is mediated by empirically well-founded belief in the speaker's trustworthiness, there is no ungrounded trust of testimony involved. If, for all the beliefs one had acquired through testimony, one had had such an empirical basis for trusting the speaker, then there would be no need to appeal to any epistemological principle concerning testimony in particular, to explain the justifiedness of our empirical beliefs. In contrast, as we saw in sect.6, when a hearer believes what she is told without possessing evidence of the speaker's trustworthiness, this is justified if and only if the PR Thesis holds. So the question we need to ask is: Can we explain the status as justified of our system of empirical beliefs, without appealing to the PR Thesis concerning testimony? Epistemological Reductionists about testimony think that we must do this: that our system of empirical beliefs is justified only if any past dependence on ungroundedly-trusted testimonial beliefs can be eliminated. That is, they hold that our system of belief is justified only if its status as such can be exhibited, without recourse to the PR Thesis: our entitlement to believe all that we've learned through past tellings can be accounted for by, reduced to, epistemological principles not including the PR Thesis.[36] Pessimistic reductionists think that this reduction is necessary, for knowledge and justified belief, but that it is not possible. A pessimistic reductionist about testimony is thus a skeptic: she must hold that our system of empirical belief, shot through as it is with uneliminated dependence on ungroundedly-trusted testimony, is unjustified and does not constitute knowledge. Optimistic reductionists think that we need such a reduction, but that we can get it. Anti-reductionists, on the other hand, think that we do gain knowledge and justified belief through testimony, and that our epistemic entitlement to the beliefs we gain through testimony does not need to be vindicated by a reduction; it is explained by the PR Thesis, which is correct.

Is the optimistic reductionist too optimistic? We examined in the previous section the prospects for eliminating dependence on what we have learned through past testimony. Now we need only to note that it is only beliefs acquired through ungroundedly-trusted testimony that are problematic, needing to have a new, alternative grounding supplied for them. This qualification makes little difference: it is the development of the conceptual foundations of our belief system which is most fundamental, and this occurs in the early period of simple trust. Thus the prospects of eliminating dependence on ungrounded trust are as we saw: it cannot be done, since the very idea of setting aside all my beliefs which depend on past simply-trusted testimony, to reconstruct my entitlement to them out of the materials that are left, is absurd. The reductionist about knowledge and justified belief from testimony can be satisfied only if she is prepared to settle for the abductive argument to truth as the best explanation of coherence in our belief system. But this does not really eliminate dependence on the PR Thesis, though it provides some support for its presupposition of the general reliability of testimony.

It seems, then, that we can explain the existence of empirical knowledge and justified belief, despite its diffused global dependence on testimony, only if we are prepared to renounce reductionist aspirations and accept the PR Thesis. However there is a further distinction to be drawn. It seems that we cannot explain the status as justified and knowledgeable of a person's system of beliefs, without invoking a PR Thesis applicable at least in her developmental phase. But this does not entail that for a mature individual, now possessed of a commonsense view of the world and her place in it, this epistemic right ungroundedly to trust still holds. One way of explaining such a distinction between the developmental phase, and the mature phase, is to hold that an infant starts off endowed with the epistemic right to trust what she is told as such; but by the time she has matured, this has been permanently defeated, by the knowledge she has acquired of the deceitfulness and folly of human nature: once one is old enough to know better, one should not trust blindly what others tell one. An alternative epistemological stance is to suggest that questions about whether our current belief system as a whole is justified are senseless or inapt; the only questions we can sensibly address are local ones about epistemic dynamics: when we should revise particular beliefs, and form new ones, within this framework. And regarding this latter local question, the correct epistemic precept is that one should trust what one is told only when one has adequate evidence that the speaker is trustworthy.

However precisely one explains this contrast, it does seem that a prudent hearer, when it really matters, will not believe what she is told without evidence of the sincerity, and the competence on her topic, of a speaker. (And when it really matters, she will not trust anyone else, but will check for herself if she possibly can: is my laptop on board the aircraft, not left in the airport?) The fact that one has ineliminable dependence on past testimony in the historical process by which one came to be a believer does not affect this essentially forward-looking practical point. But if there is a moral to be drawn for general epistemology from our consideration of testimony, it is that when we seek to explain the status as justified of our whole system of beliefs, thinking about the ineliminable role of testimony in its formation and grounding provides pressure towards a coherentist, not a foundationalist, view of the justification of empirical belief. Our best indication of the truth-in-the-main of a system of beliefs acquired in part through simply-trusted testimony, is the fact – when it is a fact – that doing so has led us to a highly coherent and integrated view of the world, including the various epistemic links to it that we enjoy – perception, memory, and testimony.

Elizabeth Fricker
Magdalen College, Oxford

NOTES

[1] Of course this does not require positing 'meanings' as dubious entities referred to by words or sentences. What we need, as a minimum, is an account of the nature of facts such as: teller T's utterance of sentence S on occasion O constituted an assertion that P. I myself think this cannot be done without invoking constant semantic properties conventionally associated with expression-types of a language – facts such as: 'dog' in English refers to an item if and only if that item is a dog.

[2] Why adopt this constraint on our epistemological theorising?- Well, surely there are better and worse ways of forming beliefs. We may think of an account of when testimony-beliefs are justified as having prescriptive force (which is not to say that we must be formulating rules that a hearer can follow infallibly, without any need for epistemic luck). If a philosopher tells us that none of our beliefs, nor our belief-forming methods, are justified, then we are left with no guidance as to how to conduct our epistemic life better or worse. Surely we do better to appraise our sources of belief judiciously, even if we acknowledge a logically possible worst case in which all of them are deceptive, and our appraisal is faulty. So we should opt to construct a notion of justifiedness of belief which is something we can aim for, try to conduct ourselves in accordance with.

[3] There is a structural parallel between testimony and memory. Neither is a source of entirely new beliefs. Testimony spreads belief from one individual to another, while memory preserves belief from one time to another, within an individual.

[4] Consider this case: A Russian national makes an assertion to me, in Russian, telling me that she is Russian. I do not understand what she says; however, I realise that she is speaking Russian, and correctly conclude from this that she is Russian. This is not a case of coming to know something through testimony, on our definition.

[5] An assertion is a saying that P in which the speaker commits herself to the truth of P – represents herself as believing, maybe as knowing, that P. Tellings are assertions made to an audience, with the intention to inform them of what is asserted. Even if assertions must be audience-directed, which is debatable, not all are aimed at informing – cf. review of facts, and exam situations.

[6] The internet provides the extreme here: no information whatsoever about the quality of its source can be gleaned from information placed on the net, apart from whatever may be inferred from the nature of the message itself. The problem of finding techniques for evaluating the quality of sources on the net is currently receiving much attention.

[7] I am here adopting a broadly 'internalist' epistemological stance: that we should be seeking to build a theory which characterises the believer's epistemic predicament 'from the inside', from her point of view. See note 2. In insisting that what a hearer initially gets, as it were, through perception, is that she has been told that P, I do not deny that, in some cases, the sincerity of the speaker, maybe even her good credentials, may be a perceptible matter. But they are not always so.

[8] Full competence as a human language-user requires inter-personal interpretative know-how as well as narrowly semantic and lexical competence, since the interpretation of particular speech acts requires more than just a knowledge of lexical meanings – appropriate sensitivity to contextual cues about likely motive and topic are needed, to disambiguate and fix the interpretation of a particular utterance of a sentence type. (Actually, this fact somewhat blurs our distinction between understanding a speech act, and evaluating the trustworthiness of the speaker: disambiguation may involve conjectures about topic and hence motive.)

[9] Here I have been influenced by McDowell (1980).

[10] There are two aspects to a general account of the grounding relation: what sorts of grounds are held to be adequate for belief; and what precise relation the believer must stand in, to those grounds. In sect.6 I maintain that adequate grounds for belief in what one is told must include justified belief that one has been told it. To cohere with the present point, our account of the second aspect must be that merely potential justified belief in an adequate ground is sufficient for actual justifiedness of the ground-needing belief. See note 20.

[11] Justifiedness is a more or less stable property of a belief, or a system of beliefs. 'Justification' I take to be a process or activity. Our main concern is with the first of these.

[12] Regarding knowledge, Williamson (1995) suggests that a founding supposition of mainstream analytic epistemology is mistaken: that the concept of knowledge, though it has some a priori necessary conditions, is not analysable, and that states of knowing are metaphysically simple. If so, the methodology proposed here is wrong, and it is not clear how much of epistemology as standardly conceived is left. I have not been convinced by Williamson's arguments.

[13] To say the conditions which characterise justified belief embody principles that a hearer should seek to follow does not commit one to a 'luck-free zone' form of internalism: one need not hold that the principles provide a canon such that it is in a hearer's power to conform to it in any possible world, regardless of how bad her epistemic luck is. I think this luck-free-zone internalism is a profoundly attractive, but equally profoundly mistaken, false goal.

[14] The property which the hearer needs to know the speaker to possess is, roughly, that expressed by this conditional: If S were to assert that P, then P. Fricker (1994) refines this further.

[15] Blind trust is at one extreme, undifferentiated distrust at the other,

of a continuum of degrees of trustfulness. These logical extremes are probably never instantiated. In particular, surely no hearer will believe an assertion that P if she already knows for certain that not-P. Generalising, the proposition asserted by a speaker having a low prior probability for the hearer is one kind of defeating cue for her.

[16] Some defeaters defeat the hypothesis that the teller is trustworthy directly – in which case belief in what she asserts is withheld, rather than active disbelief being formed. But others are defeaters of the proposition P itself, defeating the hypothesis of trustworthiness only via this fact. In this case, what is asserted is disbelieved, not merely judgement upon it suspended.

[17] A piece of testimony is flukishly true if the speaker seeks to deceive, but owing to a false belief on her part, unknowingly speaks the truth.

[18] It is debatable whether the V-conditions for seeing can be specified non-circularly: they cannot, if 'normal conditions', e.g. normal lighting, can only be specified as: conditions under which what one seems to see matches how things are. For testimony they certainly can be independently specified.

[19] To get an entailment in standard logic, one must specify conditions unpacking the concepts of assertion, sincerity, and competence with respect to P. I leave this as an exercise to the reader.

[20] The idea of one belief being grounded in others which support it has a central place in any 'internalist' account of justifiedness of belief. (This relation of grounding between beliefs is also sometimes called the 'basing relation'.) For a set of beliefs to ground a target belief, these supporting beliefs must have contents appropriately related to that of the target belief: ones which furnish the premisses of a good argument for the target belief. Different accounts of the grounding relation are possible. On one version, the grounding beliefs must be actual, and be causally sustaining the target belief. On another version, these beliefs need only be potential, beliefs which the subject would form if challenged, and which would then be justified. Clearly, which version we adopt will affect the answer we get to our question whether beliefs formed through testimony are typically justified. The stronger the requirement, the less likely that it is satisfied by ordinary folk, in their responses to testimony.

[21] On our internalist conception of justifiedness, the general conditions of justifiedness for a belief are that the subject possesses (actually, or potentially) adequate grounds for it, and that her belief is appropriately based on those grounds.

[22] Belief that the V-conditions do not hold will defeat perceptual belief, in a perceiver who appreciates the basic nature of the perceptual link. Notice that, if possession of such a disposition to accept defeat is necessary for perceptual belief to be justified, it follows that one can gain justified belief through seeing only when one has the grasp of the nature of the epistemic link of seeing which is needed, to appreciate what its veridicality conditions are, and how their absence defeats the ground for belief provided by a perceptual experience with objective content.

[23] Fricker (1994) investigates and rebuts some arguments for the PR thesis.

[24] Given our internalist conception of justifiedness as requiring an adequate ground for belief, the hearer must at least have, as ground for her testimonial belief, her perceptually-based knowledge: 'I was told that P by H.' This is how she gives an epistemically rationalising explanation of her belief: an explanation of how she came to know that P. But anyone who can give this explanation must have the concept of telling which features in it, and ipso facto appreciate that what is asserted is true if and only the speaker is trustworthy.

[25] Reid's position (Reid 1813) is arguably this, as is that of Burge (1993). It is interesting for the issues concerning global reduction considered in sect.8.

[26] See Fricker 1994 for a detailed discussion of these issues.

[27] Fricker 1998 investigates this issue in detail.

[28] This fact is familiar since Gettier 1963.

[29] Ordinary language provides some paradigm instances, but not a definition. We must theorise, to arrive at a general conception.

[30] The conclusion we have reached may be put thus: a necessary condition for a belief to be knowledge, is that the subject has the means to construct an epistemically rationalising doxastic explanation for it – that is, an explanation of why she holds it, which reveals it as connected with what it is a belief in, in a way which makes it likely to be true. Jones (1999) develops the idea of such explanations, and argues that a belief is undermined if I come to believe that the explanation of why I believe it is not epistemically rationalising.

[31] The minimal internalist condition is classically formulated: a necessary condition for a belief of mine to be knowledge, is that I have some idea of why it is likely to be true (See BonJour 1985, Ch.1). This requirement is fulfilled when I can give an epistemically rationalising explanation of my belief, and giving such explanations is our normal everyday way of justifying our beliefs. But it may not be the only way: citing evidence does not immediately have that form. However, I think it is plausible that a necessary condition for a belief to be knowledge is that the subject is able to construct an epistemically rationalising explanation of it. Even those who deny that 'knows' can be analysed, can admit that justifiedness is a necessary condition for knowledge.

[32] See McDowell (1994).

[33] Though a blind truster gains knowledge, on a reliabilist conception, if it is a law about her situation that she only encounters reliable sources.

[34] It could nonetheless be that the only way to be a reliable gainer of knowledge from testimony, is to be a monitor for trustworthiness, satisfying our requirement for justifiedness.

[35] This is what enables the independent confirmation of the trustworthiness of particular speakers!

[36] Hume (1748) provides the classic statement of the reductionist position. See Fricker (1995) for further discussion of these issues.

REFERENCES

Audi, R.: 1998, *Epistemology*, Routledge, London and New York, ch.5.

BonJour, L.: 1985, *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge, Mass.

Burge, T.: 1993, 'Content Preservation', *The Philosophical Review* **102.4**, 457-488

Burge, T: 1997, 'Interlocution, Perception, and Memory', *Philosophical Studies* **86**, 21-47

Chakrabarti, A. and B. K. Matilal (eds.): 1994, *Knowing from Words*, Synthese library, Vol 230, Dordrecht.

Coady, C. A. J.: 1992, *Testimony: a Philosophical Study*, Oxford, Clarendon Press.

Fricker, E.: 1994, 'Against Gullibility', in Chakrabarti and Matilal (eds.), op.cit., pp. 125-161

Fricker, E.: 1995, 'Telling and Trusting: Reductionism and Anti-reductionism in the Epistemology of Testimony', *Mind* **104**, 393-411.

Fricker, E.: 1998, 'Testimony and Perception: Some Contrasts', invited paper, American Philosophical Association, Western Division, meeting in Los Angeles of March 1998.

Gettier, E. L.: 1963, 'Is Justified True Belief Knowledge?', *Analysis* **23**, 121-123

Hume, D.: 1748, *An Enquiry Concerning Human Understanding*, in P. H. Nidditch (ed.), *Hume's Enquiries*, Oxford, 1975, sec. X, 'Of Miracles'.

Jones, W.: 1999, 'The View from Here: a First-Person Constraint on Believing', Doctoral Thesis, University of Oxford.

Lipton, P.: 1998, 'The Epistemology of Testimony', *Studies in the History and Philosophy of Science* **29 (1)**, 1-31

McDowell, J.: 1980, 'Meaning, Communication and Knowledge', in Z. van Straaten (ed.), *Philosophical Subjects*, Oxford, Clarendon Press, pp. 117-139.

McDowell, J.: 1994, 'Knowledge by Hearsay', in Chakrabarti and Matilal (eds.), op.cit., pp. 195-224.

Reid, T.: 1813, *An Enquiry into the Human Mind*, T. Duggan (ed.), Chicago, 1970, ch.6, sect.24.

Williamson, T.: 1995, 'Is Knowing a State of Mind?', *Mind* **104**, 533-555.

RICHARD SAMUELS, STEPHEN STICH, AND LUC FAUCHER

REASON AND RATIONALITY

1. INTRODUCTION: THREE PROJECTS IN THE STUDY OF REASON

Over the past few decades, reasoning and rationality have been the focus of enormous interdisciplinary attention, attracting interest from philosophers, psychologists, economists, statisticians and anthropologists, among others. The widespread interest in the topic reflects the central status of reasoning in human affairs. But it also suggests that there are many different though related projects and tasks which need to be addressed if we are to attain a comprehensive understanding of reasoning.

Three projects that we think are particularly worthy of mention are what we call the *descriptive, normative* and *evaluative* projects. The *descriptive project* – which is typically pursued by psychologists, though anthropologists and computer scientists have also made important contributions – aims to characterize how people *actually* go about the business of reasoning and to discover the psychological mechanisms and processes that underlie the patterns of reasoning that are observed. By contrast, the *normative project* is concerned not so much with how people actually reason as with how they *should* reason. The goal is to discover rules or principles that specify what it is to reason *correctly* or *rationally* – to specify standards against which the quality of human reasoning can be measured. Finally, the *evaluative project* aims to determine the extent to which human reasoning *accords* with appropriate normative standards. Given some criterion, often only a tacit one, of what counts as good reasoning, those who pursue the evaluative project aim to determine the extent to which human reasoning meets the assumed standard.

In the course of this paper we touch on each of these projects and consider some of the relationships among them. Our point of departure, however, is an array of very unsettling experimental results which, many have believed, suggest a grim outcome to the evaluative project and support a deeply pessimistic view of human rationality. The results that have led to this evaluation started to emerge in the early 1970s when Amos Tversky, Daniel Kahneman and a number of other psychologists began reporting findings suggesting that under quite ordinary circumstances, people reason and make decisions in ways that systematically violate familiar canons of rationality on a broad array of problems. Those first surprising studies sparked the growth of an enormously influential research program – often called the *heuristics and biases* program – whose impact has been felt in a wide range of disciplines including psychology, economics, political theory and medicine. In section 2, we provide a brief overview of some of the more disquieting experimental findings in this area.

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 131-179.
© 2004 *Kluwer Academic Publishers.*

What precisely do these experimental results show? Though there is considerable debate over this question, one widely discussed interpretation that is often associated with the heuristics and biases tradition claims that they have "bleak implications" for the rationality of the man and woman in the street. What the studies indicate, according to this interpretation, is that ordinary people lack the underlying rational *competence* to handle a wide array of reasoning tasks, and thus that they must exploit a collection of simple *heuristics* which make them prone to seriously counter-normative patterns of reasoning or *biases*. In Section 3, we set out this pessimistic interpretation of the experimental results and explain the technical notion of competence that it invokes. We also briefly sketch the normative standard that advocates of the pessimistic interpretation typically employ when evaluating human reasoning. This normative stance, sometimes called the *Standard Picture*, maintains that the appropriate norms for reasoning are derived from formal theories such as logic, probability theory and decision theory (Stein 1996).

Though the pessimistic interpretation has received considerable support, it is not without its critics. Indeed much of the most exciting recent work on reasoning has been motivated, in part, by a desire to challenge the pessimistic account of human rationality. In the latter parts of this paper, our major objective will be the consider and evaluate some of the most recent and intriguing of these challenges. The first comes from the newly emerging field of *evolutionary psychology*. In section 4 we sketch the conception of the mind and its history advocated by evolutionary psychologists, and in section 5 we evaluate the plausibility of their claim that the evaluative project is likely to have a more positive outcome if these evolutionary psychological theories of cognition are correct. In section 6 we turn our attention to a rather different kind of challenge to the pessimistic interpretation – a cluster of objections that focus on the role of pragmatic, linguistic factors in experimental contexts. According to these objections, much of the data for putative reasoning errors is problematic because insufficient attention has been paid to the way in which people interpret the experimental tasks they are asked to perform. In section 7 we focus on a range of problems surrounding the *interpretation* and *application* of the principles of the Standard Picture of rationality. These objections maintain that the paired projects of deriving normative principles from formal systems, such as logic and probability theory, and determining when reasoners have violated these principles are far harder than advocates of the pessimistic interpretation are inclined to admit. Indeed, one might think that the difficulties that these tasks pose suggest that we ought to reject the Standard Picture as a normative benchmark against which to evaluate the quality of human reasoning. Finally, in section 8 we further scrutinize the normative assumptions made by advocates of the pessimistic interpretation and consider a number of arguments which appear to show that we ought to reject the Standard Picture in favor of some alternative conception of normative standards.

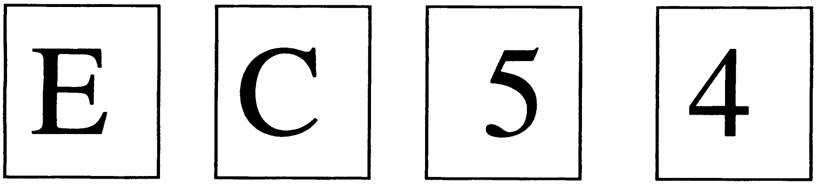
2. SOME DISQUIETING EVIDENCE ABOUT HOW HUMANS REASON

Our first order of business is to describe some of the experimental results that have been taken to support the claim that human beings frequently fail to satisfy appropriate normative standards of reasoning. The literature on these errors and biases has grown to epic proportions over the last few decades and we won't attempt to provide a comprehensive review.[1] Instead, we focus on what we think are some of the most intriguing and disturbing studies.

2.1. The Selection Task

In 1966, Peter Wason published a highly influential study of a cluster of reasoning problems that became known as the *selection task.* As a recent textbook observes, this task has become "the most intensively researched single problem in the history of the psychology of reasoning" (Evans, Newstead & Byrne 1993, 99). Figure 1 illustrates a typical example of a selection task problem.

Here are four cards. Each of them has a letter on one side and a number on the other side. Two of these cards are shown with the letter side up, and two with the number side up.



Indicate which of these cards you have to turn over in order to determine whether the following claim is true:

If a card has a vowel on one side, then it has an odd number on the other side.

Figure 1

What Wason and numerous other investigators have found is that subjects typically perform very poorly on questions like this. Most subjects respond correctly that the E card must be turned over, but many also judge that the 5 card must be turned over, despite the fact that the 5 card could not falsify the claim no matter what is on the other side. Also, a majority of subjects judge that the 4 card need *not* be turned over, though without turning it over there is no way of knowing whether it has a vowel on the other side. And, of course, if it does have a vowel on the other side then the claim is not true. It is not the case that subjects do poorly on all selection task problems, however. A wide range of variations on the basic pattern have been tried, and on some versions of the problem a much larger percentage of subjects answer correctly. These results form a bewildering pattern, since there is no obvious feature or cluster of features that separates versions on which subjects do well from those on which they do poorly. As we will see in Section 4, some evolutionary psychologists have argued that these results can be explained if we focus on the sorts of mental mechanisms that would have been crucial for reasoning about social exchange (or "reciprocal altruism") in the environment of our hominid forebears. The versions of the selection task we're good at, these theorists maintain, are just the ones that those mechanisms would have been designed to handle. But, as we will also see, this explanation is hardly uncontroversial

2. 2. The Conjunction Fallacy

Much of the experimental literature on theoretical reasoning has focused on tasks that concern *probabilistic* judgment. Among the best known experiments of this kind are those that involve so-called *conjunction problems*. In one quite famous experiment, Kahneman and Tversky (1982) presented subjects with the following task.

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Please rank the following statements by their probability, using 1 for the most probable and 8 for the least probable.
>
> (a) Linda is a teacher in elementary school.
> (b) Linda works in a bookstore and takes Yoga classes.
> (c) Linda is active in the feminist movement.
> (d) Linda is a psychiatric social worker.
> (e) Linda is a member of the League of Women Voters.
> (f) Linda is a bank teller.
> (g) Linda is an insurance sales person.
> (h) Linda is a bank teller and is active in the feminist movement.

In a group of naive subjects with no background in probability and statistics, 89% judged that statement (h) was more probable than statement (f) despite the obvious fact that one cannot be a *feminist* bank teller unless one is a *bank teller*. When the same question was presented to statistically sophisticated subjects – graduate students in the decision science program of the Stanford Business School – 85% gave the same answer! Results of this sort, in which subjects judge that a compound event or state of affairs is more probable than one of the components of the compound, have been found repeatedly since Kahneman and Tversky's pioneering studies, and they are remarkably robust. This pattern of reasoning has been labeled *the conjunction fallacy*.

2. 3. Base Rate Neglect

Another well-known cluster of studies concerns the way in which people use base-rate information in making probabilistic judgments. According to the familiar Bayesian account, the probability of a hypothesis on a given body of evidence depends, in part, on the prior probability of the hypothesis. However, in a series of elegant experiments, Kahneman and Tversky (1973) showed that subjects often seriously undervalue the importance of prior probabilities. One of these experiments presented half of the subjects with the following "cover story."

> A panel of psychologists have interviewed and administered personality tests to 30 engineers and 70 lawyers, all successful in their respective fields. On the basis of this information, thumbnail descriptions of the 30 engineers and 70 lawyers have been written. You will find on your forms five descriptions, chosen at random from the 100 available descriptions. For each description, please indicate your probability that the person described is an engineer, on a scale from 0 to 100.

The other half of the subjects were presented with the same text, except the "base-rates" were reversed. They were told that the personality tests had been administered to 70 engineers and 30 lawyers. Some of the descriptions that were provided were designed to be compatible with the subjects' stereotypes of engineers, though not with their stereotypes of lawyers. Others were designed to fit the lawyer stereotype, but not the engineer stereotype. And one was intended to be quite neutral, giving subjects no information at all that would be of use in making their decision. Here are two examples, the first intended to sound like an engineer, the second intended to sound neutral:

> Jack is a 45-year-old man. He is married and has four children. He is generally conservative, careful and ambitious. He shows no interest in political and social issues and spends most of his free time on his many hobbies which include home carpentry, sailing, and mathematical puzzles.

> Dick is a 30-year-old man. He is married with no children. A man of high ability and high motivation, he promises to be quite successful in his field. He is well liked by his colleagues.

As expected, subjects in both groups thought that the probability that Jack is an engineer is quite high. Moreover, in what seems to be a clear violation of Bayesian principles, the difference in cover stories between the two groups of subjects had almost no effect at all. The neglect of base-rate information was even more striking in the case of Dick. That description was constructed to be totally uninformative with regard to Dick's profession. Thus, the only useful information that subjects had was the base-rate information provided in the cover story. But that information was entirely ignored. The median probability estimate in both groups of subjects was 50%. Kahneman and Tversky's subjects were not, however, completely insensitive to base-rate information. Following the five descriptions on their form, subjects found the following "null" description:

> Suppose now that you are given no information whatsoever about an individual chosen at random from the sample. The probability that this man is one of the 30 engineers [or, for the other group of subjects: one of the 70 engineers] in the sample of 100 is ____%.

In this case subjects relied entirely on the base-rate; the median estimate was 30% for the first group of subjects and 70% for the second. In their discussion of these experiments, Nisbett and Ross offer this interpretation:

> The implication of this contrast between the "no information" and "totally nondiagnostic information" conditions seems clear. When *no* specific evidence about the target case is provided, prior probabilities are utilized appropriately; when *worthless* specific evidence is given, prior probabilities may be largely ignored, and people respond as if there were no basis for assuming differences in relative likelihoods. People's grasp of the relevance of base-rate information must be very weak if they could be distracted from using it by exposure to useless target case information. (Nisbett & Ross 1980, 145-6)

Before leaving the topic of base-rate neglect, we want to offer one further example illustrating the way in which the phenomenon might well have serious practical consequences. Here is a problem that Casscells et al. (1978) presented to a group of faculty, staff and fourth-year students and Harvard Medical School.

> If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms or signs? ____%

Under the most plausible interpretation of the problem, the correct Bayesian answer is 2%. But only eighteen percent of the Harvard audience gave an answer close to 2%. Forty-five percent of this distinguished group completely ignored the base-rate information and said that the answer was 95%.

2. 4. Overconfidence

One of the most extensively investigated and most worrisome cluster of phenomena explored by psychologists interested in reasoning and judgment involves the degree of confidence that people have in their responses to factual questions – questions like:

In each of the following pairs, which city has more inhabitants?

(a) Las Vegas	(b) Miami
(a) Sydney	(b) Melbourne
(a) Hyderabad	(b) Islamabad
(a) Bonn	(b) Heidelberg

In each of the following pairs, which historical event happened first?

(a) Signing of the Magna Carta	(b) Birth of Mohammed
(a) Death of Napoleon	(b) Louisiana Purchase
(a) Lincoln's assassination	(b) Birth of Queen Victoria

After each answer subjects are also asked:

How confident are you that your answer is correct?

50% 60% 70% 80% 90% 100%

In an experiment using relatively hard questions it is typical to find that for the cases in which subjects say they are 100% confident, only about 80% of their answers are correct; for cases in which they say that they are 90% confident, only about 70% of their answers are correct; and for cases in which they say that they are 80% confident, only about 60% of their answers are correct. This tendency toward overconfidence seems to be very robust. Warning subjects that people are often overconfident has no significant effect, nor does offering them money (or bottles of French champagne) as a reward for accuracy. Moreover, the phenomenon has been demonstrated in a wide variety of subject populations including undergraduates, graduate students, physicians and even CIA analysts. (For a survey of the literature see Lichtenstein, Fischoff & Phillips 1982.)

2. 5. Anchoring

In their classic paper, "Judgment under uncertainty," Tversky and Kahneman (1974) showed that quantitative reasoning processes – most notably the production of estimates – can be strongly influenced by the values that are taken as a starting point. They called this phenomenon *anchoring*. In one experiment, subjects were asked to estimate quickly the products of numerical expressions. One group of subjects was given five seconds to estimate the product of

8×7×6×5×4×3×2×1

while a second group was given the same amount of time to estimate the product of

1×2×3×4×5×6×7×8.

Under these time constraints, most of the subjects can only do some steps of the computation and then have to extrapolate or adjust. Tversky and Kahneman predicted that because the adjustments are usually insufficient, the procedure should lead to underestimation. They also predicted that because the result of the first step of the descending sequence is higher than the ascending one, subjects would produce higher estimates in the first case than in the second. Both predictions were confirmed. The median estimate for the descending sequence was 2250 while for the ascending one was only 512. Moreover, both groups systematically underestimated the value of the numerical expressions presented to them since the correct answer is 40,320.

It's hard to see how the above experiment can provide grounds for serious concern about human rationality since it results from of imposing serious constraints on the time that people are given to perform the task. Nevertheless, other examples of anchoring are genuinely bizarre and disquieting. In one experiment, for example, Tversky and Kahneman asked subjects to estimate the percentage of African countries in the United Nations. But before making these estimates, subjects were first shown an arbitrary number that was determined by spinning a 'wheel of fortune' in their presence. Some, for instance, were shown the number 65 while others the number 10. They were then asked to say if the correct estimate was higher or lower than the number indicated on the wheel and to produce a real estimate of the percentage of African members in the UN. The median estimates were 45% for subjects whose "anchoring" number was 65 and 25% for subjects whose number was 10. The rather disturbing implication of this experiment is that people's estimates can be affected quite substantially by a numerical "anchoring" value even when they must be fully aware that the anchoring number has been generated by a random process which they surely know to be entirely irrelevant to the task at hand![12]

3. THE PESSIMISTIC INTERPRETATION: SHORTCOMINGS IN REASONING COMPETENCE

The experimental results we've been recounting and the many related results reported in the extensive literature in this area are, we think, intrinsically unsettling. They are even more alarming if, as has occasionally been demonstrated, the same patterns of reasoning and judgment are to be found outside the laboratory. None of us want our illnesses to be diagnosed by physicians who ignore well-confirmed information about base-rates. Nor do we want public officials to be advised by CIA analysts who are systematically overconfident. The experimental results themselves do not entail any conclusions about the nature or the normative status of the cognitive mechanisms that underlie people's reasoning and judgment. But a number of writers have urged that these results lend considerable support to a pessimistic hypothesis about those mechanisms, a hypothesis which may be even more disturbing than the results themselves. On this pessimistic view, the examples of problematic reasoning, judgments and decisions that we've sketched are not mere *performance errors*. Rather, they indicate that most people's underlying *reasoning competence* is irrational or at least normatively problematic. In order to explain this view more clearly, we first need to explain the distinction between competence and performance on which it is based and say something about the normative standards of reasoning that are being assumed by advocates of this pessimistic interpretation of the experimental results.

3.1. Competence and Performance

The competence/performance distinction, as we will characterize it, was first introduced into cognitive science by Chomsky, who used it in his account of the explanatory strategy of theories in linguistics (Chomsky 1965, Ch. 1; 1975; 1980). In testing linguistic theories, an important source of data are the "intuitions" or unreflective judgments that speakers of a language make about the grammaticality of sentences, and about various linguistic properties and relations. To explain these intuitions, and also to explain how speakers go about producing and understanding sentences of their language in ordinary discourse, Chomsky and his followers proposed that a speaker of a language has an internally represented grammar of that language – an integrated set of generative rules and principles that entail an infinite number of claims about the language. For each of the infinite number of sentences in the speaker's language, the internally represented grammar entails that it is grammatical; for each ambiguous sentence in the speaker's language, the grammar entails that it is ambiguous, etc. When speakers make the judgments that we call linguistic intuitions, the information in the internally represented grammar is typically accessed and relied upon, though neither the process nor the internally represented grammar are accessible to consciousness. Since the internally represented grammar plays a central role in the production of linguistic intuitions, those intuitions can serve as an important source of data for linguists trying to specify what the rules and principles of the internally represented grammar are.

A speaker's intuitions are not, however, an infallible source of information about the grammar of the speaker's language, because the grammar cannot produce linguistic intuitions by itself. The production of intuitions is a complex process in which the internally represented grammar must interact with a variety of other cognitive mechanisms including those subserving perception, motivation, attention, short term memory and perhaps a host of others. In certain circumstances, the activity of any one of these mechanisms may result in a person offering a judgment about a sentence which does not accord with what the grammar actually entails about that sentence. This might happen when we are drunk or tired or in the grip of rage. But even under ordinary conditions when our cognitive mechanisms are not impaired in this way, we may still fail to recognize a sentence as grammatical due to limitations on attention or memory. For example, there is considerable evidence indicating that the short-term memory mechanism has difficulty handling center embedded structures. Thus it may well be the case that our internally represented grammars entail that the following sentence is grammatical:

What what what he wanted cost would buy in Germany was amazing.

even though our intuitions suggest, indeed shout, that it is not.

Now in the jargon that Chomsky introduced, the rules and principles of a speaker's internalized grammar constitutes the speaker's *linguistic competence*. By contrast, the judgments a speaker makes about sentences, along with the sentences the speaker actually produces, are part of the speaker's *linguistic performance*. Moreover, as we have just seen, some of the sentences a speaker produces and some of the judgments the speaker makes about sentences, will not accurately reflect the speaker's linguistic competence. In these cases, the speaker is making a *performance error*.

There are some obvious analogies between the phenomena studied in linguistics and those studied by philosophers and cognitive scientists interested in reasoning. In both cases there is spontaneous and largely unconscious processing of an open-ended class of inputs; people are able to understand endlessly many sentences, and to draw inferences from endlessly many premises. Also, in both cases, people are able to make spontaneous intuitive judgments about an effectively infinite class of cases – judgments about grammaticality, ambiguity, etc. in the case of linguistics, and judgments about validity, probability, etc. in the case of reasoning. Given these analogies, it is plausible to explore the idea that the mechanism underlying our ability to reason is similar to the mechanism underlying our capacity to process language. And if Chomsky is right about language, then the analogous hypothesis about reasoning would claim that people have an internally represented, integrated set of rules and principles of reasoning – a "psycho-logic" as it has been called – which is usually accessed and relied upon when people draw inferences or make judgments about them. As in the case of language, we would expect that neither the processes involved nor the principles of the internally represented psycho-logic are readily accessible to consciousness. We should also expect that people's inferences, judgments and decisions would not be an infallible guide to what the underlying psycho-logic actually entails about the validity or plausibility of a given inference. For here, as in the case of language, the internally represented rules and principles must interact with lots of other cognitive mechanisms – including attention, motivation, short term memory and many others. The activity of these mechanisms can give rise to *performance errors* – inferences, judgments or decisions that do not reflect the psycho-logic which constitutes a person's *reasoning competence.*

There is, however, an important difference between reasoning and language, even if we assume that a Chomsky-style account of the underlying mechanism is correct in both cases. For in the case of language, it makes no clear sense to offer a normative assessment of a normal person's competence. The rules and principles that constitute a French speaker's linguistic competence are significantly different from the rules and principles that underlie language processing in a Chinese speaker. But if we were asked which system was better or which one was correct, we would have no idea what was being asked. Thus, on the language side of the analogy, there are performance errors, but there is no such thing as a competence error or a normatively problematic competence. If two otherwise normal people have different linguistic competences, then they simply speak different languages or different dialects. On the reasoning side of the analogy, however, things look very different. It is not clear whether there are significant individual and group differences in the rules and principles underlying people's performance on reasoning tasks, as there so clearly are in the rules and principles underlying people's linguistic performance.[3] But if there are significant interpersonal differences in reasoning competence, it surely *appears* to make sense to ask whether one system of rules and principles is better than another.[4]

3.2. The Standard Picture

Clearly, the claim that one system of rules is superior to another assumes – if only tacitly – some standard or metric against which to measure the relative merits of reasoning systems. And this raises the normative question of what standards we ought to adopt when evaluating human reasoning. Though advocates of the pessimistic interpretation rarely offer an explicit and general normative theory of rationality, perhaps the most plausible reading of their work is that they are assuming some version of what Edward Stein calls the *Standard Picture*:

> According to this picture, to be rational is to reason in accordance with principles of reasoning that are based on rules of logic, probability theory and so forth. If the standard picture of reasoning is right, principles of reasoning that are based on such rules are normative principles of reasoning, namely they are the principles we ought to reason in accordance with. (Stein 1996, 4)

Thus the Standard Picture maintains that the appropriate criteria against which to evaluate human reasoning are rules derived from formal theories such as classical logic, probability theory and decision theory.[5] So, for example, one might derive something like the following principle of reasoning from the conjunction rule of probability theory:

> *Conjunction Principle:* One ought not to assign a lower degree of probability to the occurrence of event A than one does to the occurrence of A and some (distinct) event B. (Stein 1996, 6)

If we assume this principle is correct, there is a clear answer to the question of why the patterns of inference discussed in section 2.2 (on the "conjunction fallacy") are normatively problematic: they violate the conjunction principle. More generally, given principles of this kind, one can evaluate the specific judgments and decisions issued by human subjects and the psycho-logics that produce them. To the extent that a person's judgments and decisions accord with the principles of the Standard Picture, they are rational and to the extent that they violate such principles, the judgments and decisions fail to be rational. Similarly, to the extent that a reasoning competence produces judgments and decisions that accord with the principles of the Standard Picture, the competence is rational and to the extent that it fails to do so, it is not rational.

Sometimes, of course, it is far from clear how these formal theories are to be applied – a problem that we will return to in section 7. Moreover, as we'll see in section 8, the Standard Picture is not without its critics. Nonetheless, it does have some notable virtues. First, it seems to provide reasonably precise standards against which to evaluate human reasoning. Second, it fits very neatly with the intuitively plausible idea that logic and probability theory bear an intimate relationship to issues about how we *ought* to reason. Finally, it captures an intuition about rationality that has long held a prominent position in philosophical discussions, namely that the norms of reason are "universal principles" – principles that apply to all actual and possible cognizers irrespective of who they are or where they are located in space and time. Since the principles of the Standard Picture are derived from formal/mathematical theories – theories that, if correct, are *necessarily* correct – they appear to be precisely the sort of principles that one needs to adopt in order to capture the intuition that norms of reasoning are universal principles.

3.3 The Pessimistic Interpretation

We are now, finally, in a position to explain the pessimistic hypothesis that some authors have urged to account for the sorts of experimental results sketched in Section 2. According to this hypothesis, the errors that subjects make in these experiments are very different from the sorts of reasoning errors that people make when their memory is overextended or when their attention wanders. They are also different from the errors people make when they are tired, drunk or emotionally upset. These latter cases are all examples of *performance errors* – errors that people make when they infer in ways that are *not* sanctioned by their own psycho-logic. But, according to the pessimistic interpretation, the sorts of errors described in Section 2 are *competence errors*. In these cases people *are* reasoning, judging and making decisions in ways that accord with their psycho-logic. The subjects in these experiments do not use the right rules – those sanctioned by the Standard Picture – because they do not have access to them; they are not part of the subjects' internally represented reasoning competence. What they have instead is a collection of simpler rules or "heuristics" that may often get the right answer, though it is also the case that often they do not. So, according to this pessimistic hypothesis, the subjects make mistakes because their psycho-logic is normatively defective; their internalized rules of reasoning are less than fully rational. It is not at all clear that

Kahneman and Tversky would endorse this interpretation of the experimental results, though a number of other leading researchers clearly do.[6] According to Slovic, Fischhoff and Lichtenstein, for example, "It appears that people lack the correct programs for many important judgmental tasks.... We have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty." (1976, 174)

To sum up: According to the pessimistic interpretation, what experimental results of the sort discussed in section 2 suggest is that our reasoning is subject to systematic competence errors. But is this view warranted? Is it really the most plausible response to what we've been calling the evaluative project, or is some more optimistic view in order? In recent years, this has become one of the most hotly debated questions in cognitive science, and numerous challenges have been developed in order to show that the pessimistic interpretation is unwarranted. In the remaining sections of this paper we consider and evaluate some of the more prominent and plausible of these challenges.

4. THE CHALLENGE FROM EVOLUTIONARY PSYCHOLOGY

In recent years Gerd Gigerenzer, Leda Cosmides, John Tooby and other leading evolutionary psychologists have been among the most vocal critics of the pessimistic account of human reasoning, arguing that the evidence for human irrationality is far less compelling than advocates of the heuristics and biases tradition suggest. In this section, we will attempt to provide an overview of this recent and intriguing challenge. We start in section 4.1 by outlining the central theses of evolutionary psychology. Then in 4.2 and 4.3 we discuss how these core ideas have been applied to the study of human reasoning. Specifically, we'll discuss two psychological hypotheses – the *cheater detection hypothesis* and the *frequentist hypothesis* – and the evidence that's been invoked in support of them. Though they are ostensibly descriptive psychological claims, a number of prominent evolutionary psychologists have suggested that these hypotheses and the experimental data that has been adduced in support of them provide us with grounds for rejecting the pessimistic interpretation of human reasoning. In section 5, we consider the plausibility of this claim.

4.1 The Central Tenets of Evolutionary Psychology

Though the interdisciplinary field of evolutionary psychology is too new to have developed any precise and widely agreed upon body of doctrine, there are two theses that are clearly central. First, evolutionary psychologists endorse an account of the structure of the human mind which is sometimes called the *massive modularity hypothesis* (Sperber 1994; Samuels 1998). Second, evolutionary psychologists commit themselves to a methodological claim about the manner in which research in psychology ought to proceed. Specifically, they endorse the claim that adaptationist considerations ought to play a pivotal role in the formation of psychological hypotheses.

4.1.1 The Massive Modularity Hypothesis

Roughly stated, the massive modularity hypothesis (MMH) is the claim that the human mind is largely or perhaps even entirely composed of highly specialized cognitive mechanisms or *modules*. Though there are different ways in which this rough claim can be spelled out, the version of MMH that evolutionary psychologists defend is heavily informed by the following three assumptions:

Computationalism. The human mind is an information processing device that can be described in computational terms – "a computer made out of organic compounds rather than silicon chips" (Barkow et al. 1992, 7). In expressing this view, evolutionary psychologists clearly see themselves as adopting the *computationalism* that is prevalent in much of cognitive science

Nativism. Contrary to what has surely been the dominant view in psychology for most of the Twentieth Century, evolutionary psychologists maintain that much of the structure of the human mind is innate. Evolutionary psychologists thus reject the familiar empiricist proposal that the innate structure of the human mind consists of little more than a general-purpose learning mechanism. Instead they embrace the *nativism* associated with Chomsky and his followers (Pinker 1997).

Adaptationism. Evolutionary psychologists invariably claim that our cognitive architecture is largely the product of natural selection. On this view, our minds are composed of *adaptations* that were "invented by natural selection during the species' evolutionary history to produce adaptive ends in the species' natural environment" (Tooby and Cosmides, 1995, p. xiii). Our minds, evolutionary psychologists maintain, are designed by natural selection in order to solve *adaptive problems:* "evolutionary recurrent problem[s] whose solution promoted reproduction, however long or indirect the chain by which it did so" (Cosmides and Tooby 1994, 87).

Evolutionary psychologists conceive of modules as a type of computational mechanism – viz. computational devices that are *domain-specific* as opposed to domain-general.[7] Moreover, in keeping with their nativism and adaptationism, evolutionary psychologists also typically assume that modules are innate and that they are adaptations produced by natural selection. In what follows we will call cognitive mechanisms that posses these features *Darwinian modules.*[8] The version of MMH endorsed by evolutionary psychologists thus amounts to the claim that:

MMH. The human mind is largely or perhaps even entirely composed of a large number of Darwinian modules – innate, computational mechanisms that are domain-specific adaptations produced by natural selection.

This thesis is a far more radical than earlier modular accounts of cognition, such as the one endorsed by Jerry Fodor (Fodor 1983). According to Fodor, the modular structure of the human mind is restricted to input systems (those responsible for perception and language processing) and output systems (those responsible for producing actions). Though evolutionary psychologists accept the Fodorian thesis that such *peripheral* systems are modular in character, they maintain, *pace* Fodor, that many or perhaps even all so-called *central capacities,* such as reasoning, belief fixation and planning, can also "be divided into domain-specific modules" (Jackendoff 1992, p.70). So, for example, it has been suggested by evolutionary psychologists that there are modular mechanisms for such central processes as 'theory of mind' inference (Leslie 1994; Baron-Cohen 1995) social reasoning (Cosmides and Tooby 1992), biological categorization (Pinker 1994) and probabilistic inference (Gigerenzer 1994 and 1996). On this view, then, "our cognitive architecture resembles a confederation of hundreds or thousands of functionally dedicated computers (often called modules) designed to solve adaptive problems endemic to our hunter-gatherer ancestors" (Tooby and Cosmides 1995, xiv).

4.1.2 The Research Program of Evolutionary Psychology

A central goal of evolutionary psychology is to construct and test hypotheses about the Darwinian modules which, MMH maintains, make up much of the human mind. In pursuit of this goal, research may proceed in two quite different stages. The first, which we'll call *evolutionary analysis,* has as its goal the generation of plausible hypotheses about Darwinian modules. An evolutionary analysis tries to determine as much as possible about the recurrent, information processing problems that our forebears would have confronted in what is often called *the environment of evolutionary adaptation* or the EEA – the environment in which our ancestors evolved. The focus, of course, is on *adaptive* problems whose successful solution would have directly or indirectly contributed to reproductive success. In some cases these adaptive problems were posed by physical features of the EEA, in other cases they were posed by biological features, and in still other cases they were posed by the social environment in which our forebears were embedded. Since so many factors are involved in determining the sorts of recurrent information processing problems that our ancestors confronted in the EEA, this sort of evolutionary analysis is a highly interdisciplinary exercise. Clues can be found in many different sorts of investigations, from the study of the Pleistocene climate to the study of the social organization in the few remaining hunter-gatherer cultures. Once a recurrent adaptive problem has been characterized, the theorist may hypothesize that there is a module which would have done a good job at solving that problem in the EEA.

An important part of the effort to characterize these recurrent information processing problems is the specification of the sorts constraints that a mechanism solving the problem could take for granted. If, for example, the important data needed to solve the problem was almost always presented in a specific format, then the mechanism need not be able to handle data presented in other ways. It could "assume" that the data would be presented in the typical format. Similarly, if it was important to be able to detect people or objects with a certain property that is not readily observable, and if, in the EEA, that property was highly correlated with some other property that is easier to detect, the system could simply assume that people or objects with the detectable property also had the one that was hard to observe.

It is important to keep in mind that evolutionary analyses can only be used as a way of *suggesting plausible hypotheses* about mental modules. By themselves evolutionary analyses provide no assurance that these hypotheses are true. The fact that it would have enhanced our ancestors' fitness if they had developed a module that solved a certain problem is no guarantee that they *did* develop such a module, since there are many reasons why natural selection and the other processes that drive evolution may fail to produce a mechanism that would enhance fitness (Stich 1990, Ch. 3).

Once an evolutionary analysis has succeeded in suggesting a plausible hypothesis, the next stage in the evolutionary psychology research strategy is to *test* the hypothesis by looking for evidence that contemporary humans actually have a module with the properties in question. Here, as earlier, the project is highly interdisciplinary. Evidence can come from experimental studies of reasoning in normal humans (Cosmides 1989; Cosmides and Tooby 1992, 1996; Gigerenzer 1991a; Gigerenzer and Hug 1992), from developmental studies focused on the emergence of cognitive skills (Carey and Spelke 1994; Leslie 1994; Gelman and Brenneman 1994), or from the study of cognitive deficits in various abnormal populations (Baron-Cohen 1995). Important evidence can also be gleaned from studies in cognitive anthropology (Barkow 1992; Hutchins 1980), history, and even from such surprising areas as the comparative study of legal traditions (Wilson and Daly 1992). When evidence from a number of these areas points in the same direction, an increasingly strong case can be made for the existence of a module suggested by evolutionary analysis.

In 4.2 and 4.3 we consider two applications of this two-stage research strategy to the study of human reasoning. Though the interpretation of the studies we will sketch is the subject of considerable controversy, a number of authors have suggested that they show there is something deeply mistaken about the pessimistic hypothesis set out in Section 3. That hypothesis claims that people lack normatively appropriate rules or principles for reasoning about problems like those set out in Section 2. But when we look at variations on these problems that may make them closer to the sort of recurrent problems our forebears would have confronted in the EEA, performance improves dramatically. And this, it is argued, is evidence for the existence of at least two normatively sophisticated Darwinian modules, one designed to deal with probabilistic reasoning when information is presented in a frequency format, the other designed to deal with reasoning about cheating in social exchange settings.

4.2 The Frequentist Hypothesis

The experiments reviewed in Sections 2.2 and 2.3 indicate that in many cases people are quite bad at reasoning about probabilities, and the pessimistic interpretation of these results claims that people use simple ("fast and dirty") heuristics in dealing with these problems because their cognitive systems have no access to more appropriate principles for reasoning about probabilities. But, in a series of recent and very provocative papers, Gigerenzer (1994, Gigerenzer & Hoffrage 1995) and Cosmides and Tooby (1996) argue that from an evolutionary point of view this would be a surprising and paradoxical result. "As long as chance has been loose in the world," Cosmides and Tooby note, "animals have had to make judgments under uncertainty" (Cosmides and Tooby 1996, 14; for the remainder of this section, all quotes are from Cosmides and Tooby 1996, unless otherwise indicated). Thus making judgments when confronted with probabilistic information posed adaptive problems for all sorts of organisms, including our hominid ancestors, and "if an adaptive problem has endured for a long enough period and is important enough, then mechanisms of considerable complexity can evolve to solve it" (p. 14). But as we saw in the previous section, "one should expect a mesh between the design of our cognitive mechanisms, the structure of the adaptive problems they evolved to solve, and the typical environments that they were designed to operate in – that is, the ones that they evolved in" (p. 14). So in launching their evolutionary analysis Cosmides and Tooby's first step is to ask: "what kinds of probabilistic information would have been available to any inductive reasoning mechanisms that we might have evolved?" (p. 15)

In the modern world we are confronted with statistical information presented in many ways: weather forecasts tell us the probability of rain tomorrow, sports pages list batting averages, and widely publicized studies tell us how much the risk of colon cancer is reduced in people over 50 if they have a diet high in fiber. But information about the probability of single events (like rain tomorrow) and information expressed in percentage terms would have been rare or unavailable in the EEA.

> What *was* available in the environment in which we evolved was the encountered frequencies of actual events – for example, that we were successful 5 times out of the last 20 times we hunted in the north canyon. Our hominid ancestors were immersed in a rich flow of observable frequencies that could be used to improve decision-making, given procedures that could take advantage of them. So if we have adaptations for inductive reasoning, they should take frequency information as input. (pp. 15-16)

After a cognitive system has registered information about relative frequencies it might convert this information to some other format. If, for example, the system has noted that 5 out of the last 20 north canyon hunts were successful, it might infer and store the conclusion that there is a .25 chance that a north canyon hunt will be successful. However, Cosmides and Tooby argue, "there are advantages to storing and operating on frequentist representations because they preserve important information that would be lost by conversion to single-event probability. For example, ... the number of events that the judgment was based on would be lost in conversion. When the *n* disappears, the index of reliability of the information disappears as well." (p. 16)

These and other considerations about the environment in which our cognitive systems evolved lead Cosmides and Tooby to hypothesize that our ancestors "evolved mechanisms that took frequencies as input, maintained such information as frequentist representations, and used these frequentist representations as a database for effective inductive reasoning."[9] Since evolutionary psychologists expect the mind to contain many specialized modules, Cosmides and Tooby are prepared to find other modules involved in inductive reasoning that work in other ways.

> We are not hypothesizing that every cognitive mechanism involving statistical induction necessarily operates on frequentist principles, only that at least one of them does, and that this makes frequentist

> principles an important feature of how humans intuitively engage the statistical dimension of the world. (p. 17)

But, while their evolutionary analysis does not preclude the existence of inductive mechanisms that are not focused on frequencies, it does suggest that when a mechanism that operates on frequentist principles is engaged, it will do a good job, and thus the probabilistic inferences it makes will generally be normatively appropriate ones. This, of course, is in stark contrast to the pessimistic interpretation which claims that people simply do not have access to normatively appropriate strategies in this area.

From their hypothesis, Cosmides and Tooby derive a number of predictions:

> (1) Inductive reasoning performance will differ depending on whether subjects are asked to judge a frequency or the probability of a single event.
>
> (2) Performance on frequentist versions of problems will be superior to non-frequentist versions.
>
> (3) The more subjects can be mobilized to form a frequentist representation, the better performance will be.
>
> (4) ... Performance on frequentist problems will satisfy some of the constraints that a calculus of probability specifies, such as Bayes' rule. This would occur because some inductive reasoning mechanisms in our cognitive architecture embody aspects of a calculus of probability. (p. 17)

To test these predictions Cosmides and Tooby ran an array of experiments designed around the medical diagnosis problem which Casscells et. al. used to demonstrate that even very sophisticated subjects ignore information about base rates. In their first experiment Cosmides and Tooby replicated the results of Casscells et. al. using exactly the same wording that we reported in section 2.3. Of the 25 Stanford University undergraduates who were subjects in this experiment, only 3 (= 12%) gave the normatively appropriate bayesian answer of "2%", while 14 subjects (= 56%) answered "95%".[10]

In another experiment, Cosmides and Tooby gave 50 Stanford students a similar problem in which relative frequencies rather than percentages and single event probabilities were emphasized. The "frequentist" version of the problem read as follows:

> 1 out of every 1000 Americans has disease X. A test has been developed to detect when a person has disease X. Every time the test is given to a person who has the disease, the test comes out positive. But sometimes the test also comes out positive when it is given to a person who is completely healthy. Specifically, out of every 1000 people who are perfectly healthy, 50 of them test positive for the disease.
>
> Imagine that we have assembled a random sample of 1000 Americans. They were selected by lottery. Those who conducted the lottery had no information about the health status of any of these people.
>
> Given the information above:
>
> on average,
>
> How many people who test positive for the disease will *actually* have the disease? _____ out of _____.[11]

On this problem the results were dramatically different. 38 of the 50 subjects (= 76%) gave the correct bayesian answer.[12]

A series of further experiments systematically explored the differences between the problem used by Casscells, et al. and the problems on which subjects perform well, in an effort to determine which factors had the largest effect. Although a number of different factors affect performance, two predominate. "Asking for the answer as a frequency produces the largest effect, followed closely by presenting the problem information as frequencies." (p. 58) The most important conclusion that Cosmides and Tooby want to draw from these experiments is that "frequentist representations activate mechanisms that produce bayesian reasoning, and that this is what accounts for the very high level of bayesian performance elicited by the pure frequentist problems that we tested." (p. 59)

As further support for this conclusion, Cosmides and Tooby cite several striking results reported by other investigators. In one study, Fiedler (1988), following up on some intriguing findings in Tversky and Kahneman (1983), showed that the percentage of subjects who commit the conjunction fallacy can be radically reduced if the problem is cast in frequentist terms. In the "feminist bank teller" example, Fiedler contrasted the wording reported in 2.2 with a problem that read as follows:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> There are 100 people who fit the description above. How many of them are:
>
> bank tellers?
>
> bank tellers and active in the feminist movement?
>
> ...

In Fiedler's replication using the original formulation of the problem, 91% of subjects judged the feminist bank teller option to be more probable than the bank teller option. However in the frequentist version only 22% of subjects judged that there would be more feminist bank tellers than bank tellers. In yet another experiment, Hertwig and Gigerenzer (1994; reported in Gigerenzer 1994) told subjects that there were 200 women fitting the "Linda" description, and asked them to estimate the number who were bank tellers, feminist bank tellers, and feminists. Only 13% committed the conjunction fallacy.

Studies on over-confidence have also been marshaled in support of the frequentist hypothesis. In one of these Gigerenzer, Hoffrage and Kleinbölting (1991) reported that the sort of overconfidence described in 2.4 can be made to "disappear" by having subjects answer questions formulated in terms of frequencies. Gigerenzer and his colleagues gave subjects lists of 50 questions similar to those described in 2.4, except that in addition to being asked to rate their confidence after each response (which, in effect, asks them to judge the probability of that single event), subjects were, at the end, also asked a question about the frequency of correct responses: "How many of these 50 questions do you think you got right?" In two experiments, the average over-confidence was about 15%, when single-event confidences were compared with actual relative frequencies of correct answers, replicating the sorts of findings we sketched in Section 2.4. However, comparing the subjects' "estimated frequencies with actual frequencies of correct answers made 'overconfidence' *disappear*.... Estimated frequencies were practically identical with actual frequencies, with even a small tendency towards underestimation. The 'cognitive illusion' was gone." (Gigerenzer 1991a, 89)

4.3. The Cheater Detection Hypothesis

In Section 2.1 we reproduced one version of Wason's four card selection task on which most subjects perform very poorly, and we noted that, while subjects do equally poorly on many other versions of the selection task, there are some versions on which performance improves dramatically. Here is an example from Griggs and Cox (1982).

In its crackdown against drunk drivers, Massachusetts law enforcement officials are revoking liquor licenses left and right. You are a bouncer in a Boston bar, and you'll loose your job unless you enforce the following law:

"If a person is drinking beer, then he must be over 20 years old."

The cards below have information about four people sitting at a table in your bar. Each card represents one person. One side of a card tells what a person is drinking and the other side of the card tells that person's age. Indicate only those card(s) you definitely need to turn over to see if any of these people are breaking the law.

drinking beer	**drinking coke**	**25 years old**	**16 years old**

From a logical point of view, this problem would appear to be structurally identical to the problem in Section 2.1, but the *content* of the problems clearly has a major effect on how well people perform. About 75% of college student subjects get the right answer on this version of the selection task, while only 25% get the right answer on the other version. Though there have been dozens of studies exploring this "content effect" in the selection task, the results have been, and continue to be, rather puzzling since there is no obvious property or set of properties shared by those versions of the task on which people perform well. However, in several recent and widely discussed papers, Cosmides and Tooby have argued that an evolutionary analysis enables us to see a surprising pattern in these otherwise bewildering results. (Cosmides 1989, Cosmides and Tooby 1992)

The starting point of their evolutionary analysis is the observation that in the environment in which our ancestors evolved (and in the modern world as well) it is often the case that unrelated individuals can engage in "non-zero-sum" exchanges, in which the benefits to the recipient (measured in terms of reproductive fitness) are significantly greater than the costs to the donor. In a hunter-gatherer society, for example, it will sometimes happen that one hunter has been lucky on a particular day and has an abundance of food, while another hunter has been unlucky and is near starvation. If the successful hunter gives some of his meat to the unsuccessful hunter rather than gorging on it himself, this may have a small negative effect on the donor's fitness since the extra bit of body fat that he might add could prove useful in the future, but the benefit to the recipient will be much greater. Still, there is *some* cost to the donor; he would be slightly better off if he didn't help unrelated individuals. Despite this, it is clear that people sometimes do help non-kin, and there is evidence to suggest that non-human primates (and even vampire bats!) do so as well. On first blush, this sort of "altruism" seems to pose an evolutionary puzzle, since if a gene which made an organism *less* likely to help unrelated individuals appeared in a population, those with the gene would be slightly *more* fit, and thus the gene would gradually spread through the population.

A solution to this puzzle was proposed by Robert Trivers (1971) who noted that, while one-way altruism might be a bad idea from an evolutionary point of view, *reciprocal altruism* is quite a different matter. If a pair of hunters (be they humans or bats) can each count on the other to help when one has an abundance of food and the other has none, then they may both be better off in the long run. Thus organisms with a gene or a suite of genes that inclines them to engage in reciprocal exchanges with non-kin (or "social exchanges" as they are sometimes called) would be more fit than members of the same species without those genes. But of course, reciprocal exchange arrangements are vulnerable to cheating. In the business of maximizing fitness, individuals will do best if they are regularly offered and accept help when they need it, but never reciprocate when others need help. This suggests that if stable social exchange arrangements are to exist, the organisms involved must have cognitive mechanisms that enable them to detect cheaters, and to avoid helping them in the future. And since humans apparently are capable of entering into stable social exchange relations, this evolutionary analysis leads Cosmides and Tooby to hypothesize that we have one or more Darwinian modules whose job it is to recognize reciprocal exchange arrangements and to detect cheaters who accept the benefits in such arrangements but do not pay the costs. In short, the evolutionary analysis leads Cosmides and Tooby to hypothesize the existence of one or more cheater detection modules. We call this *the cheater detection hypothesis.*

If this is right, then we should be able to find some evidence for the existence of these modules in the thinking of contemporary humans. It is here that the selection task enters the picture. For according to Cosmides and Tooby, some versions of the selection task engage the mental module(s) which were designed to detect cheaters in social exchange situations. And since these mental modules can be expected to do their job efficiently and accurately, people do well on those versions of the selection task. Other versions of the task do not trigger the social exchange and cheater detection modules. Since we have no mental modules that were designed to deal with these problems, people find them much harder, and their performance is much worse. The bouncer-in-the-Boston-bar problem presented earlier is an example of a selection task that triggers the cheater detection mechanism. The problem involving vowels and odd numbers presented in Section 2.1 is an example of a selection task that does not trigger cheater detection module.

In support of their theory, Cosmides and Tooby assemble an impressive body of evidence. To begin, they note that the cheater detection hypothesis claims that social exchanges, or "social contracts" will trigger good performance on selection tasks, and this enables us to see a clear pattern in the otherwise confusing experimental literature that had grown up before their hypothesis was formulated.

> When we began this research in 1983, the literature on the Wason selection task was full of reports of a wide variety of content effects, and there was no satisfying theory or empirical generalization that could account for these effects. When we categorized these content effects according to whether they conformed to social contracts, a striking pattern emerged. Robust and replicable content effects were found only for rules that related terms that are recognizable as benefits and cost/requirements in the format of a standard social contract.... No thematic rule that was not a social contract had ever produced a content effect that was both robust and replicable.... All told, for non-social contract thematic problems, 3 experiments had produced a substantial content effect, 2 had produced a weak content effect, and 14 had produced no content effect at all. The few effects that were found did not replicate. In contrast, 16 out of 16 experiments that fit the criteria for standard social contracts ... elicited substantial content effects. (Cosmides and Tooby 1992, 183)

Since the formulation of the cheater detection hypothesis, a number of additional experiments have been designed to test the hypothesis and rule out alternatives. Among the most persuasive of these are a series of experiments by Gigerenzer and Hug (1992). In one set of experiments, these authors set out to show that, contrary to an earlier proposal by Cosmides and Tooby, *merely* perceiving a rule as a social contract was not enough to engage the cognitive mechanism that leads to good performance in the selection task, and that cueing for the possibility of *cheating* was required. To do this they created two quite different context stories for social contract rules. One of the stories required subjects to attend to the possibility of cheating, while in the other story cheating was not relevant. Among the social contract rules they used was the following which, they note, is widely known among hikers in the Alps:

> (i.) If someone stays overnight in the cabin, then that person must bring along a bundle of wood from the valley.

The first context story, which the investigators call the "cheating version," explained:

> There is a cabin at high altitude in the Swiss Alps, which serves hikers as an overnight shelter. Since it is cold and firewood is not otherwise available at that altitude, the rule is that each hiker who stays overnight has to carry along his/her own share of wood. There are rumors that the rule is not always followed. The subjects were cued into the perspective of a guard who checks whether any one of four hikers has violated the rule. The four hikers were represented by four cards that read "stays overnight in the cabin", "carried no wood", "carried wood", and "does not stay overnight in the cabin".

The other context story, the "no cheating version,"

> cued subjects into the perspective of a member of the German Alpine Association who visits the Swiss cabin and tries to discover how the local Swiss Alpine Club runs this cabin. He observes people bringing wood to the cabin, and a friend suggests the familiar overnight rule as an explanation. The context story also mentions an alternative explanation: rather than the hikers, the members of the Swiss Alpine Club, who do not stay overnight, might carry the wood. The task of the subject was to check four persons (the

> same four cards) in order to find out whether anyone had violated the overnight rule suggested by the friend. (Gigerenzer and Hug 1992, 142-143)

The cheater detection hypothesis predicts that subjects will do better on the cheating version than on the no cheating version, and that prediction was confirmed. In the cheating version, 89% of the subjects got the right answer, while in the no cheating version, only 53% responded correctly.

In another set of experiments, Gigerenzer and Hug showed that when social contract rules make cheating on both sides possible, cueing subjects into the perspective of one party or the other can have a dramatic effect on performance in selection task problems. One of the rules they used that allows the possibility of bilateral cheating was:

(ii.) If an employee works on the weekend, then that person gets a day off during the week.

Here again, two different context stories were constructed, one of which was designed to get subjects to take the perspective of the employee, while the other was designed to get subjects to take the perspective of the employer.

> The employee version stated that working on the weekend is a benefit for the employer, because the firm can make use of its machines and be more flexible. Working on the weekend, on the other hand is a cost for the employee. The context story was about an employee who had never worked on the weekend before, but who is considering working on Saturdays from time to time, since having a day off during the week is a benefit that outweighs the costs of working on Saturday. There are rumors that the rule has been violated before. The subject's task was to check information about four colleagues to see whether the rule has been violated. The four cards read: "worked on the weekend", "did not get a day off", "did not work on the weekend", "did get a day off".
>
> In the employer version, the same rationale was given. The subject was cued into the perspective of the employer, who suspects that the rule has been violated before. The subjects' task was the same as in the other perspective [viz. to check information about four employees to see whether the rule has been violated]. (Gigerenzer & Hug 1992, 154)

In these experiments about 75% of the subjects cued to the employee's perspective chose the first two cards ("worked on the weekend" and "did not get a day off") while less than 5% chose the other two cards. The results for subjects cued to the employer's perspective were radically different. Over 60% of subjects selected the last two cards ("did not work on the weekend" and "did get a day off") while less than 10% selected the first two.

4.4 How good is the case for the evolutionary psychological conception of reasoning?

The theories urged by evolutionary psychologists aim to provide a partial answer to the questions raised by what we've been calling the descriptive project – the project that seeks to specify the cognitive mechanisms which underlie our capacity to reason. The MMH provides a general *schema* for how we should think about these cognitive mechanisms according to which they are largely or perhaps even entirely *modular* in character. The frequentist hypothesis and cheater detection hypothesis, by contrast, make more specific claims about some of the particular modular reasoning mechanisms that we possess. Moreover, if correct, they provide some empirical support for MMH.

But these three hypotheses are (to put it mildly) very controversial and the question arises: *How plausible are they?* Though a detailed discussion of this question is beyond the scope of the present paper, we think that these hypotheses are important proposals about the mechanisms which subserve reasoning and that they ought to be taken very seriously indeed. As we have seen, the cheater detection and frequentist hypotheses accommodate an impressive array of data from the experimental literature on reasoning and do not seem *a priori* implausible. Moreover, empirical support for MMH comes not merely from the studies outlined in this section but also from a disparate range of other domains of research, including work in neuropsychology (Shallice 1989), research in cognitive developmental psychology on "theory of mind" inference (Leslie 1994; Baron-Cohen 1995) and arithmetic reasoning (Dehaene 1997). Further, as one of us has argued elsewhere, there are currently no good reasons to reject the MMH defended by evolutionary psychologists (Samuels 2000).

But when saying that the MMH, frequentist hypothesis and cheater detection hypothesis are plausible candidates that ought to be taken very seriously, we do *not* mean that they are highly confirmed. For, as far as we can see, *no* currently available theory of the mechanisms underlying human reasoning is highly confirmed. Nor, for that matter, do we mean that there are no plausible alternatives. On the contrary, each of the three hypotheses outlined in this section is merely one among a range of plausible candidates. So, for example, although all the experimental data outlined in 4.3 is compatible with the cheater detection hypothesis, many authors have proposed alternative explanations of these data and in some cases they have supported these alternatives with additional experimental evidence. Among the most prominent alternatives are the *pragmatic reasoning schemas* approach defended by Cheng, Holyoak and their colleagues (Cheng and Holyoak 1985 & 1989; Cheng, Holyoak, Nisbett and Oliver 1986) and Denise Cummins' proposal that we posses an innate, domain specific *deontic reasoning module* for drawing inferences about "permissions, obligations, prohibitions, promises, threats and warnings" (Cummins 1996, 166).[13]

Nor, when saying that the evolutionary psychological hypotheses deserve to be taken seriously, do we wish to suggest that they will require no further clarification and "fine-tuning" as enquiry proceeds. Quite the opposite, we suspect that as further evidence accumulates, evolutionary psychologists will need to clarify and elaborate on their proposals if they are to continue to be serious contenders in the quest for explanations of our reasoning capacities. Indeed, in our view, the currently available evidence already requires that the frequentist hypothesis be articulated more carefully. In particular, it is simply not the case that humans *never* exhibit systematically counter-normative patterns of inference on reasoning problems stated in terms of frequencies. In their detailed study of the conjunction fallacy, for example, Tversky and Kahneman (1983) reported an experiment in which subjects were asked to estimate both the number of "seven-letter words of the form '-----n-' in four pages of text" and the number of "seven letter words of the form '----ing' in four pages of text." The median estimate for words ending in "ing" was about three times *higher* than for words with "n" in the next-to-last position. As Kahneman and Tversky (1996) note, this appears to be a clear counter-example to Gigerenzer's claim that the conjunction fallacy disappears in judgments of frequency. Though, on our view, this sort of example does not show that the frequentist hypothesis is false, it does indicate that the version of the hypothesis suggested by Gigerenzer, Cosmides and Tooby is too simplistic. Since some frequentist representations do not activate mechanisms that produce good bayesian reasoning, there are presumably additional factors that play a role in the triggering of such reasoning. Clearly, more experimental work is needed to determine what these factors are and more subtle evolutionary analyses are needed to throw light on why these more complex triggers evolved.

To sum up: Though these are busy and exciting times for those studying human reasoning, and there is obviously much that remains to be discovered, we believe we can safely conclude from the studies recounted in this section that the evolutionary psychological conception of reasoning deserves to be taken very seriously. Whether or not it ultimately proves to be correct, the highly modular picture of the reasoning has generated a great deal of impressive research and will continue to do so for the foreseeable future. Thus we would do well to begin exploring what the implications would be for various claims about human rationality *if* the Massive Modularity Hypothesis turns out to be correct.

5. WHAT ARE THE IMPLICATIONS OF MASSIVE MODULARITY FOR THE EVALUATIVE PROJECT?

Suppose it turns out that evolutionary psychologists are right about the mental mechanisms that underlie human reasoning. Suppose that the MMH, the cheater detection hypothesis and the frequentist hypothesis are all true. How would this be relevant to what we have called the *evaluative project*? What would it tell us about the extent of human rationality? In particular, would this show that the pessimistic thesis often associated with the heuristics and biases tradition is unwarranted?

Such a conclusion is frequently suggested in the writings of evolutionary psychologists. On this view, the theories and findings of evolutionary psychology indicate that human reasoning is not subserved by "fast and dirty" heuristics but by "elegant machines" that were designed and refined by natural selection over millions of years. According to this optimistic view, concerns about systematic irrationality are unfounded. One conspicuous indication of this optimism is the title that Cosmides and Tooby chose for the paper in which they reported their data on the Harvard Medical School problem: "Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty." Five years earlier, while Cosmides and Tooby's research was still in progress, Gigerenzer reported some of their early findings in a paper with the provocative title: "How to make cognitive illusions disappear: Beyond 'heuristics and biases'." The clear suggestion, in both of these titles, is that the findings they report pose a head-on challenge to the pessimism of the heuristics and biases tradition. Nor are these suggestions restricted to titles. In paper after paper, Gigerenzer has said things like "more optimism is in order" (1991b, 245) and "we need not necessarily worry about human rationality" (1997, 280); and he has maintained that his view "supports intuition as basically rational" (1991b, 242). In light of comments like this, it is hardly surprising that one commentator has described Gigerenzer and his colleagues as having "taken an empirical stand against the view of some psychologists that people are pretty stupid" (Lopes, quoted in Bower 1996).

A point that needs to be made before we consider the implications of evolutionary psychology for the evaluative project, is that once we adopt a massively modular account of the cognitive mechanisms underlying reasoning, it becomes necessary to distinguish between two different versions of the pessimistic interpretation. The first version maintains that

> P1: Human beings make competence errors

while the second makes the claim that

> P2: *All* the reasoning competences that people possess are normatively problematic.

If we assume, contrary to what evolutionary psychologists suppose, that we possess only *one* reasoning competence, then there is little point in drawing this distinction since, for all practical purposes, the two claims will be equivalent. But, as we have seen, evolutionary psychologists maintain that we possess *many* reasoning mechanisms – different modules for different kinds of reasoning task. This naturally suggests – and indeed is interpreted by evolutionary psychologists as suggesting – that we possess *lots* of reasoning *competences*. Thus, for example, Cosmides and Tooby (1996) "suggest that the human mind may contain a series of well-engineered competences capable of being activated under the right conditions" (Cosmides and Tooby 1996, 17). For our purposes, the crucial point to notice is that once we follow evolutionary psychologists in adopting the assumption of multiple reasoning competences, P1 clearly doesn't entail P2. For even if we make *lots* of competence errors, it's clearly possible that we also possess *many* normatively *unproblematic* reasoning competences.

With the above distinction in hand, what should we say about the implications of evolutionary psychology for the pessimistic interpretation? First, under the assumption that both the frequentist hypothesis and cheater detection hypothesis are correct, we ought to reject P2. This is because, by hypothesis, these mechanisms embody normatively unproblematic reasoning competences. In which case, at least *some* of our reasoning competences will be normatively unproblematic. But do researchers within the heuristics and biases tradition really intend to endorse P2? The answer is far from clear since advocates of the pessimistic interpretation do not distinguish between P1 and P2. Some theorists have made claims that really do appear to suggest a commitment to P2[14]. But most researchers within the heuristics and biases tradition have been careful to avoid a commitment to the claim that we possess *no* normatively unproblematic reasoning competences. Moreover, it is clear that this claim simply isn't supported by the available empirical data, and most advocates of the heuristics and biases tradition are surely aware of this. For these reasons we are inclined to think that quotations which appear to support the adoption of P2 are more an indication of rhetorical excess than genuine theoretical commitment.[15]

What of P1 – the claim that human beings make competence errors when reasoning? This seems like a claim that advocates of the heuristics and biases approach really do endorse. But does the evolutionary psychological account of reasoning support the rejection of this thesis? Does it show that we make no competence errors? As far as we can tell, the answer is *No*. Even if evolutionary psychology is right in claiming that we possess *some* normatively unproblematic reasoning competences, it clearly does not follow that *no* errors in reasoning can be traced to a normatively problematic competence. According to MMH, people have *many* reasoning mechanisms and each of these modules has its own special set of rules. So there isn't one psycho-logic, there are many. In which case, the claim that we possess normatively appropriate reasoning competences for frequentist reasoning, cheater detection and perhaps other reasoning tasks is perfectly compatible with the claim that we also possess *other* reasoning modules that deploy normatively problematic principles which result in competence errors. Indeed, if MMH is true, then there will be lots of reasoning mechanisms that evolutionary psychologists have yet to discover. And it is far from clear why we should assume that these undiscovered mechanisms are normatively unproblematic. To be sure, evolutionary psychologists do maintain that natural selection would have equipped us with a number of well designed reasoning mechanisms that employ rational or normatively appropriate principles *on the sorts of problems that were important in the environment of our hunter/gatherer forebears*. However, such evolutionary arguments for the rationality of human cognition are notoriously problematic.[16] Moreover, even if we suppose that such evolutionary considerations justify the claim that we possess normatively appropriate principles for the sorts of problems that were important in the environment of our hunter/gatherer forebears, it's clear that there are many sorts of reasoning problems that are important in the modern world – problems involving the probabilities of single events, for example – that these mechanisms were not designed to handle. Indeed in many cases, evolutionary psychologists suggest, the elegant special-purpose reasoning mechanisms designed by natural selection will not even be able to process these problems. Many of the problems investigated in the "heuristics and biases" literature appear to be of this sort. And evolutionary psychology gives us no reason to suppose that people have rational inferential principles for dealing with problems like these.

To recapitulate: If the evolutionary psychological conception of our reasoning mechanisms is correct, we should reject P2 – the claim that human beings possess no normatively unproblematic reasoning competences. However, as we argued earlier, it is not P2 but P1 – the claim that we make competence errors – that advocates of the heuristics and biases program, such as Kahneman and Tversky, typically endorse. And evolutionary psychology provides us with no reason to reject *this* claim. As we will see in the sections to follow, however, the argument based on evolutionary psychology is not the only objection that's been leveled against the claim that humans make competence errors.

6. PRAGMATIC OBJECTIONS

It is not uncommon for critics of the pessimistic interpretation to point out that insufficient attention has been paid to the way in which pragmatic factors might influence how people understand the experimental tasks that they are asked to perform. One version of this complaint, developed by Gigerenzer (1996), takes the form of a very *general* objection. According to this objection, Kahneman, Tversky and others, are guilty "of imposing a statistical principle as a norm without examining content" – that is, without inquiring into how, under experimental conditions, subjects understand the tasks that they are asked to perform (Gigerenzer 1996, 593). Gigerenzer maintains that we cannot assume that people understand these tasks in the manner in which the experimenters intend them to. We cannot assume, for example, that when presented with the "feminist bank teller" problem, people understand the term "probable" as having the same meaning as it does within the calculus of chance or that the word "and" in English has the same semantics as the truth-functional operator "$\wedge$". On the contrary, depending on context, these words may be interpreted in a range of different ways. "Probable" can mean, for example, "plausible," "having the appearance of truth" and "that which may in view of present evidence be reasonably expected to happen" (ibid.). But if this is so, then according to Gigerenzer we cannot conclude from experiments on human reasoning that people are reasoning in a counter-normative fashion, since it may turn out that *as subjects understand the task* no normative principle is being violated.

There is much to be said for Gigerenzer's objection. First, he is clearly correct that, to the extent that it's possible, pragmatic factors should be controlled for in experiments on human reasoning. Second, it is surely the case that failure to do so weakens the inference from experimental data to conclusions about the way in which we reason. Finally, Gigerenzer is right to claim that insufficient attention has been paid by advocates of the heuristics and biases tradition to how people construe the experimental tasks that they are asked to perform. Nevertheless, we think that Gigerenzer's argument is of only limited value as an objection to the pessimistic interpretation. First, much the same criticism applies to the experiments run by Gigerenzer and other psychologists who purport to provide evidence for normatively *un*problematic patterns of inference. These investigators have done little more than their heuristics and biases counterparts to control for pragmatic factors. In which case, for all we know, it may be that the subjects in *these* experiments are not giving correct answers to the problems *as they understand them*, even though, given the *experimenters* understanding of the task, their responses are normatively unimpeachable. Gigerenzer's pragmatic objection is, in short, a double-edged one. If we take it too seriously, then it undermines both the experimental data for reasoning errors *and* the experimental data for *correct* reasoning.

A second, related problem with Gigerenzer's general pragmatic objection is that it is hard to see how it can be reconciled with other central claims that Gigerenzer and other evolutionary psychologists have made. If correct, the objection supports the conclusion that the experimental data do not show that people make systematic reasoning errors. But in numerous papers, Gigerenzer and other evolutionary psychologists have claimed that our performance *improves* – that "cognitive illusions" *disappear* – when probabilistic reasoning tasks are reformulated as frequentist problems. This poses a problem. How could our performance on frequentist problems be *superior* to our performance on single event tasks unless there was something *wrong* with our performance on single event reasoning problems in the first place? In order for performance on reasoning tasks to *improve,* it must surely be the case that people's performance *was* problematic. In which case, in order for the claim that performance improves on frequentist tasks to be warranted, it must also be the case that we are justified in maintaining that performance was problematic on nonfrequentist reasoning tasks.

Ad hominem arguments aside, however, there is another problem with Gigerenzer's general pragmatic objection. For unless we are extremely careful, the objection will dissolve into little more than a vague worry about the *possibility* of pragmatic explanations of experimental data on human reasoning. Of course, it's *possible* that pragmatic factors explain the data from reasoning experiments. But the objection does not provide any evidence for the claim that such factors *actually* account for patterns of reasoning. Nor, for that matter, does it provide an explanation of *how* pragmatic factors explain performance on reasoning tasks. Unless this is done, however, the significance of pragmatic objections to heuristics and biases research will only be of marginal interest.

This is not to say, however, that *no* pragmatic explanations of results from the heuristics and biases experiments have been proposed. One of the most carefully developed objections of this kind comes from Adler's discussion of the "feminist bank teller" experiment (Adler 1984). *Pace* Kahneman and Tversky, Adler denies that the results of this experiment support the claim that humans commit a systematic reasoning error – the conjunction fallacy. Instead he argues that Gricean principles of *conversational implicature* explain why subjects tend to make the apparent error of ranking (h) (Linda is a bank teller and is active in the feminist movement) as more probable than (f) (Linda is a bank teller.). In brief, Gricean pragmatics incorporates a *maxim of relevance* – a principle to the effect that an utterance should be assumed to be relevant in the specific linguistic context in which it is expressed. In the context of the "feminist bank teller" experiment, this means that if people behave as the Gricean theory predicts, they should interpret the task of saying whether or not (h) is more probable than (f) in such as way that the *description* of Linda is relevant. But *if* subjects interpret the task in the mannner intended by heuristics and biases researchers, such that:

The term "probable" functions according to the principles of probability theory,

(h) has the logical form ($A \wedge B$) and

(f) has the form A,

then the description of Linda is *not* relevant to determining which of (h) and (f) is more probable. On this interpretation, the judgment that (f) is more probable than (h) is merely a specific instance of the *mathematical truth* that for any A and any B, $P(A) \geq P(A\&B)$. Assuming that the class of bank tellers is not empty, *no* contingent information about Linda – *including* the description provided – is relevant to solving the task at hand. So, if subjects in the experiment behave like good Griceans, then they ought to *reject* the experimenter's preferred interpretation of the task in favor of some alternative on which the description of Linda *is* relevant. For example, they might construe (f) as meaning that Linda is a bank teller who is *not* a feminist. But when interpreted in this fashion, it need not be the case that (f) is more probable than (h). Indeed, given the description of Linda, it is surely more probable that Linda is a feminist bank teller than that she is a bank teller who's *not* a feminist. Thus, according to Adler, people do not violate the conjunction rule, but provide the correct answer to the question *as they interpret it.* Moreover, that they interpret it in this manner is explained by the fact that they are doing what a Gricean theory of pragmatics says that they should. On this view, then, the data from the "feminist bank teller" problem does not support the claim that we make systematic reasoning errors, it merely supports the independently plausible claim that we accord with a maxim of relevance when interpreting utterances.

On the face of it, Adler's explanation of the "feminist bank teller" experiment is extremely plausible. Nevertheless, we doubt that it is a decisive objection to the claim that subjects violate the conjunction rule in this experiment. First, the most plausible suggestion for how people might interpret the task so as to make the description of Linda relevant – i.e. interpret (f) as meaning "Linda is a bank teller who is *not* a feminist" – has been controlled for by Tversky and Kahneman and it seems that it makes no difference to whether or not the conjunction effect occurs (Tversky and Kahneman 1983, 95-6).[17] Thus some alternative account of how the task is interpreted by subjects needs to provided, and it is far from clear what the alternative might be. Second, Adler's explanation of the conjunction effect raises a puzzle about why subjects perform so much better on "frequentist" versions of the "feminist bank teller" problem (section 4.2). This is because Gricean principles of conversational implicature appear to treat the single event and frequentist versions of the problem in precisely the same manner. According to Adler, in the original single event experiment the description of Linda is irrelevant to ordering (h) and (f). In the frequentist version of the task, however, the description of Linda is also irrelevant to deciding whether more people are feminist bank tellers than feminists. Thus Adler's proposal appears to predict that the conjunction effect will also occur in the *frequentist* version of the "feminist bank teller" problem. But this is, of course, precisely what does *not* happen. Though this doesn't show that Adler's explanation of the results from the single event task is beyond repair, it does suggest that it can only be part of the story. What needs to added is an explanation of why people exhibit the conjunction effect in the single event version of the task but not in the frequentist version.

Finally, it is worth stressing that although the pragmatic explanations provided by Adler and others are of genuine interest, there are currently only a very small number of heuristics and biases experiments for which such explanations have been provided.[18] So, even if these explanations satisfactorily accounted for the results from some of the experiments, there would remain lots of results that are as yet unaccounted for in terms of pragmatic factors. Thus, as response to the pessimistic interpretation, the pragmatic strategy is insufficiently general.

7. OBJECTIONS BASED ON PROBLEMS WITH THE INTERPRETATION AND APPLICATION OF THE STANDARD PICTURE

Another sort of challenge to the pessimistic interpretation focuses on the problem of how to interpret the principles of the Standard Picture and how to apply them to specific reasoning tasks. According to this objection, many of the putative flaws in human reasoning turn on the way that the experimenters propose to understand and apply these normative principles. In the present section, we discuss three versions of this challenge. The first claims that there are almost invariably lots of equally correct ways to apply Standard Picture norms to a specific reasoning problem. The second concerns the claim that advocates of the pessimistic interpretation tend to adopt specific and highly contentious interpretations of certain normative principles – in particular, the principles of probability theory. The third objection is what we call the *derivation problem* -- the problem of explaining how normative principles are derived from such formal systems as logic, probability theory and decision-making theory.

7.1 On the multiple application of Standard Picture principles

When interpreting data from an experiment on reasoning, advocates of the pessimistic interpretation typically assume that there is a single best way of applying the norms of the Standard Picture to the experimental task. But opponents of the pessimistic interpretation have argued that this is not always the case. Gigerenzer (2000), for example, argues that there are usually several different and equally legitimate ways in which the principles of statistics and probability can be applied to a given problem and that these can yield different answers – or in some cases no answer at all. If this is correct, then obviously we cannot conclude that subjects are being irrational simply because they do not give the answer that the experimenters prefer.

There are, we think, some cases where Gigerenzer's contention is very plausible. One example of this sort can be found in the experiments on base rate neglect. (See section 2.3.) As Gigerenzer and others have argued, in order to draw the conclusion that people are violating Bayesian normative principles in these studies, one must assume that the prior probability assignments which subjects make are identical to the base-rates specified by the experimenters. But as Koehler observes:

> This assumption may not be reasonable in either the laboratory or the real world. Because they refer to subjective states of belief, prior probabilities may be influenced by base rates and any other information available to the decision-maker prior to the presentation of additional evidence. Thus, prior probabilities may be informed by base rates, but they need not be the same. (Koehler 1996)

If this is right, and we think it is, then it is a genuine empirical possibility that subjects are not violating Bayes' rule in these experiments but are merely assigning different prior probabilities from those that the experimenters expect. Nevertheless, we doubt that all (or even most) of the experiments discussed by advocates of the heuristics and biases program are subject to this sort of problem. So, for example, in the "feminist bank teller" problem, there is, as far as we can see, only one plausible way to apply the norms of probability theory to the task.[19]

7.2 On the rejection of non-frequentist interpretations of probability theory

Another way in which the pessimistic interpretation has been challenged proceeds from the observation that the principles of the Standard Picture are subject to different interpretations. Moreover, depending on how we interpret them, their scope of application will be different and hence experimental results that might, on one interpretation, count as a violation of the principles of the Standard Picture, will not count as a violation on some other interpretation. This kind of objection has been most fully discussed in connection with probability theory, where there has been a long-standing disagreement over how to interpret the probability calculus. In brief, Kahneman, Tversky and their followers insist that probability theory can be meaningfully applied to single events and hence that judgments about single events (e.g. Jack being a engineer or Linda being a bank teller) can violate probability theory. They also typically adopt a "subjectivist" or "Bayesian" account of probability which permits the assignment of probabilities to single events (Kahneman and Tversky 1996). In contrast, Gigerenzer has urged that probability theory ought to be given a frequentist interpretation according to which probabilities are construed as relative frequencies of events in one class to events in another.[20] As Gigerenzer points out, on the "frequentist view, one cannot speak of a probability unless a reference class is defined." (Gigerenzer 1993, 292-293) So, for example, "the relative frequency of an event such as death is only defined with respect to a reference class such as 'all male pub-owners fifty-years old living in Bavaria'." (ibid.) One consequence of this that Gigerenzer is particularly keen to stress is that, according to frequentism, *it makes no sense* to assign probabilities to single events. Claims about the probability of a single event are *literally meaningless*:

> For a frequentist ... the term "probability", when it refers to a *single event*, has no meaning at all for us (Gigerenzer 1991a, 88).

Moreover, Gigerenzer maintains that because of this "a strict frequentist" would argue that "the laws of probability are about frequencies and not about single events" and, hence, that "no judgment about single events can violate probability theory" (Gigerenzer 1993, 292-293).

This disagreement over the interpretation of probability raises complex and important questions in the foundations of statistics and decision theory about the scope and limits of our formal treatment of probability. The dispute between frequentists and subjectivists has been a central debate in the foundations of probability for much of the Twentieth century (von Mises 1957; Savage 1972). Needless to say, a satisfactory treatment of these issues is beyond the scope of the present paper. But we would like to comment briefly on what we take to be the central role that issues about the interpretation of probability theory play in the dispute between evolutionary psychologists and proponents of the heuristics and biases program. In particular, we will argue that Gigerenzer's use of frequentist considerations in this debate is deeply problematic.

As we have seen, Gigerenzer argues that if frequentism is true, then statements about the probability of single events are meaningless and, hence, that judgments about single events *cannot* violate probability theory (Gigerenzer 1993, 292-293). Gigerenzer clearly thinks that this conclusion can be put to work in order to dismantle part of the evidential base for the claim that human judgments and reasoning mechanisms violate appropriate norms. Both evolutionary psychologists and advocates of the heuristics and biases tradition typically view probability theory as the source of appropriate normative constraints on probabilistic reasoning. And if frequentism is true, then no probabilistic judgments about single events will be normatively problematic (by this standard) since they will not violate probability theory. In which case Gigerenzer gets to exclude all experimental results involving judgments about single events as evidence for the existence of normatively problematic probabilistic judgments and reasoning mechanisms.

On the face of it, Gigerenzer's strategy seems quite persuasive. Nevertheless we think that it is subject to serious objections. Frequentism itself is a hotly contested view, but even if we grant, for argument's sake, that frequentism is correct, there are still serious grounds for concern. First, there is a serious tension between the claim that subjects don't make errors in reasoning about single events because single event judgments do not violate the principles of probability theory (under a frequentist interpretation) and the claim – which, as we saw in section 4, is frequently made by evolutionary psychologists – that human probabilistic reasoning *improves* when we are presented with frequentist rather than single event problems. If there was *nothing wrong* with our reasoning about single event probabilities, then how could we *improve* – or do *better* – when performing frequentist reasoning tasks? As far as we can tell, this makes little sense. In which case, irrespective of whether or not frequentism is correct as an interpretation of probability theory, evolutionary psychologists cannot comfortably maintain both (a) that we don't violate appropriate norms of rationality when reasoning about the probabilities of single events and (b) that reasoning improves when single event problems are converted into a frequentist format.

A second and perhaps more serious problem with Gigerenzer's use of frequentist considerations is that it is very plausible to maintain that *even if* statements about the probabilities of single events really are meaningless and hence do not violate the probability calculus, subjects are still guilty of making *some sort of error* when they deal with problems about single events. For if, as Gigerenzer would have us believe, judgments about the probabilities of single events are meaningless, then surely the correct answer to a (putative) problem about the probability of a single event is not some numerical value or rank ordering, but rather: "Huh?" or "That's utter nonsense!" or "What on earth are you talking about?" Consider an analogous case in which you are asked a question like: "Is Linda taller than?" or "How much taller than is Linda?" Obviously these questions are nonsense because they are incomplete. In order to answer them we must be told what the other relatum of the "taller than" relation is supposed to be. Unless this is done, answering "yes" or "no" or providing a numerical value would surely be normatively inappropriate. Now according to the frequentist, the question "What is the probability that Linda is a bank teller?" is nonsense for much the same reason that "Is Linda taller than?" is. So when subjects answer the single event probability question by providing a *number* they are doing something that is clearly normatively inappropriate. The normatively appropriate answer is "Huh?", not "Less than 10 percent".

It might be suggested that the answers that subjects provide in experiments involving single event probabilities are an artifact of the demand characteristics of the experimental context. Subjects (one might claim) know, if only implicitly, that single event probabilities are meaningless. But because they are presented with forced choice problems that require a probabilistic judgment, they end up giving silly answers. Thus one might think the take-home message is "Don't blame the subject for giving a silly answer. Blame the experimenter for putting the subject in a silly situation in the first place!" But this proposal is implausible for two reasons. First, as a matter of fact, ordinary people use judgments about single event probabilities in all sorts of circumstances outside of the psychologist's laboratory. So it is implausible to think that *they* view single event probabilities as meaningless. But, second, even if subjects really did think that single event probabilities were meaningless, presumably we should expect them to provide more or less random answers and not the sorts of systematic responses that are observed in the psychological literature. Again, consider the comparison with the question "Is Linda taller than?" It would be a truly stunning result if everyone who was pressured to respond said "Yes."

7.3. The "Derivation" Problem

According to the Standard Picture, normative principles of reasoning are *derived* from formal systems such as probability theory, logic and decision theory. But this idea is not without its problems. Indeed a number of prominent epistemologists have argued that it is sufficiently problematic to warrant the rejection of the Standard Picture (Harman 1983; Goldman 1986).

One obvious problem is that there is a wide range of formal theories which make incompatible claims, and it's far from clear how we should decide which of these theories are the ones from which normative principles of reasoning ought to be derived. So, for example, in the domain of deductive logic there is first order predicate calculus, intuitionistic logic, relevance logic, fuzzy logic, paraconsistent logic and so on (Haack 1978, 1996; Priest et al. 1989; Anderson et al. 1992). Similarly, in the probabilistic domain there are, in addition to the standard probability calculus represented by the Kolmogorov axioms, various nonstandard theories, such as causal probability theory and Baconian probability theory (Nozick 1993; Cohen 1989).

Second, even if we set aside the problem of selecting formal systems and assume that there is some class of canonical theories from which normative standards ought to be derived, it is still unclear how and in what sense norms can be *derived* from these theories. Presumably they are not derived in the sense of *logically implied* by the formal theories (Goldman 1986). The axioms and theorems of the probability calculus do not, for example, logically imply we *should* reason in accord with them. Rather they merely state truths about probability – e.g. P(a) ≥ 0. Nor are normative principles "probabilistically implied" by formal theories. It is simply not the case that they make it *probable* that we ought to reason in accord with the principles. But if normative principles of reasoning are not *logically* or *probabilistically* derivable from formal theories, then in what sense *are* they derivable?

A related problem with the Standard Picture is that even if normative principles of reasoning are in *some sense* derivable from formal theories, it is far from clear that the principles so derived would be *correct.* In order to illustrate this point consider an argument endorsed by Harman (1986) and Goldman (1986) which purports to show that correct principles of reasoning cannot be derived from formal logic because the fact that our current beliefs entail (by a principle of logic) some further proposition doesn't always mean that we should believe the entailed proposition. Here's how Goldman develops the idea:

> Suppose p is entailed by q, and S already believes q. Does it follow that S ought to believe p: or even that he may believe p? Not at all… Perhaps what he ought to do, upon noting that q entails p, is abandon his belief in q! After all, sometimes we learn things that make it advisable to abandon prior beliefs. (Goldman 1986, 83)

Thus, according to Goldman, not only are there problems with trying to characterize the sense in which normative principles are *derivable* from formal theories, even if they were derivable *in some sense*, "the rules so derived would be wrong" (Goldman1986, 81).

How might an advocate of the Standard Picture respond to this problem? One natural suggestion is that normative principles are derivable *modulo* the adoption of some schema for *converting* the rules, axioms and theorems of formal systems into normative principles of reasoning – i.e. a set of rewrite or *conversion rules.* So, for example, one might adopt the following (fragment of a) conversion schema:

> Prefix all sentences in the formal language with the expression "S believes that."
> Convert all instances of "cannot" to "S is not permitted to."

Given these rules we can rewrite the conjunction rule – It cannot be the case that P(A) is less than P(A&B) – as the normative principle:

> S is not permitted to believe that P(A) is less than P(A&B).

This proposal suggests a sense in which normative principles are derivable from formal theories – a normative principle of reasoning is what one gets from applying a set of conversion rules to a statement in a formal system. Moreover, it also suggests a response to the Goldman objection outlined above. Goldman's argument purports to show that the principles of reasoning "derived" from a formal logic are problematic because it's simply not the case that we ought always to accept the logical consequences of the beliefs that we hold. But once we adopt the suggestion that it is the conjunction of a formal system *and* a set of conversion rules that permits the derivation of a normative principle, it should be clear that this kind of argument is insufficiently general to warrant the rejection of the idea that normative principles are derived from formal theories, since there may be some conversion schema which do not yield the consequence that Goldman finds problematic. Suppose, for example, that we adopt a set of conversion rules that permit us to rewrite *modus ponens* as the following principle of inference:

> If S believes that P and S believes that (If P then Q), then S should not believe that not-Q.

Such a principle does not commit us to believing the logical consequence of the beliefs that P and (If P then Q) but only requires us to avoid believing the negation of what they entail. So it evades Goldman's objection.

Nevertheless, although the introduction of conversion rules enables us to address the objections outlined above, it also raises problems of its own. In particular, it requires advocates of the Standard Picture to furnish us with an account of the correct conversion schema for rewriting formal rules as normative principles. Until such a schema is presented, the normative theory of reasoning which they purport to defend is profoundly underspecified. Moreover – and this is the crucial point – there are clearly *indefinitely many* rules that one might propose for rewriting formal statements as normative principles. This poses a dilemma for the defenders of the Standard Picture: Either they must propose a *principled* way of selecting conversion schemas or else face the prospect of an indefinitely large number of "standard pictures," each one consisting of the class of formal theories conjoined to one specific conversion scheme. The second of these options strikes us as unpalatable. But we strongly suspect that the former will be very hard to attain. Indeed, we suspect that many would be inclined to think that the problem is sufficiently serious to suggest that the Standard Picture ought to be rejected.

8. REJECTING THE STANDARD PICTURE: THE CONSEQUENTIALIST CHALLENGE

We've been considering responses to the pessimistic interpretation that assume the Standard Picture is, at least in broad outline, the correct approach to normative theorizing about rationality. But although this conception of normative standards is well entrenched in certain areas of the social sciences, it is not without its critics. Moreover, if there are good reasons to reject it, then it may be the case that we have grounds for rejecting the pessimistic interpretation as well, since the argument from experimental data to the pessimistic interpretation almost invariably assumes the Standard Picture as a normative benchmark against which our reasoning should be evaluated. In this section, we consider two objections to the Standard Picture. The first challenges the *deontological* conception of rationality implicit in the Standard Picture. The second focuses on the fact that the Standard Picture fails to take into consideration the considerable resource limitations to which human beings are subject. Both objections are developed with an eye to the fact that deontology is not the only available approach to normative theorizing about rationality.

8.1 Why be a deontologist?

According to the Standard Picture, *what it is* to be rational is to reason in accord with principles derived from formal theories, and where we fail to reason in this manner our cognitive processes are, at least to that extent, irrational. As Piattelli-Palmarini puts it:

> The universal principles of logic, arithmetic, and probability calculus ...tell us what we *should* ...think, not what we in fact think... If our intuition does in fact lead us to results incompatible with logic, we conclude that our intuition is at fault. (Piattelli-Palmarini 1994, 158)

Implicit in this account of rationality is, of course, a general view about normative standards that is sometimes called *deontology*. According to the deontologist, *what it is* to reason correctly – what's *constitutive* of good reasoning – is to reason in accord with some appropriate set of rules or principles.

However, deontology is not the only conception of rationality that one might endorse. Another prominent view, which is often called *consequentialism,* maintains that *what it is* to reason correctly, is to reason in such a way that you are likely to attain certain goals or outcomes.[21] Consequentialists are *not* rule-adverse: They do *not* claim that rules have *no* role to play in normative theories of reasoning. Rather they maintain that reasoning in accordance with some set of rules is not *constitutive* of good reasoning (Foley 1993) Though the application of rules of reasoning may be a *means* to the attainment of certain ends, what's *constitutive* of being a rational reasoning process on this view, is being an effective means of achieving some goal or range of goals. So, for example, according to one well-known form of consequentialism – *reliabilism* – a good reasoning processes is one that *tends to lead to true beliefs and the avoidance of false ones* (Goldman 1986; Nozick 1993). Another form of consequentialism – which we might call *pragmatism* – maintains that what it is for a reasoning process to be a good one is for it to be an efficient means of attaining the *pragmatic* objective of satisfying one's personal goals and desires (Stich 1990; Baron 1994).

With the above distinction between consequentialism and deontology in hand, it should be clear that one way to challenge the Standard Picture is to reject deontology in favor of consequentialism. But on what grounds might such a rejection be defended? Though these are complex issues that require more careful treatment than we can afford here, one consideration that might be invoked concerns the *value* of good reasoning. If issues about rationality and the quality of our reasoning are worth worrying about, it is presumably because whether or not we reason correctly really *matters*. This suggests what is surely a plausible desideratum on any normative theory of reasoning:

The Value Condition. A normative theory of reasoning should provide us with a *vindication* of rationality. It should explain why reasoning in a normatively correct fashion *matters* – why good reasoning is desirable.

It would seem that the consequentialist is at a distinct advantage when it comes to satisfying this desideratum. In constructing a consequentialist theory of reasoning we proceed by first identifying the goals or ends – the *cognitive goods* – of good reasoning (Kitcher 1992). So, for example, if the attainment of personal goals or the acquisition of true beliefs are of value, then they can be specified as being among the goods that we aim to obtain.[22] Having specified the appropriate ends, in order to complete the project, one needs to specify methods or processes that permit us to efficiently obtain these ends. The consequentialist approach to normative theorizing thus furnishes us with a clear explanation of why good reasoning matters: Good reasoning is reasoning that tends to result in the possession of things that we *value*.

In contrast to the consequentialist, it is far from clear how the deontologist should address the Value Condition. The reason is that it is far from clear why we should be concerned at all with reasoning according to some set of prespecified normative principles. The claim that we are concerned to accord with such principles just for the sake of doing so seems implausible.[23] Moreover, any appeal by the deontologist to the *consequences* of reasoning in a rational manner appears merely to highlight the superiority of consequentialism. Since deontologists claim that reasoning in accord with some set of rules R is *constitutive* of good reasoning, they are committed to the claim that a person who reasons in accordance with R is reasoning correctly *even if* there are more efficient ways – even better *available* ways – to attain the desirable ends. In other words, if there are contexts in which according with R is not the most efficient means of achieving the desirable ends, the deontologist is still committed to saying that it would be irrational to pursue a more efficient reasoning strategy for attaining these ends. And this poses a number of problems for the deontologist. First, since it's presumably more desirable to attain desirable ends than merely accord with R, it's very hard indeed to see how the deontologist could explain why, in this context, being rational is more valuable than not being rational. Second, the claim that rationality can mandate that we avoid efficient means of attaining desirable ends seems deeply counter-intuitive. Moreover, in contrast to the deontological conception of rationality, consequentialism seems to capture the correct intuition, namely that we should not be rationally required to accord with reasoning principles in contexts where they are ineffective as means to attaining the desirable goals. Finally, the fact that we are inclined to endorse this view suggests that we primarily value principles of reasoning only to the extent that they enable us to acquire desirable goals. It is, in short, rationality in the consequentialists sense that really matters to us.

One possible response to this challenge would be to deny that there are *any* (possible) contexts in which the rules specified by the deontological theory are *not* the most efficient way of attaining the desirable ends. Consider, for example, the claim endorsed by advocates of the Standard Picture, that what it is to make decisions rationally is to reason in accord with the principles of decision theory. If it were the case that decision theory is also the most efficient possible method for satisfying one's desires, then there would never be a context in which the theory would demand that you avoid using the most efficient method of reasoning for attaining desire-satisfaction. Moreover, the distinction between a pragmatic version of consequentialism and the deontological view under discussion would collapse. They would be little more than notational variants. But what sort of argument might be developed in support of the claim that decision theory is the most efficient means of satisfying our desires and personal goals? One interesting line of reasoning suggested by Baron (1994) is that decision theoretic principles specify the best method of achieving one's personal, pragmatic goals because a system that always reasons in accordance with these principles is *guaranteed* to maximize subjective expected utility – i.e. the subjective probability of satisfying its desires. But if this is so, then utilizing such rules provides, *in the long run*, the most likely way of satisfying one's goals and desires (Baron 1994, 319-20).[24] Though perhaps initially plausible, this argument relies heavily on an assumption that has so far been left unarticulated, namely that in evaluating a normative theory we should ignore the various *resource limitations* to which reasoners are subject. To use Goldman's term, it assumes that normative standards are *resource-independent*; that they abstract away from issues about the resources available to cognitive systems. This brings us our second objection to the Standard Picture: It ignores the resource limitations of human reasoners, or what Cherniak calls our *finitary predicament* (Cherniak 1986).

8.2 The Finitary Predicament: Resource-Relative Standards of Reasoning

Over the past thirty years or so there has been increasing dissatisfaction with resource independent criteria of rationality. Actual human reasoners suffer, of course, from a wide array of resource limitations. We are subject to limitations of time, energy, computational power, memory, attention and information. And starting with Herbert Simon's seminal work in the 1950's (Simon 1957), it has become increasingly common for theorists to insist that these limitations ought to be taken into consideration when deciding which normative standard(s) of reasoning to adopt. What this requires is that normative theories should be *relativized* to specific kinds of cognitive systems with specific resources limitations – that we should adopt a *resource-relative* or *bounded* conception of rationality as opposed to a resource-*independent* or *un*bounded one (Goldman 1986; Simon 1957). But why adopt such a conception of normative standards? Moreover, what implications does the adoption of such a view have for what we've been calling the normative and evaluative projects?

8.2.1. Resource-Relativity and the Normative Project

Though a number of objections have been leveled against resource-independent conceptions of rationality, perhaps the most commonly invoked – and to our minds most plausible – relies on endorsing some version of an *ought implies can principle* (OIC-principle). The rough idea is that just as in ethical matters our obligations are constrained by what we can do, so too in matters epistemic we are not obliged to satisfy standards that are beyond our capacities (Kitcher 1992). That is: If we cannot do A, then it is not the case that we ought to do A.[25] The adoption of such a principle, however, appears to require the rejection of the resource-independent conception of normative standards in favor of a resource-relative one. After all, it is clearly not the case that all actual and possible cognizers are able to perform the same reasoning tasks. Human beings do not have the same capacities as God or a Laplacian demon, and other (actual or possible) beings – e.g. great apes – may well have reasoning capacities that fall far short of those possessed by ordinary humans. In which case, if ought implies can, then there may be normative standards that one kind of being is obliged to satisfy where another is not. The adoption of an epistemic OIC-principle thus requires the rejection of resource-*independent* standards in favor of resource-*relative* ones.

Suppose for the moment that we accept this argument for resource-relativity. What implications does it have for what we are calling the normative project – the project of specifying how we ought to reason? One implication is that it undercuts some prominent arguments in favor of adopting the normative criteria embodied in the Standard Picture. In 8.1, for example, we outlined Baron's argument for the claim that decision theory is a normative standard because in the *long run* it provides the most likely way of satisfying one's goals and desires. Once we adopt a resource-relative conception of normative standards, however, it is far from clear that such an argument should be taken seriously. In the present context, "long run" means *in the limit* – as we approach infinite duration. But as Keynes famously observed, in the long run we will all be dead. The fact that a method of decision-making or reasoning will make it more probable that we satisfy certain goals in the long run is of little practical value to finite beings like ourselves. On a resource-relative conception of normative standards, we are concerned only with what reasoners ought to do given the resources that they possess. And infinite time is surely not one of these resources.

A second consequence of endorsing the above argument for resource-relativity is that it provides us with a *prima facie* plausible objection to the Standard Picture itself. If ought implies can, we are not obliged to reason in ways that we cannot. But the Standard Picture appears to require us to perform reasoning tasks that are far beyond our abilities. For instance, it seems to be a principle of the Standard Picture that we ought to preserve the truth-functional consistency of our beliefs. As Cherniak (1986) and others have argued, however, given even a conservative estimate of the number of beliefs we possess, this is a computationally intractable task – one that we *cannot* perform (Cherniak 1986; Stich 1990). Similar arguments have been developed against the claim, often associated with the Standard Picture, that we ought to revise our beliefs in such a way as to ensure *probabilistic coherence*. Once more, complexity considerations strongly suggest that we cannot satisfy this standard (Osherson 1996). And if we cannot satisfy the norms of the Standard Picture, then given that ought implies can, it follows that the Standard Picture is not the correct account of the norms of rationality.

Suppose, further, that we combine a commitment to the resource-relative conception of normative standards with the kind of consequentialism discussed in 8.1. This seems to have an important implication for how we think about normative standards of rationality. In particular, it requires that we *deny* that normative principles of reasoning are *universal* in two important senses. First, we are forced to deny that rules of good reasoning are universal in the sense that the same class of rules ought to be employed by all actual and possible reasoners. Rather, rules of reasoning will only be normatively correct relative to a specific kind of cognizer. According to the consequentialist, good reasoning consists in deploying efficient cognitive processes in order to achieve certain desirable goals – e.g. true belief or desire-satisfaction. The adoption of resource-relative consequentialism does *not* require that the *goals* of good reasoning be relativized to different classes of reasoners. A reliabilist can happily maintain, for example, that acquiring true beliefs and avoiding false ones is *always* the goal of good reasoning. Resource-relativity does force us, however, to concede that a set of rules or processes *for achieving this end* may be normatively appropriate for one class of organisms and not for another. After all, the rules or processes might be an *efficient* means of achieving the goal (e.g. true belief) for one kind of organism but not for the other. This, of course, is in stark contrast to the Standard Picture, which maintains that the same class of rules is the normatively correct one irrespective of the cognitive resources available to the cognizer. Thus, resource-relativity undermines one important sense in which the Standard Picture characterizes normative reasoning principles as universal, namely that they apply to *all* reasoners.

The adoption of resource-relative consequentialism also requires us to relativize our evaluations to specific ranges of *environments*. Suppose, for example, we adopt a resource-relative form of reliabilism. We will then need to specify the kind of environment relative to which the evaluation is being made in order to determine if a reasoning process is a normatively appropriate one. This is because, for various reasons, different environments can affect the efficiency of a reasoning process. First, different environments afford reasoners different kinds of *information*. To use an example we've already encountered, some environments might only contain probabilistic information that is encoded in the form of frequencies, while others may contain probabilistic information in a nonfrequentist format. And presumably it is a genuine empirical possibility that such a difference can affect the efficiency of a reasoning process. Similarly, different environments may impose different *time constraints*. In some environments there might be *lots* of time for a cognizer to execute a given reasoning procedure while in another there may be insufficient time. Again, it is extremely plausible to maintain that this will affect the efficiency of a reasoning process in attaining such goals as acquiring true beliefs or satisfying personal goals. The adoption of a resource-relative form of consequentialism thus requires that we reject the assumption that the same standards of good reasoning apply in all environments – that they are *context invariant*.

8.2.2. Resource-Relativity and the Evaluative Project

We've seen that the adoption of a resource-relative conception of normative standards by itself or in conjunction with the adoption of consequentialism has some important implications for the normative project. But what ramifications does it have for the evaluative project – for the issue of how good our reasoning is? Specifically, does it have any implications for the pessimistic interpretation?

First, does resource-relativity entail that the pessimistic interpretation is false? The short answer is clearly *no*. This is because it is perfectly compatible with resource-relativity that we fail to reason as we ought to. Indeed the adoption of a resource-relative form of *consequentialism* is entirely consistent with the pessimistic interpretation since even if such a view is correct, we might fail to satisfy the normative standards that we ought to.

But perhaps the adoption of resource-relativity implies – either by itself or in conjunction with consequentialism – that that the experimental evidence from heuristics and biases studies fails to *support* the pessimistic interpretation. Again, this strikes us as implausible. If the arguments outlined in 8.2.1 are sound, then we are not obliged to satisfy certain principles of the Standard Picture – e.g. the maintenance of truth functional consistency – since it is beyond our capacities to do so. However, it does not *follow* from this that we ought *never* to satisfy any of the principles of the Standard Picture. Nor does it follow that we ought not to satisfy them on the sorts of problems that heuristics and biases researchers present to their subjects. Satisfying the conjunction rule in the "feminist bank teller" problem, for example, clearly is not an impossible task for us to perform. In which case, the adoption of a resource-relative conception of normative standards does not show that the experimental data fails to support the pessimistic interpretation.

Nevertheless, we do think that the adoption of a resource-relative form of consequentialism renders it *extremely difficult* to see whether or not our reasoning processes are counter-normative in character. Once such a conception of normative standards is adopted, we are no longer in the position to confidently invoke familiar formal principles as benchmarks of good reasoning. Instead we must address a complex fabric of broadly conceptual and empirical issues in order to determine what the relevant standards are relative to which the quality of our reasoning should be evaluated. One such issue concerns the fact that we need to specify various parameters – e.g. the set of reasoners and the environmental range – before the standard can be applied. And it's far from clear how these parameters ought to be set or if, indeed, there is any principled way of deciding how this should be done. Consider, for example, the problem of specifying the range of environments relative to which normative evaluations are made. What range of environments should this be? Clearly there is a wide range of options. So, for instance, we might be concerned with how we perform in "ancestral environments" – the environments in which our evolutionary ancestors lived (Tooby and Cosmides 1998). Alternatively, we might be concerned with *all possible* environments in which humans might find themselves – including the experimental conditions under which heuristics and biases research is conducted. Or we might be concerned to exclude "artificial" laboratory contexts and concern ourselves only with "ecologically valid" contexts. Similarly, we might restrict contemporary environments for some purposes to those in which certain (minimal) educational standards are met. Or we might include environments in which no education whatsoever is provided. And so on. In short: there are lots of ranges of environments relative to which evaluations may be relativized. Moreover, it is a genuine empirical possibility that our evaluations of reasoning processes will be substantially influenced by how we select the relevant environments.

But even once these parameters have been fixed – even once we've specified the environmental range, for example – it still remains unclear what rules or processes we ought to deploy in our reasoning. And this is because, as mentioned earlier, it is largely an empirical issue which methods will prove to be efficient means of attaining normative ends for beings like us within a particular range of environments. Though the exploration of this empirical issue is still very much in its infancy, it is the focus of what we think is some of the most exciting contemporary research on reasoning. Most notably, Gigerenzer and his colleagues are currently exploring the effectiveness of certain reasoning methods which they call *fast and frugal algorithms* (Gigerenzer et al. 1999). As the name suggests, these reasoning processes are intended to be both speedy and computationally inexpensive and, hence, unlike the traditional methods associated with the Standard Picture, easily utilized by human beings. Nevertheless, Gigerenzer and his colleagues have been able to show that, in spite of their frugality, these algorithms are extremely reliable at performing some reasoning tasks within certain environmental ranges.[26] Indeed, they are often able to outperform computationally expensive methods such as Bayesian reasoning or statistical regression (Gigerenzer et al. 1999). If we adopt a resource-relative form of consequentialism, it becomes a genuine empirical possibility that fast and frugal methods will turn out to be the normatively appropriate ones – the ones against which our own performance ought to be judged (Bishop 2000).

9. CONCLUSION

The central goal of this paper has been to consider the nature and plausibility of the pessimistic view of human rationality often associated with the heuristics and biases tradition. We started by describing some of the more disquieting results from the experimental literature on human reasoning and explaining how these results have been taken to support the pessimistic interpretation. We then focused, in the remainder of the paper, on a range of recent and influential objections to this view that have come from psychology, linguistics and philosophy. First, we considered the evolutionary psychological proposal that human beings possess many specialized reasoning modules, some of which have access to normatively appropriate reasoning competences. We noted that although this view is not at present highly confirmed it is nevertheless worth taking very seriously indeed. Moreover, we argued that if the evolutionary psychological account of reasoning is correct, then we have good reason to reject one version of the pessimistic interpretation but *not* the version that most advocates of the heuristics and biases program typically endorse – the thesis that human beings make *competence errors*. Second, we considered a cluster of *pragmatic objections* to the pessimistic interpretation. These objections focus on the role of pragmatic, linguistic factors in experimental contexts and maintain that much of the putative evidence for the pessimistic view can be explained by reference to facts about how subjects interpret the tasks that they are asked to perform. We argued that although there is much to be said for exploring the pragmatics of reasoning experiments, the explanations that have been developed so far are not without their problems. Further, we maintained that they fail to accommodate most of the currently available data on human reasoning and thus constitute an insufficiently general response to the pessimistic view. Next, we turned our attention to objections which focus on the paired problems of interpreting and applying Standard Picture norms. We considered three such objections and suggested that they may well be sufficient to warrant considering alternatives to the Standard Picture. With this in mind, in section 8, we concluded by focusing on objections to the Standard Picture that motivate the adoption of a *consequentialist* account of rationality. In our view, the adoption of consequentialism does not imply that the pessimistic interpretation false, but it does make the task of evaluating this bleak view of human rationality an extremely difficult one. Indeed, if consequentialism is correct, we are surely a long way from being able to provide a definite answer to the central question posed by the evaluative project: We are, in other words, still unable to determine the extent to which human beings are rational.

Richard Samuels
King's College London

Stephen Stich
Rutgers University

Luc Faucher
Rutgers University

NOTES

[1] For detailed surveys of these results see Nisbett and Ross, 1980; Kahneman, Slovic and Tversky, 1982; Baron, 1994; Piattelli-Palmarini, 1994; Dawes, 1988 and Sutherland, 1994.

[2] Plous 1989 replicated this finding with an experiment in which the subjects were asked to estimate the likelihood of a nuclear war – an issue which people are more likely to be familiar with and to care about. He also showed that certain kinds of mental operations – e.g. imagining the result of a nuclear war just before making your estimate – fail to influence the process by which the estimate is produced.

[3] Though see Peng & Nisbett (in press) and Norenzayan, et al. 1999 for some intriguing evidence for the claim that there are substantial inter-cultural differences in the reasoning of human beings.

[4] Though at least one philosopher has argued that this appearance is deceptive. In an important and widely debated article, Cohen 1981 offers an account of what it is for reasoning rules to be normatively correct, and his account entails that a normal person's reasoning competence *must* be normatively correct. For discussion of Cohen's argument see Stich (1990, chapter 4) and Stein (1996, Chapter 5).

[5]Precisely what it is for a principle of reasoning to be *derived from* the rules of logic, probability theory and decision theory is far from clear, however. See section 7.3 for a brief discussion of this problem.

[6] In a frequently cited passage, Kahneman and Tversky write: "In making predictions and judgments under uncertainty, people do not appear to follow the calculus of chance or the statistical theory of prediction. Instead, they rely on a limited number of heuristics which sometimes yield reasonable judgments and sometimes lead to severe and systematic errors" (1973, p. 237). But this does not commit them to the claim that people do not follow the calculus of chance or the statistical theory of prediction *because these are not part of their cognitive competence*, and in a more recent paper they acknowledge that in *some* cases people *are* guided by the normatively appropriate rules (Kahneman and Tversky, 1996, p. 587). So presumably they do not think that people are simply ignorant of the appropriate rules, but only that they often do not exploit them when they should.

[7] To say that a cognitive structure is domain-specific means (roughly) that it is dedicated to solving a restricted class of problems in a restricted domain. For instance, the claim that there is a domain-specific cognitive structure for vision implies that there are mental structures which are brought into play in the domain of visual processing and are not recruited

in dealing with other cognitive tasks. By contrast, a cognitive structure that is *domain-general* is one that can be brought into play in a wide range of different domains.

[8] It is important to note that the notion of a Darwinian module differs in important respects from other notions of modularity to be found in the literature. First, there are various characteristics that are deemed crucial to some prominent conceptions of modularity that are not incorporated into the notion of a Darwinian module. So, for example, unlike the notion of modularity invoked in Fodor 1983, evolutionary psychologists do not insist – though, of course, they permit the possibility – that modules are *informationally encapsulated* and, hence, have access to less than all the information available to the mind as a whole. Conversely, there are features of Darwinian modules that many modularity theorists do *not* incorporate into their account of modularity. For instance, unlike to the notions of modularity employed by Chomsky and Fodor, a central feature of Darwinian modules is that they are adaptations produced by natural selection (Fodor, 1983; Chomsky, 1988). (For a useful account of the different notions of modularity see Segal, 1996. Also, see Samuels, 2000)

[9] Cosmides and Tooby call "the hypothesis that our inductive reasoning mechanisms were designed to operate on and to output frequency representations" *the frequentist hypothesis* (p. 21), and they give credit to Gerd Gigerenzer for first formulating the hypothesis. See, for example, Gigerenzer (1994, p. 142).

[10] Cosmides and Tooby use 'bayesian' with a small 'b' to characterize any cognitive procedure that reliably produces answers that satisfy Bayes' rule.

[11] This is the text used in Cosmides & Tooby's experiments E2-C1 and E3-C2.

[12] In yet another version of the problem, Cosmides and Tooby explored whether an even greater percentage would give the correct bayesian answer if subjects were forced "to actively construct a concrete, visual frequentist representation of the information in the problem." (34) On that version of the problem, 92% of subjects gave the correct bayesian response.

[13] Still other hypotheses that purport to account for the content effects in selection tasks have been proposed by Oaksford and Chater 1994, Manktelow and Over 1995 and Sperber, Cara and Girotto 1995.

[14] So, for example, Slovic, Fischhoff and Lichtenstein (1976, p. 174) claim that "It appears that people lack the correct programs for many important judgmental tasks.... We have not had the opportunity to evolve an intellect capable of dealing conceptually with uncertainty." Piattelli-Palmarini 1994 goes even further when maintaining that "we are ... blind not only to the extremes of probability but also to intermediate probabilities" – from which one might well adduce that we are simply blind about probabilities (Piattelli-Palmarini, 1994, p.131).

[15] See Samuels et al. (In press) for an extended defense of these claims.

[16] For critiques of such arguments see Stich 1990 and Stein 1996.

[17] Though, admittedly, Tversky and Kahneman's control experiment has a between-subjects design, in which (h) and (f) are not compared directly.

[18] Schwartz 1996 has invoked a pragmatic explanation of base-rate neglect which is very similar to Adler's critique of the "feminist bank teller problem" and is subject to very similar problems. Sperber et al. 1995 have provided a pragmatic explanation of the data from the selection task..

[19] This is assuming, of course, that (a) these principles apply at all (an issue we will address in section 7.2) and (b) people are not interpreting the problem in the manner suggested by Adler.

[20] On occasion, Gigerenzer appears to claim not that frequentism is *the* correct interpretation of probability theory but that it is merely one of a number of legitimate interpretations. As far as we can tell, however, this makes no difference to the two objections we consider below.

[21] Though we take consequentialism to be the main alternative to deontology, one might adopt a "virtue-based" approach to rationality. See, for example, Zagzebski 1996.

[22] Though see Stich 1990 for a challenge to the assumption that truth is something we should care about.

[23] And even if there is *some* intrinsic value to reasoning in accord with the deontologists rules, it is surely plausible to claim that the value of attaining desirable ends is greater.

[24] Actually, this argument depends on the additional assumption that one's subjective probabilities are well-calibrated – that they correspond to the objective probabilities.

[25] Though OIC-principles are widely accepted in epistemology, it is possible to challenge the way that they figure in the argument for resource-relativity. Moreover, there is a related problem of precisely which *version(s)* of this principle should be deployed in epistemic matters. In particular, it is unclear how the modal expression "can" should be interpreted. A detailed defense of the OIC-principle is, however, a long story that cannot be pursued here. See Samuels (in preparation) for a detailed discussion of these matters.

[26] One example of a fast and frugal algorithm is what Gigerenzer et al. call the *recognition heuristic*. This is the rule that: If one of two objects is recognized and the other is not, then infer that the recognized object has the higher value (Gigerenzer, et al., 1999). What Gigerenzer et al. have shown is that this very simple heuristic when combined with an appropriate metric for assigning values to objects can be remarkably accurate in solving various kinds of judgmental tasks. To take a simple example, they have shown that the recognition heuristic is an extremely reliable way of deciding which of two cities is the larger. For instance, by using the recognition heuristic a person who has never heard of Dortmund but has heard of Munich would be able to infer that Munich has the higher population, which happens to be correct. Current research suggests, however, that the value of this heuristic is not restricted to such 'toy' problems. To take one particularly surprising example, there is some preliminary evidence which suggests that people with virtually no knowledge of the stock market, using the recognition heuristic, can perform at levels equal to or better than major investment companies!

REFERENCES

Adler, J.: 1984, 'Abstraction is uncooperative', *Journal for the Theory of Social Behavior* **14**, 165-181.

Anderson, A., N. Belnap, and M. Dunn (eds.): 1992, *Entailment: The Logic of Relevance and Necessity*, Princeton University Press, Princeton.

Barkow, J.: 1992, 'Beneath new culture is old psychology: Gossip and social stratification', in Barkow, Cosmides and Tooby, 1992, pp. 627-637.

Barkow, J., L. Cosmides, and J. Tooby (eds.): 1992, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, Oxford University Press, Oxford.

Baron, J.: 1994, *Thinking and Deciding*, second edition, Cambridge University Press, Cambridge.

Baron-Cohen, S.: 1995, *Mindblindness: An Essay on Autism and Theory of Mind*, MIT Press, Cambridge.

Bishop, M. (2000): 'In praise of epistemic irresponsibility: How lazy and ignorant can you be?', in M.Bishop, R. Samuels, and S. Stich, *Perspectives on Rationality*, special issue of *Synthese* **122**, 179-208.

Bower, B.: 1996, 'Rational mind design: research into the ecology of thought treads on contested terrain', *Science News* **150**, 24-25.

Carey, S. and E. Spelke: 1994, 'Domain-specific knowledge and conceptual change', in Hirschfeld and Gelman, 1994, pp. 169-200.

Carruthers, P. and P. K. Smith: 1996, *Theories of Theories of Mind*, Cambridge University Press, Cambridge.

Casscells, W., A. Schoenberger, and T. Grayboys: 1978, 'Interpretation by physicians of clinical laboratory results', *New England Journal of Medicine* **299**, 999-1000.

Cheng, P. and K. Holyoak: 1985, 'Pragmatic reasoning schemas', *Cognitive Psychology* **17**, 391-416.

Cheng, P. and K. Holyoak: 1989, 'On the natural selection of reasoning theories', *Cognition* **33**, 285-313.

Cheng, P., K. Holyoak, R. Nisbett, and L. Oliver: 1986, 'Pragmatic versus syntactic approaches to training deductive reasoning', *Cognitive Psychology*, **18**, 293-328.

Cherniack, C.: 1986, *Minimal Rationality*, MIT Press, Cambridge.

Chomsky, N.: 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge.

Chomsky, N.: 1975, *Reflections of Language*, Pantheon Books, New York.

Chomsky, N.: 1980, *Rules and Representations*, Columbia University Press, New York.

Chomsky, N.: 1988, *Language and Problems of Knowledge: The Managua Lectures*, MIT Press, Cambridge.

Cohen, L.: 1981, 'Can human irrationality be experimentally demonstrated?', *Behavioral and Brain Sciences* **4**, 317-370.

Cohen, L.: 1989, *An Introduction to the Philosophy of Induction and Probability*, Clarendon Press, Oxford.

Cosmides, L.: 1989, 'The logic of social exchange: Has natural selection shaped how humans reason? Studies with Wason Selection Task', *Cognition* **31**, 187-276.

Cosmides, L. and J. Tooby: 1992, 'Cognitive adaptations for social exchange', in Barkow, Cosmides and Tooby, 1992, pp. 163-228.

Cosmides, L. and J. Tooby: 1994, 'Origins of domain specificity: The evolution of functional organization', in Hirschfeld and Gelman, 1994, pp. 85-116.

Cosmides, L. and J. Tooby: 1996, 'Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty', *Cognition* **58**, 1, 1-73.

Cummins, D.: 1996, 'Evidence for the innateness of deontic reasoning', *Mind and Language*, **11**, 160-190.

Deheane, S.: 1997, *The Number Sense: How the Mind Creates Mathematics*, Oxford University Press, Oxford.

Dawes, R. M.: 1988, *Rational Choice in an Uncertain World*, Harcourt, Harcourt.

Evans, J. S., S. E. Newstead, and R. M. Byrne: 1993, *Human Reasoning: The Psychology of Deduction*, Lawrence Erlbaum Associates, Hove, England.

Fiedler, K.: 1988, 'The dependence of the conjunction fallacy on subtle linguistic factors', *Psychological Research* **50**, 123-129.

Fodor, J.: 1983, *The Modularity of Mind*, MIT Press, Cambridge.

Foley, R.: 1993, *Working Without a Net: A Study of Egocentric Epistemology*, Oxford University Press, New York.

Gelman, S. and K. Brenneman: 1994, 'First principles can support both universal and culture-specific learning about number and music', in Hirschfeld and Gelman, 1994, pp. 369-387.

Gigerenzer, G.: 1991a, 'How to make cognitive illusions disappear: Beyond 'heuristics and biases'', *European Review of Social Psychology* **2**, 83-115.

Gigerenzer, G.: 1991b, 'On cognitive illusions and rationality', *Poznan Studies in the Philosophy of the Sciences and the Humanities* Vol. **21**, 225-249.

Gigerenzer, G.: 1993, 'The bounded rationality of probabilistic models', in K. I. Manktelow and D. E. Over (eds.), *Rationality: Psychological and Philosophical Perspectives*, Routledge, London.

Gigerenzer, G.: 1994, 'Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa)', in G. Wright and P. Ayton (eds.), *Subjective Probability*, John Wiley, New York.

Gigerenzer, G.: 1996, 'On narrow norms and vague heuristics: A reply to Kahneman and Tversky 1996', *Psychological Review* **103**, 592-596.

Gigerenzer, G.: 1997, 'The modularity of social intelligence', in A. Whiten and R. Byrne (eds.), *Machiavellian Intelligence* II, Cambridge University Press, Cambridge.

Gigerenzer, G.: 1998, 'Ecological intelligence: An adaptation for frequencies', in D. Cummins and C. Allen (eds.), *The Evolution of Mind*, Oxford University Press, New York.

Gigerenzer, G.: 2000, *Adaptive Thinking: Rationality in the Real World*, Oxford University Press, New York.

Gigerenzer, G. and K. Hug: 1992, 'Domain-specific reasoning: Social contracts, cheating and perspective change' *Cognition* **43**, 127-171.

Gigerenzer, G., and U. Hoffrage: 1995, 'How to improve Bayesian reasoning without instruction: Frequency formats', *Psychological Review* **102**, 684-704.

Gigerenzer, G., U. Hoffrage, and H. Kleinbslting: 1991, 'Probabilistic mental models: A Brunswikean theory of confidence', *Psychological Review* **98**, 506-528.

Gigerenzer, G., P. Todd, and the ABC Research Group: 1999, *Simple Heuristics That Make Us Smart*. Oxford University Press, New York.

Goldman, A.: 1986, *Epistemology and Cognition*, Harvard University Press, Cambridge.

Griggs, R. and J. Cox: 1982, 'The elusive thematic-materials effect in Wason's selection task', *British Journal of Psychology* **73**, 407-420.

Haack, S.: 1978, *Philosophy of Logics*, Cambridge University Press, Cambridge.

Haack, S.: 1996, *Deviant Logic, Fuzzy Logic: Beyond Formalism*, Chicago University Press, Chicago.

Harman, G.: 1983, 'Logic and probability theory versus canons of rationality', *Behavioral and Brain Sciences* **6**, p. 251.

Harman, G.: 1986, *Change of View*, MIT Press, Cambridge.

Hertwig, R. and G. Gigerenzer: 1994, 'The chain of reasoning in the conjunction task', unpublished manuscript.

Hirschfeld, L. and S. Gelman: 1994, *Mapping the Mind*, Cambridge University Press, Cambridge.

Hutchins, E.: 1980, *Culture and Inference: A Trobriand Case Study*, Harvard University Press, Cambridge.

Jackendoff, R.: 1992, *Languages of the Mind*, MIT Press, Cambridge.

Kahneman, D., P. Slovic and A. Tversky (eds.): 1982, *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

Kahneman, D. and A. Tversky: 1973, 'On the psychology of prediction', *Psychological Review* **80**, 237-251, reprinted in Kahneman, Slovic and Tversky, 1982.

Kahneman, D. and A. Tversky: 1982, 'Judgments of and by representativeness', in Kahneman, Slovic and Tversky (eds.), 1982, pp. 84-98.

Kahneman, D. and A. Tversky: 1996, 'On the reality of cognitive illusions: A reply to Gigerenzer's critique', *Psychological Review* **103**, 582-591.

Kitcher, P.: 1992, 'The naturalists return', *The Philosophical Review* **101**, no. 1, 53-114.

Koehler, J.: 1996, 'The Base-Rate Fallacy Reconsidered', *Behavioral and Brain Sciences* **19**, 1-53.

Leslie, A.: 1994, 'ToMM, ToBY, and agency: Core architecture and domain specificity', in Hirschfeld and Gelman 1994, pp. 19-148.

Lichtenstein, S., B. Fischoff and L. Phillips: 1982, 'Calibration of probabilities: The state of the art to 1980', in Kahneman, Slovic, and Tversky, 1982, pp. 306-334.

Manktelow, K. and D. Over: 1995, 'Deontic reasoning', in S. Newstead and J. St. B. Evans (eds.), *Perspectives on Thinking and Reasoning*, Erlbaum, Hillsdale, N.J.

Nisbett, R. and L. Ross: 1980, *Human Inference: Strategies and Shortcomings of Social Judgment*, Prentice-Hall, Englewood Cliffs, NJ.

Norenzayan, A., R. E. Nisbett, E. E. Smith, and B. J. Kim: 1999, *Rules vs. Similarity as a Basis for Reasoning and Judgment in East and West*, University of Michigan, Ann Arbor.

Nozick, R.: 1993, *The Nature of Rationality*, Princeton University Press, Princeton.

Oaksford, M. and N. Chater: 1994, 'A rational analysis of the selection task as optimal data selection', *Psychological Review* **101**, 608-631.

Osherson, D. N.: 1996, 'Judgement', in E.E. Smith and D. N. Osherson (eds.), *Thinking: Invitation to Cognitive Science*, MIT Press, Cambridge, MA.

Peng, K., and R. E. Nisbett (in press): 'Culture, dialectics, and reasoning about contradiction', *American Psychologist.*

Piattelli-Palmarini, M.: 1994, *Inevitable Illusions: How Mistakes of Reason Rule Our Minds*, John Wiley & Sons, New York.

Pinker, S.: 1994, *The Language Instinct*, William Morrow and Co, New York.

Pinker, S.: 1997, *How the Mind Works*, W. W. Norton, New York.

Plous, S.: 1989, 'Thinking the unthinkable: the effects of anchoring on the likelihood of nuclear war', *Journal of Applied Social Psychology* **19**, 1, 67-91.

Priest, G., R. Routley, and J. Norman (eds.): 1989, *Paraconsistent Logic: Essays on the Inconsistent*, Philosophia Verlag, München.

Samuels, R.: 1998, 'Evolutionary psychology and the massive modularity hypothesis', *British Journal for the Philosophy of Science* **49**, 575-602.

Samuels, R.: 2000, 'Massively modular minds: Evolutionary psychology and cognitive architecture', in P. Carruthers and A. Chamberlain (eds.), *Evolution and the Human Mind*, Cambridge University Press, Cambridge.

Samuels, R. (in preparation): 'Naturalism and normativity: Descriptive constraints on normative theories of rationality'.

Samuels R., S. Stich and M. Bishop (in press): 'Ending the rationality wars: How to make disputes about human rationality disappear', in R. Elio (ed.), *Common Sense, Reasoning and Rationality,* Vancouver Studies in Cognitive Science, Vol. 11, Oxford University Press, New York.

Savage, L. J.: 1972, *The Foundations of Statistics*, J. Wiley, London.

Schwarz, N.: 1996, *Cognition and Communication: Judgmental Biases, Research Methods and the Logic of Conversation*, Erlbaum, Hillsdale, NJ.

Segal, G.: 1996, 'The modularity of theory of mind', in Carruthers and Smith, 1995, pp. 141-157.

Shallice, T.: 1989, *From Neuropsychology to Mental Structures*, Cambridge University Press, Cambridge.

Simon, H. A.: 1957, *Models of Man: Social and Rational*, Wiley, New York.

Slovic, P., B. Fischhoff, and S. Lichtenstein: 1976, 'Cognitive processes and societal risk taking', in J. S. Carol and J. W. Payne (eds.), *Cognition and Social Behavior*. Erlbaum, Hillsdale, NJ.

Sperber, D.: 1994, 'The modularity of thought and the epidemiology of representations', in Hirschfeld and Gelman, 1994, pp. 39-67.

Sperber, D., F. Cara, and V. Girotto: 1995, 'Relevance theory explains the selection task', *Cognition* **57**, 1, 31-95.

Stein, E.: 1996, *Without Good Reason*, Clarendon Press, Oxford.

Stich, S.: 1990, *The Fragmentation of Reason*, MIT Press, Cambridge.

Sutherland, S.: 1994, *Irrationality: Why We Don't Think Straight!*, Rutgers University Press New Brunswick, NJ.

Tooby, J. and L. Cosmides: 1995, 'Foreword', in Baron-Cohen, 1995.

Tooby, J. and L. Cosmides: 1998, *Ecological Rationality and the Multimodular Mind*, manuscript.

Trivers, R.: 1971, 'The evolution of reciprocal altruism', *Quarterly Review of Biology* **46**, 35-56.

Tversky, A. and D. Kahneman: 1974, 'Judgment under uncertainty: Heuristics and biases', *Science* **185**, 1124-1131, reprinted in Kahneman, Slovic and Tversky, 1982.

Tversky, A. and D. Kahneman: 1983, 'Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement', *Psychological Review* **90**, 293-315.

von Misses, R.: 1957, *Probability, Statistics and Truth*, Second edition, prepared by Hilda Geiringer, Macmillan, New York.

Wilson, M. and M. Daly: 1992, 'The man who mistook his wife for a chattel', in Barkow, Cosmides and Tooby, 1992, pp. 289-322.

Zagzebski, L.: 1996, *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*, Cambridge University Press, New York.

PART II: KNOWLEDGE ACQUISITION

KEVIN KELLY

LEARNING THEORY AND EPISTEMOLOGY

1 INTRODUCTION

Learning is the acquisition of new knowledge and skills. It spans a range of processes from practice and rote memorization to the invention of entirely novel abilities and scientific theories that extend past experience. Learning is not restricted to humans: machines and animals ran learn, social organizations can learn, and a genetic population can learn through natural selection. In this broad sense, learning is adaptive change, whether in behavior or in belief.

Learning can occur through the receipt of unexpected information, as when a detective learns where the suspect resides from an anonymous informant. But it can also be a process whose arrival at a correct result is in some sense guaranteed before the new knowledge is acquired. Such a learning process may be said to be *reliable* at the time it is adopted. *Formal Learning Theory* is an a priori, mathematical investigation of this strategic conception of reliability. It does not examine how people learn or whether people actually know but rather, how reliable any system, human or otherwise, could possibly be. Thus, learning theory is related to traditional psychological and epistemological issues, but retains its own, distinct emphasis and character.

Reliability is a notoriously vague concept, suggesting a disposition to acquire new knowledge or skill over a broad range of relevantly possible environments. Learning theory deals with the vagueness not by insisting on a single, sharp "explication" of reliability, but by studying a range of possible explications, no one of which is insisted upon. This approach subtly shifts the focus from intractable debates about what reliability *is* to the more objective task of determining which precise senses of reliability are achievable in a given, precisely specified learning *problem.*

A learning problem specifies (1) what is to be learned, (2) a range of relevantly possible environments in which the learner must succeed, (3) the kinds of inputs these environments provide to the learner, (4) what it means to learn over a range of relevantly possible environments, and (5) the sorts of learning strategies that will be entertained as solutions. A learning strategy *solves* a learning problem just in case it is admitted as a potential solution by the problem and succeeds in the specified sense over the relevant possibilities. A problem is *solvable* just in case some admissible strategy solves it.

Solvability is the basic question addressed by formal learning theory. To establish a positive solvability result, one must construct an admissible learning strategy and prove that this strategy succeeds in the relevant sense. A negative result

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 183-203.
© 2004 *Kluwer Academic Publishers.*

requires a general proof that every allowable learning strategy fails. Thus, the positive results appear "methodological" whereas the negative results look "skeptical''. Negative results and positive results lock together to form a whole that is more interesting than the sum of its parts. For example, a learning method may appear unimaginative and pedestrian until it is shown that no method could do better (i.e., no harder problem is solvable). And a notion of success may sound too weak until until it is discovered that some natural problem is solvable in this sense but not in the more ambitious senses we would prefer.

There are so many different parameters in a learning problem that it is common to hold some of them fixed (e.g., the notion of success) and to allow others to vary (e.g., the set of relevantly possible environments). A partial specification of the problem parameters is called a learning *paradigm* and any problem agreeing with these specifications is an *instance* of the paradigm.

The notion of a paradigm raises more general questions. After several solvability and unsolvability results have been established in a paradigm, a pattern begins to emerge and one would like to know what it is about the combinatorial structure of the solvable problems that makes them solvable. A rigorous answer to this question is called a *characterization theorem.*

Many learning theoretic results concern the relative difficulty of two paradigms. Suppose we change a parameter (e.g., success) in one paradigm to produce another paradigm. There will usually remain an obvious correspondence between problems in the two paradigms (e.g., identical sets of serious possibilities). A *reduction* of paradigm P to another paradigm P' transforms a solution to a problem in P' into a solution to the corresponding problem in P. Then we may say that P is *no harder* than P''. Inter-reducible paradigms are *equivalent.* Equivalent paradigms may employ intuitively different standards of success, but the equivalence in difficulty shows that the quality of information provided by the diverse criteria is essentially the same. Paradigm equivalence results may therefore be viewed as epistemic analogues of the conservation principles of physics, closing the door on the temptation to get something (more reliability) for nothing by fiddling with the notion of success.

2 LEARNING IN EPISTEMOLOGY

Epistemology begins with the irritating stimulus of unlearnability arguments. For example. Sextus Empiricus records the classical problem of inductive justification as follows:

> [Dogmatists] claim that the universal is established from the particulars by means of induction. If this is so, they will effect it by reviewing either all the particulars or some of them. But if they review only some their induction will be unreliable, since it is possible that some of the particulars omitted in the induction may contradict the universal. If, on the other hand, their review is to include all the particulars, theirs will be an impossible task, because particulars are infinite and indefinite. (Sextus 1985, 105.)

This argument may be modelled in the following *data stream paradigm.* A *data stream* is just an infinite sequence e of natural numbers encoding discrete "observations". By stage n of inquiry the learner has seen observations e(0), $e(1)$,.. $e(n — 1)$. An *empirical proposition* is a proposition whose truth or falsity depends only on the data stream, and hence may be identified with a set of data. streams. A learning strategy *decides* a given empirical proposition *with certainty* just in case in each relevantly possible data stream, it eventually halts and returns the truth value of the proposition.

Let the hypothesis to be assessed be "zeros will be observed forever", which corresponds to the empirical proposition whose only element is the everywhere zero data stream. Let every Boolean-valued data stream be a relevant alternative. To show that no possible learning strategy decides the hypothesis with certainty over these alternatives, we construct a "demonic strategy" for presenting data in response to the successive outputs of an arbitrary learning strategy in such a way that the learner fails to halt with the right answer on the data stream presented. The demon presents the learner with the everywhere zero stream until the learner halts and returns "true". If this never happens, the learner fails on the everywhere zero data stream. If the learner halts with "true", there is another relevantly possible data stream that agrees with the everywhere zero data stream up to the present and that presents only ones thereafter. The demon then proceeds to present this alternative data stream, on which the learner has already halted with the wrong answer. So whatever the learner's strategy does, it fails on some relevantly possible data stream and hence does not decide the hypothesis with certainty. This is the simplest example of a negative learning theoretic argument.

The argument actually shows something stronger. *Verification* with certainty requires, asymmetrically, that the learner's strategy halt with the output "true" if the hypothesis under assessment is true and that the strategy always say "false" otherwise, possibly without ever halting. The preceding argument shows that the "zeros forever" hypothesis is not verifiable with certainty.

Karl Popper's falsificationist epistemology was originally based on the observation that although universal hypotheses cannot be verified with certainty, they can be *refuted* certainty, meaning that a method exists that halts with "false" if the hypothesis is false and that always says "true" otherwise. In the "zeros forever" example, the refutation method simply returns "true" until a nonzero value is observed and then halts inquiry with "false".

When reliability demands verification with certainty, there is no tension between the static concept of conclusive justification and the dynamical concept of reliable success, since convergence to the truth occurs precisely when conclusive justification is received. Refutation with certainty severs this tie: the learner reliably stabilizes to the truth value of *h* but when *h* is true there is no time at which this guess is certainly justified. The separation of reliablity from complete justification was hailed as a major epistemological innovation by the American Pragmatists.[1] In light of it, one may either try to invent some notion of *partial* empirical justification (e.g., a theory of *confirmation*), or one may, like Popper, side entirely with reliability.[2] Learning theory has nothing to say about whether partial epistemic justification exists or what it might be. Insofar as such notions are entertained at all, they are assessed either as components of reliable learning strategies or as extraneous constraints on admissible strategies that may make reliability more difficult or even impossible to achieve. Methodological principles with the latter property are said to be *restrictive.*[3]

"Hypothetico-deductivism" is sometimes viewed as a theory of partial inductive support (Glymour 1980), but it can also been understood as a strategy for *reducing* scientific discovery to hypothesis assessment (Popper 1968, Kemeny 1953, Putnam 1963). Suppose that the relevant possibilities are covered by a countable family of hypotheses, each of which is refutable with certainty and informative enough to be interesting. A *discovery* method produces empirical hypotheses in response to its successive observations. A discovery method *identifies* these hypotheses *in the limit* just in case on each relevantly possible data stream, the method eventually stabilizes to some true hypothesis in the family. Suppose that we have an assessment method that refutes each hypothesis with certainty. The corresponding hypothetico-deductive method is constructed as follows. It enumerates the hypotheses (by "boldness", "abduction", "plausibility", "simplicity", or the order by which they are produced by "creative intuition") and outputs the first hypothesis in the enumeration that is not rejected by the given refutation method. This reduction has occurred to just about everyone who has ever thought about inductive methodology. But things needn't be quite so easy. What if the hypotheses aren't even refutable with certainty? Could enumerating the right hypotheses occasion computational difficulties? These are just the sorts of questions of principle that are amenable to learning theoretic analysis, as will be seen below.

Another example of learning theoretic thinking in the philosophy of science is Hans Reichenbach's "pragmatic vindication" of the "straight rule" of induction (Reichenbach 1938). Reichenbach endorsed Richard von Mises' frequentist interpretation of probability. The relative frequency of an outcome in a data stream at position *n* is the number of occurrences of the outcome up to position *n* divided by *n*. The *probability* of an outcome in a data stream is the limit of the relative frequencies as *n* goes to infinity. Thus, a probabilistic statement determines an empirical proposition: the set of all data streams in which the outcome in question has the specified limiting relative frequency.

To discover limiting relative frequencies, Reichenbach recommended using the *straight rule,* whose guess at the probability of an outcome is the currently observed relative frequency of that outcome. It is immediate by definition that if the relevant possibilities include only data streams in which the limiting relative frequency of an event type is defined, then following the straight rule *gradually identifies* the true probability value, in the sense that on each relevantly possible data stream, for each nonzero distance from the probability, the conjectures of the rule eventually stay within that distance.

If the straight rule is altered to output an open interval of probabilities of fixed width centered on the observed relative frecluency, then the modified method evidently identifies a true interval in the limit (given that a probability exists). This is the same property that hypothetico-deductive inquiry has over countable collections of refutable hypotheses.

So are probability intervals refutable with certainty? Evidently not, for each finite data sequence is consistent with each limiting relative frecluency: simply extend the finite sequence with an infinite data sequence in which the probability claim is true. Is there any interesting sense in which open probability intervals can be reliably assessed? Say that a learner *decides* a hypothesis *in the limit* just in case in each relevantly possible environment, the learner eventually stabilizes to "true" if the hypothesis is true and to "false" if the hypothesis is false. According to this notion of success, the learner is guaranteed to end up with the correct truth values even though no relevantly possible environment affords certain verification or refutation. But even assuming that some limiting relative frequency exists, open probability intervals are not decidable even in this weak, limiting sense (Kelly 1996). A learner *verifies* a hypothesis *in the limit* just in case on each relevantly possible data stream, she converges to "true" if the hypothesis is true and fails to converge to "true" otherwise. This even weaker notion of success is "one sided", for when the hypothesis is true, it is only guaranteed that "false" is produced infinitely often (possibly at ever longer intervals).[4] Analogously, *refutation in the limit* requires convergence to "false" when the hypothesis is false and anything but convergence to "false" otherwise. It turns out that open probability intervals are verifiable but not decidable in the limit given that some probability (limiting relative frequency) exists.[5]

Thus, identification in the limit is possible even w-hen the possible hypotheses are merely verifiable in the limit. Indeed, identification in the limit is in general reducible to limiting verification, but the requisite reduction is a bit more complicated than the familiar hypothetico-deductive construction. Suppose we have a countable family of hypotheses covering all the relevant possibilities and a limiting verifier for each of these hypotheses. Enumerate the hypotheses so that each hypothesis occurs infinitely often in the enumeration. At a given stage of inquiry, find the first remaining hypothesis whose limiting verifier currently returns "true". If there is no such, output the first hypothesis and go to the next stage of inquiry. If there is one, output it and delete all hypotheses occurring prior to it from the hypothesis enumeration. It is an exercise to check that this method identifies a true hypothesis in the limit. So although limiting verification is an unsatisfying sense of reliable assessment, it sufficees for limiting identification. If the hypotheses form a partition, the limiting verifiability of each cell is also necessary for limiting identification (Kelly 1996). So limiting verification is perhaps more important than it might first have appeared.

Neyman and Pearson justified their theory of statistical testing in terms of the frequentist interpretation of probability:

It may often be proved that if we behave according to such a rule, then in the long run we shall reject *h* when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject *h* sufficiently often when it is false (Neyman and Pearson 1933, 142).

The significance level of a test is a fixed upper bound on the limiting relative frequency of false rejection of the hypothesis under test over all possible data streams. A test is "useless" if the limiting frequency of mistaken acceptances exceeds one minus the significance, for then we could have done better at reducing the limiting relative frequency of error by ignoring the data and flipping a coin biased according to the significance level. "Useful" testability can be viewed as a learning paradigm over data streams. How does it relate to the "qualitative" paradigms just discussed? It turns out that the existence of a useful test for a hypothesis is equivalent to the hypothesis being either verifiable or refutable in the limit (Kelly 1996). This is an example of a paradigm equivalence theorem, showing that useful statistical tests provide essentially no more "information" than limiting verification or refutation procedures, assuming the frequentist interpretation of probability.

It is standard to assume in statistical studies that the relevant probabilities exist, but is there a sense in which this claim could be reliably assessed? Demonic arguments reveal the existence of a limiting relative frequency to be neither verifiable in the limit nor refutable in the limit over arbitrary data streams. But this hypothesis is *gradually verifiable* in the sense that there is a method that outputs numbers in the unit interval such that these numbers approach one just if the hypothesis is true (Kelly 1996). A demonic argument shows that the existence of a limiting relative frequency is not *gradually refutable,* in the sense of producing a sequence of numbers approaching zero just in case the hypothesis is false.

Gradual decidability requires that the learner's outputs gradually converge to the truth value of the hypothesis whatever this truth value happens to be. Unlike gradual verification and refutation, which we have just seen to be weaker than their limiting analogues, gradual decision is inter-reducible with limiting decision: simply choose a cutoff value (e.g. 0.5) and output "true" if the current output is less than 0.5 and "false" otherwise. Gradual decision is familiar as the sense of success invoked in Bayesian convergence arguments. Since Bayesian updating by conditionalization can never retract a zero or a one on data of nonzero probablity, these outputs indicate certainty (inquiry may as well be halted), so limiting decision may only be accomplished gradually.

This short discussion illustrates how familiar epistemological issues as diverse as the problem of induction, Popper's falsificationism, Reichenbach's vindication of the straight rule, statistical testability, and Bayesian convergence all fit within a single, graduated system of learnability concepts.

3 COMPUTABLE LEARNING

The preceding discussion framed traditional epistemological topics in learning theoretic terms. But despite its ancient pedigree, the focus of formal learning theory on computational issues anchors it squarely in the present.

One of the earliest examples of a computationally driven unlearnability argument was presented by Hilary Putnam in 1963 in an article criticizing Rudolph Carnap's (1950) approach to inductive logic. Following suggestions by Wittgenstein, Carnap viewed inductive logic as a theory of "partial entailment", in which the conditional probability of the hypothesis given the data is interpreted as the proportion of logical possibilities satisfying the "premise" that also satisfy the intended "conclusion".

An inductive logic determines a *prediction* function: given the data encountered so far, output the most probable guess at the next datum to be seen. If there is a tie, we interpret this as a refusal to select a prediction and view it as a failure at this round. Since the relevant probabilities are computable in Carnap's inductive logic, so is the induced prediction function.

In the *extrapolation paradigm* the goal in each relevantly possible data stream is to eventually produce only correct predictions. Putnam showed that no computable prediction function can extrapolate the set of all total computable data streams, from which it follows that Carnap's inductive logic cannot extrapolate the computable data streams. Let an arbitrary, computable prediction strategy be given. At each stage, the demon calculates the computable prediction strategy's next prediction in light of the data already presented. If the prediction is one or greater, the demon presents a zero. If the prediction is zero, the demon presents a, one. Evidently. every prediction made by the computable extrapolator along the resulting data stream is wrong. Since both the demon's strategy and the learner's strategy are computable, this data stream is computable and hence relevantly possible.[6]

On the other hand, the problem is solved by the obvious, but noncomputable, hypothetico-deductive method. Enumerate a set of computer programs computing all and only the total computable functions (i.e., no programs that go into infinite loops are included). Each such program is computably refutable with certainty by calculating its prediction for the current stage of inquiry and rejecting it if this prediction does not agree with what is observed. This method identifies a correct program in the limit. To turn it into a reliable extrapolator, just compute what the currently output hypothesis says will happen at the next stage (another example of a paradigm reduction).

The only part of this procedure that is not computable is enumerating a collection of programs covering exactly the total computable functions. Since the prediction problem is computably unsolvable, it follows immediately that no such program enumeration is computable. So computable predictors fail on this problem "because" they cannot enumerate the right collection of hypotheses.[7]

The *computable function identification* paradigm poses the closely related problem of identifying in the limit a computer program correctly predicting each position in the data stream. The preceding hypothetico-deductive method noncomputably identifies the computable data streams in this sense, but in a seminal paper, the computer scientist E. M. Gold (1965) showed that the problem is not computably solvable. The computable demonic construction employed in the proof of this result is more subtle than in the extrapolation case, because it is a nontrivial matter for a computable demon to figure out what the computable learner's current hypothesis predicts the next datum to be. For all the demon knows, the prediction may be undefined (i.e., the hypothesis may go into an infinite loop).

The demon proceeds in stages as follows.[8] At a given stage, some data points have already been presented to the learner. The demon employs a fixed, computable enumeration of all the ordered pairs of natural numbers. He then seeks the first pair (i,j) in the enumeration such that after reading the current data followed by i zeros, the learner outputs a program that halts in j steps of computation with a prediction of zero for the next datum. If the search terminates with some such pair (i,j), then the demon adds i zeros to the data presented so far, and then presents a one (falsifying the hypothesis output by the learner after seeing the last zero). Otherwise, the demon continues searching forever and never proceeds to the next stage.

Suppose the demon's construction runs through infinitely many stages. Then the search for a pair always terminates, so the resulting data stream falsifies the learner's conjecture infinitely often. The data stream is computable because it is produced by the interaction of two computable strategies. Suppose, then, that the demon's construction eventually gets stuck at a given stage. Then the demon's search for a pair fails. So on the data stream consisting of the data presented so far followed by all zeros, the learner never produces a hypothesis that correctly predicts the next zero. This data stream is also computable: use a finite lookup table to handle the data presented so far and output zero thereafter. So in either case, the demon never identifies a correct program along some relevantly possible data stream.

Since the demon makes the learner's conjecture false infinitely often, his strategy wins even if we weaken the criterion of success to *unstable* identification in the limit, according to which the learner must eventually output only true hypotheses but need not stabilize to a particular hypothesis.[9]

Each total computer program is computably refutable with certainty (compute its successive predictions and compare them to the data), so we now know that computable refutability with certainty reduces neither computable extrapolation nor computable limiting identification. Does computable identification in the limit reduce computable extrapolation? One might suppose so: just compute tlie prediction of the limiting identifier's current conjecture, which must eventually be right since the identifiers conjectures are eventually correct. But although the limiting identifier eventually produces programs without infinite loops, nothing prevents it from producing defective programs in the short run. If a computer attempts to derive predictions from these conjectures in the manner just described, it may get caught in an infinite loop and hang for eternity.

Blum and Blum (1975) constructed a learning problem that is computably identifiable in the limit but not computably extrapolable for just this reason. Consider a problem in which an unknown Turing machine without infinite loops is hidden in a box and the successive data are the (finite) runtimes of this program on successive inputs. The learner's job is to guess some computer program whose runtimes match the observed runtimes for each input (a task suggestive of fitting a computational model to psychological reaction time data). In this problem, every program is computably refutable with certainty: simulate it and see if it halts precisely when the data say it should. Infinite loops are no problem, for one will observe in finite time that the program doesn't halt when it should have. Since the set of all programs is computably enumerable (we needn't restrict the enumeration to *total* programs this time), a computable implementation of the hypothetico-deductive strategy identifies a correct hypothesis in the limit.

Nonetheless, computable extrapolation of runtimes is not possible. Let a computable extrapolator be given. The demon is a procedure that wastes computational cycles in response to the computable predictor's last prediction. So at a given stage, the demonic program simulates the learner's program on the successive runtimes of the demonic program on earlier inputs. Whatever the prediction is, the demon goes into a wasteful subroutine that uses at least one more step of computation than the predictor expected.

Another question raised by the preceding discussion is whether stable identification is equivalent to or harder than unstable identification for computable learners in the computable function identification paradigm. This question is answered affirmatively by Case and Smith (1983). To see why the answer might be positive, consider the function identification problem in which the relevant possibilities are the "almost self-describing data streams". A unit variant of a data, stream is a partial computable function that is just like the data stream except that it may disagree or be undefined in at most one position. A data stream is *almost self-describing just* in case it is a unit variant of the function computed by the program whose index (according to a fixed, effective encoding of Turing programs into natural numbers) occurs in the data stream's first position. In other words, an "almost self-describing" data stream "gives away" a nearly correct hypothesis, but it doesn't say where the possible mismatch might be. An unstable learner can succeed by continually patching the "given away" program with ever larger lookup tables specifying what has been seen so far, since eventually the lookup table corrects the mistake in the "given away" program. But a stable learner would have to know *when* to stop patching, and this information was not given away.

In the problem just described, it is trivial to stably identify an almost correct program (just output the first datum) whereas no computable learner can stably identify an exactly correct program. Indeed, for each finite number of allowed errors there is a learning problem that is computably solvable under that error allowance but not with one fewer error (Case and Smith 83). This result, known as the *anomaly hierarchy theorem,* can be established by means of functions that are self-describing up to *n* possible errors.

There are many more sophisticated results of the kind just presented, all of which share the following points in common. (1) Uncomputability is taken just as seriously as the problem of induction from the very outset of the analysis. This is different from the approach of traditional epistemology, in which idealized logics of justification are proposed and passed along to experts in computation for advice on how to satisfy them (e.g., Levi 1991). (2) When computability is taken seriously, the *halting problem* (the formal problem of determining whether a computer program is in an infinite loop on a given input) is very similar to the classical problem of induction: for as soon as one is sure that a computation will never end, it might, for all the simulator knows *a priori,* halt at the next stage. (3) Thus, computable learners fail when ideal ones succeed because computable solvability requires the learner to solve an *internalized* problem of induction (Kelly and Schulte 1997).

4 SOME OTHER PARADIGMS

E. M. Gold's *language learnability* paradigm (1967) was intended to model child language ac-quition. In this setting, a *language* is just a computably enumerable set and a hypothesis is a code number (index) of a procedure that *accepts* all and only the members of the set.[10] Different kinds of relevantly possible environments are considered. An *informant* for a language is an enumeration of all possible strings labelled as positive or negative examples of the language. A *text* for a language is an enumeration of the elements of the language, and hence provides only positive information about membership.

Gold showed a number of results that attracted wide attention from cognitive scientists. The results for informant are similar to those for computable function identification. For example, (1) the obvious hypothetico-deductive method (non-computably) identifies all languages and (2) even the set of all computably decidable languages is not computably identifiable in the limit (the proof is similar to the one showing that the total computable functions are not identifiable in the limit). But the results for text are much weaker. For example, no collection of languages containing one infinite language and all finite subsets of that language is identifiable in the limit, even by non-computable learners.[11] Since children seem to learn language with fewer negative examples or corrections (Brown and Hanlon 1970), there have been attempts to obtain stronger positive results. For example, Wexler and Culicover (1980) modelled the environment as a presentation of context-utterance pairs, exchanging language learning from positive examples for the easier problem of computable function identification. Many other variations of the language learnability paradigm have been examined.[12]

The special difficulty with learning from text is "over-generalization", or leaping to a language that properly extends the actual language, for then no further data will correct the error. If there is no way to avoid positioning a language prior to one of its proper subsets (e.g., an infinite language must occur prior to all but finitely many of its finite subsets), hypothetico-deductivism must fail, since it will converge to the large language when one of its subsets is true. What is required is a way to use evidence to avoid overgeneralizing. This can be accomplished if (f) each possible language has a finite, characteristic sample such that once that sample is seen, the language can be produced without risk of overgeneralization. Then one may proceed by enumerating the relevantly possible grammars and conjecturing the first in the enumeration that is consistent with the data and whose characteristic sample has been observed. If no such grammar exists, stick with the preceding conjecture. Condition (f) is both necessary and sufficient for a collection of languages to be identifiable in the limit from text (Angluin 1980, Osherson et al. 1996), providing our first example of a learning theoretic *characterisation theorem.* Computable identification from text is characterized by the existence of a procedure that enumerates the characteristic sample for a language when provided with the index of a formal verification program for that language.

The *logical paradigm* (Shapiro 1981, Osherson and Weinstein 1986, 1989, Kelly and Glymour 1989, 1990), situates learning theoretic ideas in a more traditional epsistemological setting. In this paradigm, there is a, first-order language in which to frame hypotheses and the underlying world is a countable relational structure interpreting this language. An environment consists of such a structure together with a variable assignment onto the domain of the structure and an enumeration of the set of all quantifier-free formulas true under that assignment.[13] The relevant possibilities are all the environments presenting models of some theory representing the learner's background knowledge.

An hypothesis assessment method tries to guess the truth value of a particular sentence or theory in light of the increasing information provided by the environment, and successful assessment can be interpreted in any of the senses introduced above. So for example, the dense order postulate (each pair of points has a point between them) is refutable but not verifiable in the limit given as background the theory of total orders with endpoints (Osherson and Weinstein 1989).

The characterization theorem for this paradigm explains the grain of truth in the positivist's program of linking "cognitive significance" to logical form. An hypothesis is refutable (respectively, verifiable) with certainty given background theory K just in case the hypothesis is equivalent in to a sentence in prenex normal form[14] with a purely universal (respectively, existential) quantifier prefix. Similarly, an hypothesis is refutable (respectively, verifiable) in the limit given K just in case it is equivalent in K to a prenex sentence with a prefix of form $\forall\exists$ (respectively, $\exists\forall$) (Osherson and Weinstein 1989, Kelly and Glymour 1990). As one might expect, decision with certainty is possible just in case the hypothesis is equivalent to a quantifier-free sentence in K and decision in the limit (and hence gradual decision) is possible just in case the hypothesis is equivalent in K to a finite Boolean combination of purely universal and existential sentences.

A discovery method outputs theories in response to the information provided. As the goal of discovery, one can require that the method converge to the complete true theory in some fragment of the language (e.g., the purely universal sentences). *Uniform* theory identification requires that after some time the outputs of the method are true and entail the complete theory of the required fragment. For example, the complete truth is uniformly identifiable in the limit in a language with only unary predicates, but if there is a binary predicate or a unary predicate and a function symbol in the language, then neither the purely universal nor the purely existential fragment of the complete truth is identifiable in the limit (Kelly and Glymour 1989. Kelly 1996). *Nonuniform* or *pointwise* theory identification requires only that each true sentence in the specified fragment is eventually always entailed by the scientist's successive conjectures and each false sentence is eventually never entailed. The theory of all true Boolean combinations of universal and existential sentences is identifiable in the limit in this sense. Thus, nonuniform theory identification provides a logical conception of scientific progress that, unlike Popper's "deductivist" epistemology, treats verifiable and refutable hypotheses symmetrically.

Nonuniform theory identification bears on another Popperian difficulty. Popper held that hypothetico-deductivism leads us ever closer to the truth in the limit. David Miller (1974) argued that "closeness" to the truth is not a semantic notion since it is not preserved under translation. Thomas Mormann (1988) traced the difficulty to mathematics: translation is a type of topological equivalence, but topological equivalence permits "stretching" and hence does not preserve distance (e.g., verisimilitude). Nonuniform identification is a topological rather than a metrical notion, and hence is preserved under translation, thereby avoiding Miller-style objections. Nonetheless it constitutes a nontrivial account of scientific progress toward the complete truth that does not imply that any future theory produced by science will be literally true.

5 RELIABILITY AND COMPLEXITY

Learnability is a matter of how the possible futures making different hypotheses correct branch off from one another through time. The more complex the temporal entanglement of the futures satisfying incompatible hypotheses, the more difficult learning will be. Learnability is governed by the *topological* complexity of the possible hypotheses and computable learnability depends on their *computational* complexity.[15]

Data streams can be topologized in an epistemologically relevant manner as follows. A *fan* of data streams is the set of all data streams extending some finite data sequence, which we may call the *handle* of the fan. A fan with a given handle is just the empirical proposition asserting that the handle has occurred in the data. An empirical proposition is *open* just in case it is a union of fans and is *closed* just in case its complement is open.[16] Then we have the following characterization: an empirical proposition is verifiable with certainty just in case it is open, is refutable with certainty just in case it is closed, and is decidable with certainty just in case it is both closed and open. For suppose that a hypothesis is open. To verify it with certainty, just wait until the observed data sequence is the handle of a fan contained in the hypothesis and halt inquiry with "true". Conversely, if a given method verifies a hypothesis with certainty, the hypothesis can be expressed as the union of all fans whose handles are finite data sequences on which the method halts with "true".

To characterize limiting and gradual success, topological generalizations of the open and closed propositions are required. Call the open and closed propositions the Σ_1 and Π_1 propositions, respectively. For each n, the Σ_{n+1} propositions are countable unions of Π_n propositions and the Π_{n+1} propositions are countable intersections of Σ_n propositions. At each level n, a proposition is Δ_n just in case it is both Π_n and Σ_n. These are known as the *finite Borel* complexity classes, which have been familiar in functional analysis since early in this century (Hinman 1978). Then it can be shown that limiting verifiability, refutability, and decidability are characterized by Π_2, Σ_2 and Δ_2, respectively. It can also be shown that when the hypotheses are mutually incompatible, stable identification in the limit is characterized by each hypothesis being Σ_2.[17]

In computable inquiry, attaching hypotheses to propositions is a nontrivial matter, so instead of bounding the complexity of empirical propositions, we must consider the overall *correctness relation* $C(e, h)$ indicating that hypothesis h is correct in environment e. In computable function identification, for example, correctness requires that h be the index of a computer program that computes e. In language learning from text, h must be the index of a positive test procedure for the range of e. By suitable coding conventions, language learning from informant and logical learning can also modelled with correctness relations in the data stream paradigm. Computational analogs of the Borel complexity classes can be defined for correctness relations, in which case analogous characterization theorems hold for computable inquiry (Kelly 1996).

The moral of this discussion is that the problem of induction, or empirical underdeterination, comes in degrees corresponding to standard topological and computational complexity classes, which determine the objective sense in which reliable inquiry is possible.

6 A FOOLISH CONSISTENCY

A *consistent* learner never produces an output that is incorrect of every relevantly possible data stream extending the current data sequence. For non-computable learners, consistency makes a great deal of sense: why should someone who aims to find the truth say what has to be wrong? On the other hand, we have seen that formal relations can pose an "internal" problem of induction for computable learners. Since we do not require omniscience on the empirical side, why should we do so on the formal side when the underlying structure of the problem of induction is the same on both sides?

This raises an interesting question. Could insistence on computationally achievable consistency *preclude* computationally achievable empirical reliability? The answer is striking. One can construct an empirical proposition with the following properties. (1) The proposition is computably refutable with certainty. (2) Some computable, consistent method exists for the proposition (the method that always says "false" suffices since the proposition is never verified). But (3) a consistent, computable method of even a highly idealized, uncomputable kind[18] can even gradually decide the hypothesis. Thus, where traditional epistemology sees consistency as a *means* for finding the truth sooner, enforcing *achievable* consistency may prevent computable learners from finding truths they could otherwise have reliably found. So if the aim of inquiry is to find the truth, inconsistency may be an epistemic *obligation* (rather than a merely forgivable lapse) for computable agents. Such results exemplify the sharp difference in emphasis between computational learning theory and traditional, justificationist epistemology.[19]

7 GAMBLING WITH SUCCESS

Suppose that each learning problem comes equipped with an assignment of probabilities to empirical propositions. More precisely, suppose that the probability assignment is defined on the set of all *Borel propositions* (i.e., the least set that contains all the open (Σ_i) propositions and that is closed under countable union and complementation). A *probability assignment* on the Borel propositions is a function taking values in the unit interval that assigns unity to the vacuous proposition and that is *finitely additive* in the sense that the probability of a finite union of mutually incompatible Borel propositions is the sum of the probabilities of the propositions the union is taken over. *Countable additivity* extends finite additivity to countable, disjoint unions. While Kolmogorov's familiar mathematical theory of probability assumes countable additivity as a postulate, limiting relative frecinencies do not satisfy it and the usual foundations of Bayesian probability theory do not entail it (e.g., DeFinetti 1990. Savage 1972).

Say that an hypothesis is gradually decidable *with probability r* just in case there exists some empirical proposition of probability *r* over which the hypothesis is gradually decidable in the usual sense, and similarly for the other assessment criteria. Probabilistic success can be much easier to achieve than success in each relevant possibility. If the probability assignment is countably additive, then, remarkably, every Borel hypothesis is (1) decidable in the limit with unit probability and (2) decidable with certainty with arbitrarily high but non-unit probability. (1) can be improved to the result that tlie method of updating the given probability measure by conditionalization gradually decides the hypothesis with unit prior probability (e.g., Halmos 1970). This is a very general version of the familiar Bayesian claim that prior probabilities are eventually "swamped" by the data.

Compared with the purely topological analysis of section 5, these probabilistic results seem almost too good to be true, since Borel propositions can be infinitely more complex than Δ_2 propositions (Hinman 1978). What accounts for the dramatic difference? Suppose we want to decide the "zeros forever" hypothesis with a given, nonzero probability r. The negation of this hypothesis is the countable, disjoint union of the hypotheses h_i = "the first nonzero occurs at position i". So by countable additivity, the probability that the "zeros forever" hypothesis is false is the sum of the probabilities of the propositions h_i. Since the infinite sum converges to a finite value, there is some position n such that the sum of the probabilities of $h_n, h_{n+1}, \ldots$ is less than r. So our probability of failure is less than r if we halt with "true" at stage n if no nonzero datum lias been seen by position n and halt with "false" as soon as a nonzero datum is seen. In other words, countable additivity asserts that when a high prior probability of successful learning suffices, only finitely many of the demon's opportunities to make the hypothesis false matter.

Without countable additivity, it is possible that the probability that the hypothesis is false exceeds the mass distributed over the $h_{n,\ldots}$, say by a value of r. Since this "residual" probability mass is not distributed over the propositions h_i, the learner never "gets past" it, so whenever the learner halts inquiry with "true", the probability that this conclusion was in error remains at least as high as r. The residual probability reflects the demon's *inexhaustible* opportunities to falsify the hypothesis in tlie infinite future, providing a probabilistic model of Sextus' demonic argument. In fact, both (1) and (2) can fail when countable additivity is dropped (Kelly 1996) highlighting the pivotal epistemological significance of this questionable and somewhat "technical" looking assumption.

8 CONCEPT LEARNING AND THE PAC PARADIGM

In the *Meno,* Plato outlined what has come to be known as the *concept learning paradigm,* which has captured the imagination of philosophers, psychologists, and artificial intelligence researchers ever since. A concept learning problem specifies a domain of *examples* described as vectors of *values* (e.g., blue, five kilos) of a corresponding set of *attributes* (e.g., color, weight), together with a set of possible *target concepts,* which are sets of examples. The learner is somehow presented with examples labelled either as positive or as negative examples of the concept to be learned, and the learner's task is to converge in some specified sense to a correct definition. In contemporary artificial intelligence and cognitive science, the "concepts" to be learned are defined by neural networks, logic circuits, and finite state automata, but the underlying paradigm would still to familiar to Socrates.

Socrates ridiculed students who proposed disjunctive concept definitions, which suggests that he admitted only conjunctively definable concepts as relevant possibilities. Socrates' solution to the problem was to have the environment "give away" the answer in a mystical flash of insight. But J. S. Mill's (i.e., Francis Bacon's) well-known inductive methods need no mystical help to identify conjunctive concepts with certainty: the first conjecture is the first positive example sampled. On each successive positive example in the sample, delete from the current conjecture each conjunct that disagrees with the corresponding attribute value of the example (the "method of difference"). On each successive negative example that agrees with the current conjecture everywhere except on one attribute, underline the value of that attribute in the current conjecture (the "method of similarity"). When all conjuncts in the current conjecture are underlined, halt inquiry.

Boolean concepts are also identifiable with certainty over a finite set of attribute values: wait for all possible examples to come in and then disjoin the positive ones. Bacon's methods sound plausible in the conjunctive case, but this "jerrymandering" procedure for learning Boolean concepts sounds hopeless (it is, in fact, just what Socrates ridiculed). Yet both procedures identify the truth with certainty since the set of examples is finite. The PAC (Probably Approximately Correct) paradigm distinguishes such "small" problems in terms of *tractable* rather than merely *computable* inquiry.[20]

In the PAC paradigm, examples are sampled with replacement from an urn in which the probability of selecting an example is unknown. There is a collection of relevantly possible concepts and also a collection of hypotheses specifying the possible forms in which the learner is permitted to define a relevantly possible concept. Say that a hypothesis is ε-accurate just in case the sampling probability that a single sampled individual is a counterexample is less than ε. The learner is given a *confidence* parameter φ and an *error* parameter ε. From these parameters, the learner specifies a sample size and upon inspecting the resulting sample, she outputs a hypothesis. A learning strategy is *probably approximately correct* (PAC') just in case for each probability distribution on the urn and for each ε, φ exceeding zero, the strategy has a probability of at least 1 — ε of producing an ε-accurate hypothesis.

It remains to specify what it means for a PAC learning strategy to be *efficient.* Computational complexity is usually analyzed in terms of asymptotic growth rate over an infinite sequence of "similar" but "ever larger" examples of the problem. Tractability is understood as resource consumption bounded almost everywhere by some polynomial function of problem size. The size of a concept learning problem is determined by (1) the number of attributes, (2) the size of the smallest definition of the target concept, (3) the reciprocal of the confidence parameter, and (4) the reciprocal of the error parameter (higher accuracy and reliability requirements make for a "bigger" inference problem). A *data efficient* PAC learner takes a sample in each problem whose size is bounded by a polynomial in these four arguments.

There is an elegant combinatorial characterization of how large the sample required for PAC learning should be. Say that a concept class *shatters* a set *S* of examples just in case each subset of S is the intersection of *S* with some concept in the class. The *Vapnik-Chervonenkis* (VC) dimension of the concept class is the cardinality of the largest set of instances shattered by the class. There exists a fixed constant c such that if the VC dimension of the concept class is d, it suffices for PAC learnability that a sample of size s be taken, where

$$s \geq c(\frac{1}{\varepsilon}\log\frac{1}{\delta}+\frac{d}{\varepsilon}\log\frac{1}{\varepsilon}).$$

For example, the VC dimension of the conjunctive concepts over n Boolean attributes is $2n$ (and in fact is just n if $n > 1$) so the problem is data-efficiently solvable by setting the sample size according to the above formula and then using any method producing conjectures consistent with the data (e.g., Bacon's method of similarity). Calculating the VC' dimension of the concepts decidable by neural networks reveals that they are also data-efficiently learnable.

On the negative side, it can be shown that if the *VC* dimension of a concept class is d, then on some concept and in some sampling distribution, a sample size of at least $^{d}/_{\varepsilon}$ is required. Since the VC dimension of the Boolean concepts over n Boolean attributes is 2^n, exponentially large samples will sometimes be required. Thus, any algorithm that takes a sample whose size depends only on the problem and not the size of the (unknown) target concept itself will be data-inefficient (since the sample size grows non-polynomially when concept size is held fixed at the minimum value).

A *computationally efficient* PAC learner is a PAC learner whose runtime is bounded by a polynomial of the sort described in the definition of data efficiency. Since scanning a sampled instance takes time, computational efficiency implies data efficiency. Since Bacon's method is computationally trivial and requires small samples, it is a computationally efficient PAC learner. This method can be generalized to efficiently PAC learn k-CNF concepts (i.e., conjunctions of k-ary disjunctions of atomic or negated atomic sentences), for fixed k.

Sometimes computational difficulties arise entirely because it is hard for the learner to frame her conjecture in the required hypothesis language. It is known, for example, that the k-term DNF concepts (i.e., disjunctions of k purely conjunctive concepts) are not efficiently PAC learnable using k-term DNF hypotheses (when $k \geq 2$),[21] whereas they are efficiently PAC learnable using k-CNF hypotheses

For some time it was not known whether there exist efficiently solvable PAC problems that are unsolvable neither due to sample-size compexity nor due to output representation. It turns out (Kearns and Valiant 1994) that under a standard cryptographic hypothesis,[22] the Boolean concepts of length polynomial in the number of attributes have this property, as does the neural network training problem.

An alternative way to obtain more refined results in a non-probabilistic context is to permit the learner to ask questions. A *membership oracle* accepts an example from the learner and returns "in" or "out" to indicate whether it is a positive or a negative example. A *Socratic oracle* responds to an input conjecture with a counterexample, if there is one.[23] One such result is that Socratic and membership queries suffice for identification of finite state automata with certainty in polynomial time (Angluin 1987).

9 LEARNING THEORY AND EPISTEMOLOGY

To coherentists, learning theory looks like a naive form of foundationalism, in which incorrigible beliefs are the fulcrum driving inquiry to the truth. But foundationalists are also disappointed because positive learning theoretic results depend on substantial, contingent assumptions such as the nature of the signals from the environment, the structure of time, and the range of relevant possibilities. Externalists would prefer to investigate our reliability directly, instead of taking a mathematical detour into possible methods and problems. And contextualists will object to the fixity of truth through time, ignoring the possibility of meaning shifts due to conceptual change.

But on a more careful examination, learning theory reinforces recent epistemological trends. The search for incorrigible foundations for knowledge is no longer considered a serious option, so the fact that reliability depends on contingent assumptions is hardly a penetrating objection. Indeed, it can be shown by learning theoretic means that if some background knowledge is necessary for reliability, this knowledge can sometimes be reliably assessed according to the same standard, providing a learning-theoretic account of empirical regress.

Externalist epistemologies sidestep the foundational demand that the conditions for reliability be known by requiring only that we be reliable, without necessarily being aware of this fact. Knowledge attributions are then empirical hypotheses that can be studied by ordinary empirical means. But empirical science is not the same as behavioristic science. Mature empirical investigations are always focused by general mathematical constraints on what is possible. Accordingly, learning theoretic results constrain naturalistic epistemology by specifying how reliable an arbitrary system, whether computable or otherwise, could possibly be in various learning situations.

Externalism has encountered the objection (Lehrer 1990) that reliability is insufficient for knowledge if one is not justified in believing that one is reliable (e.g., someone has a thermometer implanted in her brain that suddenly begins to produce true beliefs about the local temperature). The intended point of such objections is that reliable belief-forming processes should be embedded in a coherent belief system incorporating beliefs about the agent's own situation and reliability therein. Learning theory may then be viewed as defining the crucial relation of *methodological coherence* between epistemic situations, ambitions, and means. Unlearnability arguments isolate methodological incoherence and positive arguments suggest methods, background assumptions, or compromised ambitions which, if adopted, could bring a system of beliefs into methodological coherence.

Incorporating learning theoretic structure into the concept of coherence addresses what some coherentists take to be the chief objection to their position.

> ... [Although any adequate epistemological theory must confront the task of bridging t lie gap between justification and truth, the adoption of a nonstandard conception of truth, such as a coherence theory of truth, will do no good unless that conception is independently motivated. Therefore, it seems that a coherence theory of justification has no acceptable way of establishing the essential connection with truth (Bonjour 1985, 110).

Whether a methodological principle guarantees or prevents reliable convergence to the truth is, of course, the unshakable focus of learning theoretic analysis. Where coherence is at issue, one must consider a multitude of possible interpretations of reliability and of one's epistemic situation, backing and filling until the analysis seems apt and fits with the rest of one's beliefs. This pluralistic attitude is reflected in the wide variety of success criteria, paradigms and problems considered in the learning theoretic literature.

Contextualists may also find some value in learning theoretic results. The first moral of the subject is that reliability is highly sensitive to the finest details of the data presentation, the range of possible alternatives, the kinds of hypotheses or skills at issue, the learner's cognitive powers and resources, and the methodological principles to which she is committed. Reliable methodology is unavoidably piecemeal, contextual methodology, optimized to the special features of the problem at hand.

A remaining contextualist objection is that learning theory presupposes a fixed "conceptual scheme" in which truth is a fixed target, whereas in light of conceptual revolutions, meaning and hence truth changes as the beliefs of the learner change through time. This objection does apply to the usual learning theoretic paradigms, but the concept of reliability is flexible enough to accommodate it. If truth feints as inquiry lunges, then success can be defined as a methodological *fixed point* in which the beliefs of the learner are eventually true *with respect to themselves* (Kelly 1996, Kelly and Glymour 1992). Unlike norms of justification, which may change through time, convergence to the *relative* truth provides a strategic aim that plausibly survives successive changes in the underlying scientific tradition.

Kevin Kelly
Carnegie Mellon University

NOTES

[1] 'We may talk of the *empiricist* and the *absolutist* way of believing the truth. The absolutists in this matter say that we not only can attain to knowing truth, but we can know when we have attained to knowing it; while the empiricists think that although we may attain it, we cannot infallibly know when." (James 1945): 95-96.

[2] 'Of course theories which we claim to be no more than conjectures or hypotheses need no justification (and least of all a justification by a nonexistent 'method of induction', of which nobody has ever given a sensible description)." (Popper 1982): 79.

[3] Cf. section 6 below.

[4] If there were any schedule governing the rate at which the the outputs "false" spread apart through time, this schedule could be used to produce a method that decides the hypothesis in the limit: the new rule outputs "false" until the simulated rule produces more "true"s than the schedule allows for. Thus the potential for ever rarer "false" outputs when the hypothesis is false is crucial to the extra lenience of this criterion.

[5] Conjecturing "true" while the observed frequency is in the interval and "false" otherwise does suffice unless we exclude possible data streams in which the limiting relative frequency approaches its limit from one side, for all but finitely many stages along the data stream. A reliable method is presented in (Kelly 1996).

[6] Putnam's actual argument was more complicated.

[7] Putnam concluded that a scientific method should always be equipped with an extra input slot into which hypotheses that occur to us during the course of inquiry can be inserted. But such an 'open minded" method must hope that the external hypothesis source (e.g., 'creative intuition") does not suggest any programs that go into infinite loops, since the inability to distinguish such programs from 'good" ones is what restricted the reliability of computable predictors to begin with!

[8] This construction (Case and Smith 1983) is a bit stronger than Gold's. It produces a data stream on which infinitely many outputs of the learner are wrong. Gold's construction merely forces the learner to vacillate forever (possibly among correct conjectures).

[9] Cf. the preceding footnote. In the learning theoretic literature unstable identification is called BC identification for 'behaviorally correct", whereas stable identification is called EX identification for 'explanatory". Osherson et. al. (1986) call stable identification 'intensional" and unstable identification 'extensional".

[10] I.e., the procedure halts on members of the set (indicating acceptance) and not on any other inputs.

[11] The demon presents a text for the infinite language until the learner outputs a grammar for it, then keeps repeating the preceding datum until the learner produces a grammar for the data presented so far, then starts presenting the text from where he left off last etc.

[12] A systematic compendium of results on language learnability is (Osherson et. al. 1986).

[13] The 'onto" assumption can be dropped if empirical adequacy rather than truth is the goal (Lauth 1993).

[14] I.e., the sentence has the form of a quantifier-free sentence preceded by a sequence of quantifiers.

[15] 'The computational versions of these ideas are in (Gold 1965, Putnam 1965, Kugel 1977). The topological space is introduced in (Osherson et. al. 1986) and the characterizations are developed in (Kelly 1992, 1996) Logical versions of the characterizations are developed in (Osherson and Weinstein 1991) and (Kelly and Glymour 1990).

[16] These are, in fact, the open sets of an extensively studied topological space known as the *Baire space* (Hinman 1978).

[17] Necessity of the condition fails if the hypotheses are mutually compatible or if we drop the stability requirement.

[18] i.e., hyperarithmetically definable

[19] (Osherson et. al. 1986) contains many restrictiveness results carrying a similar moral. Also, see (Osherson and Weinstein 1988).

[20] An excellent source presenting all of the results mentioned here is (Kearns and Vazirani 1994), which provides detailed descriptions and bibliographic notes for all the results mentioned below.

[21] This negative result holds only under the familiar complexity-theoretic hypothesis that $P \neq NP$.

[22] I.e. that computing discrete cube roots is intractable even for random algorithms.

[23] In the learning theoretic literature, Socratic queries are referred to as 'equivalence" queries.

REFERENCES

Angluin, D.: 1987, 'Learning Regular Sets from Queries and Counterexamples', *Information and Computation* **75**, 87-106.

Angluin, D.: 1989, 'Inductvive Inference of Formal Languages from Positive Data', *Information and Control* **49**, 117-135.

Blum, M. and L. Blum: 1975, 'Toward a Mathematical Theory of Inductive Inference', *Information and Control* **28**, 125-155.

Bonjour, L.: 1985, *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge.

Brown, R. and C. Hanlon: 1970, 'Derivational Complexity and the Order of Acquisition of Child Speech', in J. Hayes (ed.), *Cognition and the Development of Language*, Wiley, New York.

Carnap, R.: 1950, *The Logical Foundations of Probability*, University of Chicago Press, Chicago.

Case, J. and C. Smith: 1983, 'Comparison of Identification Criteria for Machine Inductive Inference', *Theoretical Computer Science* **24**, 193-220.

DeFinetti: 1990, *The Theory of Probability*, Wiley, New York.

Glymour, C.: 1980, *Theory and Evidence*, MIT Press, Cambridge.

Gold, F. M.: 1965, 'Limiting Recursion', *Journal of Symbolic Logic* **30**, 27-48.

Gold, E. M.: 1967, 'Language Identification in the Limit', *Information and Control* **10**, 447-474.

Halmos, P.: 1974, *Measure Theory*, Springer, New York.

Hinman. P.: 1978, *Recursion Theoretic Hierarchies*, Springer, New York.

James, W.: 1948, 'The Will to Believe', in A. Castell (ed.), *Essays in Pragmatism*, Collier Macmillan, New York.

Kearns, M. and L. Valiant: 1994, 'Cryptographic limitations on learning boolean formulae and finite automata', *Journal of the ACM.* **41**, 57-95.

Kearns, M. and U. Vazirani: 1994, *An Introduction to Computational Learning Theory*, MIT Press, Cambridge.

Kelly, K.: 1992, "Learning Theory and Descriptive Set Theory", *Logic and Computation* **3**, 27-45.

Kelly, K.: 1996, *The Logic of Reliable Inquiry*, Oxford University Press, New York.

Kelly, K. and C. Glymour: 1989, 'Convergence to the Truth and Nothing But the Truth', *Philosophy of Science* **56**, 185-220.

Kelly, K. and C. Glymour: 1990, 'Theory Discovery from Data with Mixed Quantifiers', *Journal of Philosophical Logic* **19**, 1-33.

Kelly. K. and C. Glymour: 1992, 'Inductive Inference from Theory-Laden Data', *Journal of Philosophical Logic* **21**, 391-444.

Kelly, K. and O. Schulte: 1995, 'The Computable Testability of Theories Making Uncomputable Predictions', *Erkenntnis* **43**, 29-66.

Kelly K. and O. Schulte: 1997, 'Church's Thesis and Hume's Problem, in M. L. Dalla Chiara et. al. (eds.), *Logic and Scientific Methods*, Kluwer, Dordrecht.

Kemeny, J.: 1953, 'The Use of Simplicity in Induction', *Philosophical Review* **62**, 391-408.

Kugel, P.: 1977, 'Induction Pure and Simple', *Information and Control* **33**, 236-336.

Lauth, B.: 1993, 'Inductive Inference in the Limit for First-Order Sentences', *Studia Logica 52*, 491-517.

Lehrer, K.: 1990, *Theory of Knowledge*, Westview, San Francisco.

Levi, I.: 1991, *The Fixation of Belief and It's Undoing*, Cambridge University Press, Cambridge.

Miller, D.: 1974, 'On Popper's Definitions of Verisimilitude', *British Journal of the Philosophy of Science* **25**, 155-188.

Mormann, T.: 1988, 'Are All False Theories Equally False?', *British Journal for the Philosophy of Science* **39**, 505-519.

Neyman, J. and E. Pearson: 1933: 'On the Problem of the Most Efficient Tests of Statistical Hypotheses', *Philosohical Transactions of the Royal Society* **231 A**, 289-337.

Osherson, D. and S. Weinstein: 1986, *Systems that Learn*, MIT Press, Cambridge.

Osherson, D. and S. Weinstein: 1987, 'Paradigms of Truth Detection', *Journal of Philosophical Logic* **18**, 1-41.

Osherson, D. and S. Weinstein: 1988, 'Mechanical Learners Pay a Price for Bayesianism', *Journal of Symbolic Logic* **56**, 661-672.

Osherson, D. and S. Weinstein: 1989, 'Identification in the Limit of First Order Structures', *Journal of Philosophical Logic* **15**, 55-81.

Osherson, D, and S. Weinstein: 1991, 'A Universal Inductive Inference Machine', *Journal of Symbolic Logic* **56**, 661-672.

Popper, K.: 1982, *Unended Quest: an Intellectual Autobiography,* Open Court, LaSalle.

Popper, K.: 1968, *The Logic of Scientific Discovery*, Harper, New York.

Putnam, H.: 1963, ''Degree of confirmation' and inductive logic', in A. Schilpp (ed.), *The Philosophy of Rudolph Carnap*, Open Court, LaSalle.

Putnam, H.: 1965, 'Trial and Error Predicates and a Solution to a Problem of Mostowski', *Journal of Symbolic Logic* **30**, 49-57.

Reichenbach, H.: 1938: *Experience and Prediction*, University of Chicago Press, Chicago.

Savage, L.: 1972: *The Foundations of Statistics,* Dover, New York.

Sextus Empiricus: 1985, *Selections from the Major Writings on Scepticism, Man and God,* P. Hallie (ed.), trans. S. Etheridge, Hackett, Indianapolis.

Shapiro, E.: 1981, 'Inductive Inference of Theories from Facts', *Report YLU* **192**, Department of Computer Science, Yale University, New Haven.

Wexler, K. and P. Culicover: 1980, *Formal Principles of Language Acquisition*, MIT Press, Cambridge.

MATTI SINTONEN AND MIKA KIIKERI

SCIENTIFIC DISCOVERY

1. INTRODUCTION

A logic or method for the discovery of new knowledge is an old epistemological dream. Knowledge of general or singular truths of course is not the only possible type of object of discovery. One can discover new things or phenomena, such as so-far undetected quasars or unconquered continents or undescribed species of micro-organisms, although it may be argued that such discoveries employ particular classificatory schemes, concepts and some particular language. Technological innovations, especially in the modern information society, occupy a half-way house. Although technology aims at designing (commercially valuable) technical devices and their systems (where systems are in seamless interaction with their human users), these innovations rely heavily on basic and applied research.

What would a logic or method of discovery look like? In the old dream it would have been a mechanical or nearly mechanical step-by-step procedure which, when given data as input, would have guaranteed a steady flow of informative truths. Ideally, it would also have been perspicuous and unambiguous in its manner of operation, easy to apply, context-independent or at least suitable for various types of knowledge from mathematics to physics and history, and effective. Perception and introspection (and possibly memory), are obvious sources of knowledge in general, and it might therefore be suggested that the perception and introspection also serve as methods of discovery – after all, although not suitable for knowledge acquisition in all areas, especially perception is reliable, almost mechanical and easy to apply. However, the discovery programme in epistemology has not been interested in the emergence of truths within the reach of unaided senses, the lower levels of the mental faculty. Rather, it has focused on the design of a method which could produce significant and systematically organized hypotheses and theories which surpass observations and introspective reports.

Many of the philosophically loaded notions in the vicinity of the notion of discovery, most notably reason, rationality, logic and method, have gone through a multitude of conceptual upheavals. It is not possible, within the bounds of a short review, to trace these changes through the past centuries and millennia. Instead, we will approach the issues of discovery and the current debates on the possibility and nature of a method of discovery through the idealized and official 20th Century story.[1] Reference is made to previous centuries through the lenses of the official story: this means that we confine to illustrate how the current agenda perceives its problems.

According to this official story the classical discovery programme was part of the scientific revolution and hence part of the project of the modern mind. It was a

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 205-253.
© 2004 *Kluwer Academic Publishers.*

reaction against the teaching of the Schools which focused on interpreting, systematizing and trasmitting knowledge already available. The motto of the programme was to read and interpret the Book of Nature and not (only) the Bible or Aristotle. It reflected the need to rely on man's own powers of mind and the ability to distinguish between certainty on one hand and hearsay and received opinion on the other hand. To keep the mind or reason on this narrow road to truth, a method, analytic, inductive or deductive, was needed. The results we saw in the methodological writings of Bacon, Descartes and Newton.

This discovery programme fell into disfavour sometime in the 19th and 20th Centuries, although the official anti-discovery programme did not harden into the current dogma until the first half of this century. The main reasons were that the proposed logics had not lived up to their promises, and that they were claimed to be epistemologically irrelevant anyway. The dogma said that there simply can be no logical way of having ideas and hence that the generation of new ideas and specific hypotheses ultimately is a matter of intuition, guesswork and luck. The dominant analytic or logically oriented philosophy of science considered discovery arational if not irrational and largely ignored the topic. It was often added, somewhat mysteriously, that learnedness and hard work help, but how they conspire to produce informed, true, truthlike or at least practically useful guesses was left open. In any case, the received view had it, the answers to these questions belonged to the psychology of creative thinking and, perhaps, history of science but not to the logic and methodology of science.

After this period of neglect the interest in scientific discovery has been in the rise again. There had always been some dissenting voices, amongst scientists and philosophers alike (see Nickles 1980a, 1980b, 1985). The main systematic discussions have been conducted in areas where cognitive development and conceptual change can be studied. These areas include the study of machine or artificial intelligence (AI) as well as cognitive psychology and cognitive science, philosophy of science, and the field of science and technology studies (STS), the more encompassing interdiciplinary enterprise which looks at science with tools obtained from history, cognitive science, psychology, and the social sciences.

Both the philosophically oriented friends of discovery and their AI allies have suggested that the real issue is not logic but something closer to heuristics and looser rational guidance. Furthermore, the historicist critics of the positivistic or logical-empiricist picture of the logic of science have been keen to emphasize the historical and social dimensions of scientific discovery. It has turned out that we need to distinguish between individual creativity and its products on one hand, and scientific discoveries on the other hand. It has been argued that the philosophical understanding of discovery and problem solving is too restricted, since the really important questions, or at least those answerable by the existing historical evidence, are not cognitive but social: How is the status of discovery attributed to a results? How are priority disputes settled and how is prestige distributed among competitors and collaborators? As many case studies suggest, there is no straighforwards answer to questions of the form: Who discovered what, where, when and how?

We shall start with a *précis* of the modern history of the discovery programme, and move on (in section 3) to the motivation for distinguishing a logic or methodology of discovery, separate from the logic of justification. Section 4 explores the rehabilitation of the discovery programme by Charles Peirce and Norwood Russell Hanson, and the replacement of the dichotomous distinction between the two 'contexts' of discovery and justification by a more fine-grained analyses in the more recent work of the 'friends of discovery' (and their allies). Section 5 addresses the question if there are methods of generating ideas or novel hypotheses, and section 6 the question if such a method would have *epistemological* relevance. What, if any, support does a hypothesis acquire by being conceived in some particular way? Section 7 discusses two more recent proposals for a logic of discovery, the adaptive logic and the interrogative view advanced by a number of writers. Section 8 is dedicated to cognitive problem solving models and the computational tradition, and the concluding section 9 to the social models of discovery.

2. THE EARLY HISTORY OF THE DISCOVERY PROGRAMME: FROM BACON TO NEWTON

The urgency of a method of discovery has not always been the same. Although novelties have always been appreciated, the focus in previous centuries or millennia was rather on systematizing and transmitting truths somehow already known than on the acquisition of new ones. The obsession of exploring completely new terrains of truths, and especially the discovery of new explanatory theories which refer to unobservable entity is a modern one.

The romantic view of scientific genius, popular in the nineteenth century, was preceded by an age of methodological optimism (Schaffer, 1990). Many philosophers and scientists (e.g., Bacon, Descartes) thought that there is a general method which enables its user to achieve more or less certain results from sensory experiences and experiments. In this picture no heroic genius is needed, for a good method and ability to use it effectively is sufficient. Larry Laudan (1980) distinguished two aspects in the 17th century discovery program. The practical problem concerned the efficient means to achieve useful practical inventions, the epistemological problem stressed the need to achieve an indubitable basis for empirical claims.

Francis Bacon's views on the new scientific method clearly focused on both aspects of the discovery programme (see Hesse 1964, Gower 1997, Urbach 1987). Although Bacon was not a professional philosopher but rather a man of action he had a profound influence on philosophical and scientific thought as well as the organization of inquiry. Criticizing the Aristotelians he wanted to see himself as a reformer who emphasized the need of a method of acquiring new knowledge and not just of representing and organizing knowledge already available. Yet, Lisa Jardine (1974) writes, although his method was not to confine to the traditional art of discourse, to him all methods were still concerned with an effective way of teaching and persuading an audience.

Bacon's theory of scientific method was part of a larger project, *The Great Instauration*, whose purpose was to collect and order all empirical knowledge by the new inductivist and experimental method. He presented the new method for finding forms that produced or constituted given natures or instances of the natural phenomena in *Novum Organum* ([1620] 1994), a book in which his philosophy of discovery is most thoroughly discussed. The method was inductivist in the eliminative sense: from lists of phenomena (natural histories) one constructs tables that show which natures are present and which ones absent when a studied nature is present or absent (i.e., the Baconian tables are an early version of Mill's methods of causal inference). This qualitative analysis is supplemented by one in which quantitative changes in a phenomenon are compared with changes in other variables. These tables, however, were only "Tables of First Presentation", and the whole procedure only yielded preliminary results, or "First Vintage". Bacon then discusses at length how the true forms can be discovered from these preliminary ones by considering especially important "prerogative instances", and how competing hypotheses could be compared by the "Instances of the Fingerpost", or crucial experiments. Although Bacon himself, unlike Robert Boyle, did not make important empirical discoveries, he illustrated the method by a lengthy examination of the form of heat.

It is easy to see that Bacon's method could not support the strong epistemological claims. Only if one could give complete lists (or histories) of natural phenomenona (and barring other errors) could the eliminative method give certain results. But there are no such lists. Moreover, as Peter Dear (1998) has argued, Bacon's methodological optimism commits its advocate to an essentialist metaphysics with a finite set of natural kinds, as well as to easy access to their complete inventory. Basically the same diagnosis applies to René Descartes' famous "rules for the direction of mind" which suggested a kind of down-to-top method: inquiry starts with simple objects and evident knowledge and proceeds, step-by-step, to more complex objects and knowledge claims. Complex problems should be divided into simpler ones whose solutions are easier, and all relevant cases should be examined. (Descartes [1637] 1968, 41) All and all, Descartes' method contains familiar heuristic elements which are widely used in the current problem solving models (see section 8 below). Dear (1998) argues that also Descartes' trust on method depended on an essentialist metaphysics which assumed close ties between all fields of natural knowledge. Reality was thought to be a simply arranged totality so that its secrets could, given the rules, be revealed by relatively small efforts.

Although attempts to formulate proper discovery methods were more systematic in the 17th century than before, they relied heavily on ancient and medieval contributions. It has been shown convincingly that Bacon, Descartes, and Isaac Newton as well as numerous 17th and 18th Century natural philosophers built on the ancient method of analysis and synthesis, or the method of "resolution and composition" (see Polya 1945, Lakatos 1976, Sczabo 1974, Hintikka and Remes 1974, Koertge 1980, Mäenpää 1993). The origin of the method is in ancient geometry where it was used as a heuristic method in the discovery of proofs. The Greek mathematician Pappus (c. 300 AD) wrote that the method of analysis is a method of mathematical discovery which starts from "the thing sought" (i.e. the geometrical figure to be constructed or the theorem to be proved) and proceeds backwards until something already known is achieved or the solution is found to be impossible. Succesfull analysis is then followed by a synthesis which purports to show that the analytically constructed proof is valid.

Although the basic idea is simple one can ask what "proceeding backwards" means in this context. Hintikka and Remes (1974) have suggested that auxiliary constructions have a key role here. They distinguish two parts in the analysis of a theorem: the given (*dedomena*, the assumptions of a theorem) and the thing sought (*zetoumenon*, the thing to be proved or constructed). The analysis proceeds deductively from the given and the thing sought, utilizing already proven results and introducing suitable auxiliary constructions (e.g., auxiliary figures in geometrical proofs). The analysis stops when the result which follows only from the given and the additional results is arrived. In synthesis auxiliary constructions are added to the proof schema and each step is shown to be convertible.

It is evident that the method of analysis and synthesis had a great impact on early modern science. Newton's unpublished writings repeatedly refer to it and his famous Query 31 in the *Opticks* makes it explicit:

> As in Mathematicks, so in Natural Philosophy, the Investigation of difficult Things by the Method of Analysis, ought ever to precede the Method of Composition. This Analysis consists in making Experiments and Observations, and in drawing general Conclusions from them by Induction, and admitting of no Objections against the Conclusions, but such as are taken from Experiments, or other certain Truths. For Hypotheses are not to be regarded in experimental Philosophy. (*Opticks*, p. 404)

But this is not the whole story. Newton developed a personal variant of the methods of the Ancients, relating them to 17th Century experimental philosophy and mathematical mechanics. His most puzzling methodological remarks, expressed in *Principia* and *Opticks*, are important because Newton claims that there indeed is a procedure which amounts to a method of discovery:

> Whatever is not deduced from the phenomena is to be called an hypothesis; and hypotheses, whether metaphysical or physical, whether of occult qualities or mechanical, have no place in experimental philosophy. In this philosophy, particular propositions are inferred from the phenomena, and afterwards rendered general by induction (*Principia*, Book III, General Scholium)

His claim thus to have "Deduced from Phaenomena" the law of universal gravitation and other general results goes against the received view that ampliative arguments could not be deductive ones. And even his concept of induction is puzzling since Newton not only claims that the results of experimental philosophy (in contrast with metaphysical and physical hypotheses) are deduced from the phenomena but also that they are "rendered general by induction".

One plausible interpretation of these puzzles is provided by Hintikka and Remes (1974, 110). They summarise the Newtonian method as follows: "(i) an analysis of a certain situation into its ingredients and factors -> (ii) an examination of the interdependencies between these factors -> (iii) a generalisation of the relationships so discovered to all similar situations -> (iv) deductive applications of these general laws to explain and predict other situations". According to this interpretation an investigation starts with an analysis of singular experimental and observational results. The analysis is then finished and the generalization deduced, and there is an attempt to generalize it by examining whether it applies to similar situations. This illustrates Newton's peculiar notion of induction[2]. For instance, he first showed that the many familiar physical interactions in Earth were instances of the inverse square law of gravitation and then generalized the law to stellar events by showing that the physical situations are similar or analogical and therefore also obey the inverse square law. Moreover, the results of experimental deductions are certain but not incorrigible generalizations, since later experiments may reveal exceptions, forcing the inquirer to develop a more refined version. There are degrees of certainty, and a generalization "may be looked upon as so much the stronger by how much the induction is more general." (*Opticks*, 404). The more a generalization could be stretched to explain apparent exceptions, the stronger and more certain it is. Finally, synthesis shows how the generalization explains all the experimental phenomena in its range.

The view that Newton's deduction amounts to a logic of discovery has also been challenged. Howard Stein (1991) has argued that Newton used the term "demonstratur" and "probare" to mean "prove", i.e., to denote purely mathematical reasoning. This is what in modern terms is usually meant by deduction. The term "deducere" (translated as "deduction") was a broader term and referred to "reasoning competent to establish a conclusion as warranted (in general, on the basis of available evidence)," i.e., to reasoning which used ordinary consequential testing, and not to deduction in our sense. On this reading, Newton's method comes close to the hypothetico-deductivist method after all.

While evidence seems to confirm that Newton was serious in his insistence that he deduced general propositions from phenomena, we can still ask if Newtonian generative methods in fact suffice for generating plausible hypotheses from empirical evidence and, if so, whether they have been applied succesfully. Dorling (1973) and Norton (1994) have argued that there are strong and commonly employed generative methods in physics. Their accounts are based on demonstrative and eliminative induction in which general results are deduced from strong theoretical premises and only a few (possibly only one) experimental results. John Worrall (2000) is more sceptical and stresses the role of background assumptions in Newtonian deductions, concluding that since it is implausible that all the background items are deduced from phenomena, the Newtonian method supplements rather than replaces the hypothetico-deductive method (cf. also Laymon 1994).

To sum up, the details of the proposed methods of discovery varied. Bacon, Descartes and Newton, the key figures in the classical discovery programme, consistently stressed the most creative part of inquiry, hypothesis generation. In retrospect we could say that the programme was based on exaggarated methodological optimism. While the state of knowledge and the general world-view in 17th century gave it some initial plausibility (though the aims of the programme were controversial even then), the development of science in the subsequent centuries lead to its abandonment.

3. THE EMERGENCE OF THE HYPOTHETICO-DEDUCTIVE VIEW

The classical discovery programme was gradually given up. We can see in the background at least three developments. First, there was the rise of the hypothetico-deductive method in the 19th Century and its canonization in the 20th Century. Secondly, there was the romantic view of science and art which emphasized the role of literally inexplicable genius in all truly creative work. And third, there was a persuasive redefinition of the proper task of philosophy of science and epistemology, aided by the 'linguistic turn' around the turn of the 20th Century: the task was to be normative justification rather than non-normative description, and this was to be done *via* an analysis of linguistically constructed inferences and rational reconstructions.

Take the hypothetico-deductive view of method first. Both Descartes' *Rules* and Bacon's *Novum Organum* were simultaneously procedures for discovering new ideas *and for justifying them* (Laudan 1980). But the Baconian and Cartesian dream, a mechanical method of discovering new truths, turned out to be impossible. Already Descartes acknowledged the difficulty of underdetermination: it is possible that one and the same set of facts is explainable by competing and incompatible assumptions.

The problem was highlighted by theories which employed unobservable entities. The atomists had already proposed minute but unobservable hooks to explain the observable properties of various types of matter, and such postulational theories multiplied in the 18th and 19th centuries. There is a method of obtaining reliable observation reports, viz. by employing ones senses. However, descriptive theories which make the inductive leap beyond singular observation claims, and especially explanatory theories which characteristically make another leap by resorting to unobservables, are in a disadvantageus position. Descriptive theories are underdetermined, and explanatory theories doubly underdetermined, by the necessarily finite observations. There are a great many (indeed, countless) rivals which, for all we know and perhaps ever will know, could systematise and explain these observations equally well (or if one theory is better than another one, superiority must be established by such pragmatic criteria as simplicity etc). And precisely because theories are risky extrapolations or conjectures, there can be no demonstrative or logical way to derive true theories.

This fallibilism was according to Laudan (1980) the first idea which contributed to hypothetico-deductive view of knowledge, as in the thought of William Whewell: empirical facts, phenomenological laws and especially explanatory theories, whatever their mode of conception (that is, intuition, analysis, synthesis, induction or whatever), could not be proved the way mathematical truths could be proved. The second idea which contributed was consequentialism according to which there is no direct way of comparing theories with nature. The best we can do is try and deduce (or induce) empirical consequences and thereby assess, albeit indirectly, their tenability. Combined with fallibilism consequentialism yields the hypothetico-deductive view of inquiry. Although Whewell was greatly influenced by Kant's views on the aim and manner of progress of scientific knowledge, he clearly formulated an empiricist view of the method which was to become the official backbone of scientific rationality, and which indeed became deeply entrenched in the scientists' self-understanding. As Whewell put it (1847, vol II, 20), "scientific discovery must ever depend upon some happy thought, of which we cannot trace the origin; some fortunate cast of the intellect, rising above rules." No maxims can be given which inevitably lead to discovery. Verification (and falsification) through observations and by help of experiments was the true mark of science.

The quotation from Whewell also shows traces of another factor which contributed to the downfall of the classical discovery programme, viz. the romantic view that art and science are allied enterprises, both with respect to their aim and the method, or rather, lack of method. On the romantic view major scientific discoveries are not systematizations of what is already known but literally conceptual novelties.

Their emergence could not be given an exhaustive account, and there could not be a mechanical step-by-step procedure from what was already known. After all, if such a method were available now, we could know now what we are supposed to learn tomorrow, which obviously is absurd. What an innovative *savant* needs is imagination which is free of ordinary mental constraints, i.e., creative genius. And it is not even clear that the logical gap could be bridged by psychology and history. For if psychological theories with their laws were sufficient to explain how creative innovations arise, they could in principle be used to anticipate these innovations before they see the daylight. Similarly, if a historian of ideas or of science could explain how the ingredients of a major innovation came together he would accomplish the impossible: he would be able to explain something which by definition is inexplicable. From these premises the alternatives seem to be luck and mysterious genius, as was emphasized by the romantic poet-scientists such as Goethe, Schelling and others. The romantic view of creative genius and the hypothetico-deductive view of the scientific method nicely complement one another.

The third reason for the downfall of the discovery programme was a new view of the task of philosophy of science. The fallibilism-cum-consequentialism view cannot of course demonstrate that there is no logic of discovery in some other than logical sense, only that it would be epistemologically secondary. Nor can it show that issues of discovery are somehow uninteresting or irrelevant and therefore not worth studying. The impossibility and irrelevant argument needed backing from the emerging division of labour between normative philosophy of science and descriptive or naturalistic doctrines. Philosophy is, as one way to draw the distinction has it, interested in reasons and grounds and normative justification, and not with the way people actually think.

The classic stand in the discovery debate is usually said to derive from Hans Reichenbach's distinction between a 'context of discovery' which comprises actual psychological (and sociological) facts on the way to discovery, and a 'context of justification' (See Reichenbach 1938, and Gutting 1980 for discussion). But what was the distinction? In a survey of the discovery debate Nickles (1980) singles out several claims which have dominated the debate and helped to establish the 'standard' distinction. First, there is the distinction between the scientist's actual psychological processes and the logically tidied and edited argument which display the evidential credentials of the 'finished report' to the scientific community. Logic and philosophy are in their nature normative disciplines which aim at rational reconstructions and not at descriptions of actual descovery processes. Secondly, there is a temporal distinction, for an hypothesis must be discovered or invented before it can be tested. Third, there is the view that logic must be denied recognition in the context of discovery either because all logical considerations by definition are justificatory or because there is no algorithmic procedure for scientific problem solving. Finally, a deep discovery can be said to be illogical or non-discursive because it characteristically involves conceptual innovations and a holistic transition to "a new way of seeing things".

A closer reading of Reichenbach shows that only the first one of these ideas can be attributed to him (see Gutting 1980). Two important developments can be seen in the background of the distinction. First, Frege (and Husserl) had completed a devastating critique of psychologism in logic, and established that logic does not deal with actual thinking. Secondly, in part due to Frege and Wittgenstein, philosophy had gone through the linguistic turn according to which the perspicuous way to formulate problems and their solutions was within some well-defined language. Reichenbach was a leading proponent of the scientific empiricists in Berlin, and like the Viennese positivists he was dedicated to purging all subjective and psychological elements from the foundations of science. On Reichenbach's view the validity of a scientific claim did not depend on who had proposed it but rather on the logical relation of a theory to facts (see Giere 1996 for discussion). And indeed, Reichenbach's idea of giving a rational reconstruction of scientific inquiry owed its main character and main tools to the logical positivists and especially Carnap. The general idea was to give these notions purely syntactic and semantic explicates, to make them non-pragmatic relationships between sentences and hence to avoid mention of time, persons, background knowledge and contexts in general. Since discovery episodes, unlike justificatory arguments, cannot be presented as logically neat transitions from available premisses to conclusions, they were epistemologically unimportant or irrelevant.

The outcome was the received view of discovery according to which there is a logical or category difference between discovery and justification, since one deals with sequences of ideas and the other with propositions. Moreover, there is a associated temporal difference because hypotheses must first be generated before they can be tested. And since all empirical knowledge is fallible the rationality of science hinges on the way it is justified. A classic source in which many of these claims can be found is Sir Karl Popper who perhaps more than anyone else contributed to its propagation. Popper (1959, 31) put the sorry state of the discovery programme succinctly:

> The initial stage, the act of conceiving or inventing a theory, seems to me neither to call for logical analysis nor to be susceptible of it. The question how it happens that a new idea occurs to a man . . . may be of great interest to empirical psychology; but it is irrelevant to the logical analysis of scientific knowledge. This latter is concerned not with questions of fact ..., but only with questions of justification or validity.

And Popper then proceeded to distinguish sharply between "the process of conceiving a new idea, and the methods and results of examining it logically". This, likewise, develops Reichenbach's ideas, but there is also another idea in Popper which puts a discovery idea into jeopardy, viz. his fallibilism: it might be possible to design heuristic rules for inventing ideas, but there can be no demonstrative or logical way to derive theories, simply because all theories are conjectural and open to refutation. Popper's fallibilism thus provides one source for those who doubt the existence of a logic of discovery.

And when we come to Carl Hempel's (1966) classic critique of 'narrow inductivism' as a description of the process of inquiry we finally meet the canonical work which established the impossibility of a logic of discovery. Induction, Hempel wrote, can refer to the evidential relationship between a hypothesis and evidence, but it has no say in the description of the actual path into that discovery. Not only is the path to established results beyond any conceivable inductive principles: the process of inquiry cannot even start without specific hypotheses so that the first step in the narrow inductivist's discovery programme is impossible. This is a refinement of Reichenbach's distinction and already contains a basic challenge to a discovery program: consequentialism and the ideal that logic of science must confine to the logic of hypothetico-deductive testing.

4. THE REHABILITATION OF THE DISCOVERY PROGRAMME: ABDUCTION AND THE CONTEXT OF PURSUIT

One of the turning points in the philosophical re-evaluation of the discovery programme occurred in 1978 when the so-called "friends of discovery" and allies took stock in Reno, Nevada (see Nickles 1980, where the label "friends of discovery" is also attributed to Gary Gutting). The simplistic dichotomy between discovery and justification was rejected, and it was acknowledged that discovery can either refer narrowly to original generation or broadly to the process which starts with generation (or even before) and ends up with final acceptance. As Gutting (1980a, 222) put it, if we do not confine to finished products, "discovery (properly speaking) is identical with the very process of scientific inquiry". As a result a viable discovery model should address both the question how ideas are generated and what happens to them later in the process. The wider sense leads to a more fine-grained account than Reichenbach's, both logically and temporally.

As a result of this development many "friends of discovery" replaced the two-stage view by a tripartite account consisting of generation, pursuit, and justification stages, and argued that there is an interesting middle ground between discovery and justification (see Laudan 1977, Kordig 1978, Curd 1980, McLaughlin 1982a). If discovery is defined narrowly as an original psychological act of conceiving an idea, there seems to be room for the pursuit of the idea before it is tested. Indeed, Carl Kordig (1978) elaborated the mainstream thought by suggesting that of the three contexts plausibility assessment and testing and acceptance fall within the area of logic and rational method, while initial generation belongs to psychology. This also follows from the observation that before the hypothesis is consider to be worthy of often very costly testing, there has to be some amount of evidence which supports it. This initial justification makes hypothesis believable and testworthy, and the relevant evidence on its behalf is gathered during the pursuit stage. Now the problem was: What can we say about this context of pursuit? The main impact of the friends of discovery has been the clarification of this context or element of scientific activity.

We should note, however, that this development started earlier in the century. That something constructive could be said, specifically, about discovery surfaced anew into the consciousness of the philosophical community in the late fifties when Norwood Russell Hanson revived Charles Sanders Peirce's retroductive or abductive inference. He also strongly criticized the popular hypothetico-deductive method (Hanson 1958), and argued that if we focus on hypothesis testing only we miss many of the most interesting aspects of scientific research. We also need an account of context of discovery which shows that initial generation is rational activity and adheres to strict arguments. To this effect, Hanson adopted the Peircean notion of retroductive or abductive inference and claimed that this schema forms a kind of logic of discovery. Abduction, Peirce and Hanson claimed, differs from deductive and inductive inference. It moves from an explanation-requiring observation to an explanation-giving hypothesis along the following pattern (Peirce 1931-58, Vol. 5, 189):

> The surprising fact, *C*, is observed
> But if *A* were true, *C* would be a matter of course
> Hence, there is reason to suspect that *A* is true.

It soon became evident that what Hanson actually proposed was, at best, a logic for pursuing competing hypotheses. The standard complaint was that the abductive schema cannot account for the initial generation since the hypothesis-to-be-discovered *A* was already mentioned in the premises. And as Peter Achinstein (1970, 1980) remarked, the inference from the succesful explanation to the truth of the hypothesis is not valid, if even reasonable. There are always numerous wild hypotheses which, if true, would explain observations. Hence, the natural way to interpret the situation is that this pattern of reasoning focuses on reasons for entertaining or suspecting an hypothesis instead on reasons for accepting it as true or adequate.

Hanson later elaborated on the distinction by allowing for the possibility that abductive inference is triggered by more than one fact and by adding an important note on the identity or type of the hypothesis: reasons for suggesting a hypothesis are, rather, reasons for adopting one kind of hypothesis, whereas reasons for accepting a hypothesis are reasons which favour some particular well-circumscribed "minutely specified hypothesis *H*" (Hanson 1961, 22). The result is the following pattern:

> 1. Some surprising, astonishing phenomena p_1, p_2, p_3 ... are encountered.
> 2. But p_1, p_2, p_3 ... would not be surprising were a hypothesis of *h*'s type to obtain. They would follow as a matter of course from something like *h* and would be explained by it.
> 3.Therefore there is good reason for elaborating a hypothesis of the type of *h*; for proposing it as a possible hypothesis from whose assumption p_1, p_2, p_3 .... might be explained.

The reasons for suggesting a hypothesis belong to plausibility considerations, and they operate prior to reasons for acceptance. They are characteristically based on similarities or analogies between known systems and unknown or partially unknown systems. Although analogies are not sufficient to justify hypothesis as true they do suffice to make them plausible or at least worthy of further inquiry.

As we have seen, Peirce's retroductive inference to explanation presupposes that the explanatory hypothesis is available (as well that, in what became to known as inference to best explanation or IBE, the rival candidates are available), and he later submitted that abduction does not deliver what is needed – an orinatory logic or logic of generation (see Tursman 1987, 18). Similarly, Hanson acknowledged that analogies can only justify a kind of hypothesis as reasonable. What an analogy or abduction cannot do is generate a specific explanatory (or any other) hypothesis, nor guarantee the truth (or truthlikeness) of either the generic or specific hypothesis. Hanson also charaterizes the discovery process in terms of conceptual *Gestalt* switches in analogy with that of perceiving patterns. But as critics have proposed, such sudden perceptual reorganizations are the opposites of what we usually refer to by reasoning (for a summary of the discussion, see Nickles 1980).

In this light Hanson's "logic of scientific discovery" is best interpreted as a logic of pursuing hypotheses, not as the logic of their initial generation (which was Hanson's original intent). This comes close to the later tripartite analysis of research process. How could we then analyse this new middle stage, the context of pursuit? Although opinions vary as to its nature, there seem to be (at least) two types of activities, both of which can be analysed further still. First, once an initial idea is generated it is subject to a preliminary assessment, and secondly, plausible or promising ideas are developed into more detailed hypotheses.

Preliminary assessment seems to involve several types of considerations. Some ideas – the non-starters – are ruled out at the outset as unlikely or impossible. This could happen through conscious reasoning or on the level of unconscious processing. Theories or their parts serve in the dual role of a positive and negative heuristics (as in Lakatos's (1970) methodology of scientific research programmes), weeding out implausible ideas and bringing forth plausible ones. Here the contexts of discovery and justification seem impossible to separate: a theoretical background may give rise, e.g. through a similarity or an analogy, to an idea which nevertheless is rejected because it goes against the fundamental ontological assumptions of a field or because it contradicts accepted facts within the field or in the neigbouring field. Such preliminary assessment could in principle be presented in the form of explicit arguments but often they are more like routine "expert judgments".

Preliminary assessment can also build on Peirce's and Hanson's two types of reasons, i.e. reasons for suggesting a hypothesis in the first place and reasons for accepting it. Now for acceptance in the context of final assessment there are a number of tools, such as those canvassed within the Bayesian strategy, but they presuppose that the hypothesis and the evidence are clearly articulated. Salmon (1966, 118) suggested that Hanson's plausibility arguments could be explicated by help of the prior probabilities of the hypotheses: "They are logically prior to the confirmatory data emerging from the hypothetico-deductive schema, and they involve direct consideration of whether the hypothesis is of a type likely to be successful". At the same time, plausibility arguments help to settle one of the main problems of Bayesian model of scientific inference, namely, the determination of prior probabilities. Salmon goes on to expand on the features which bear on such plausibility assessment, such as the relations with already established theories and pragmatic considerations of simplicity. There seems to be a consensus that similarities, analogies and models as well as pragmatic and even aesthetic factors function both on the level of generation and pursuit, i.e., preliminary assessment and theory development (cf. section 5).

One example of this line of argument is Robert McLaughlin's (1982a,b) account of plausibility arguments. A new hypothesis *h* is subject to plausibility assesment which normally precedes testing, and which, on the other hand, is intimately linked with invention. The outcome of plausibility assesment can back the decision to pursue further (and eventually to test) *h*, or it could help to assign a prior probability for *h* in the Bayesian scheme, as Salmon suggested. Depending on the outcomes, there are two (not necessarily disjoint) types of plausibility arguments.

Advancement arguments function in the context of invention, enhancement arguments in the context of appraisal. They can be based on enumerative induction or some more sophisticated principle of inference such as analogy, symmetry or some other form of simplicity. An example of analogical advancement argument concerning the structure of DNA molecule is the following (McLaughlin 1982b, 88):

(Q1) The structure of DNA molecule is unknown, but one of its major chemical constituents is a form of nucleic acid. (Background information)
(Q2) In chemical composition, DNA is analogous to TMV (tobacco mosaic virus), which also has a form of nucleic acid as a major chemical constituent. (Analogy claim)
(Q3) The TMV molecule is helical in structure. (Datum)

(*h*) The DNA molecule is helical in structure.

By this kind of arguments, McLaughlin concludes, advancement arguments can give inductive support for *h*, i.e., plausibility is a measure of initial inductive support for *h*.

The second type of activity in the context of pursuit concerns the development of the idea into a full-blown hypothesis. To the extent scientific achievements are planned, or can be prepared for, this stage must overlap theory construction and goal-directed problem-solving. An example is given by Robert Monk (1977, 1980) who has studied what he calls productive reasoning which draws on written records (scientists's publications etc), interviews and the monitoring of research in progress to articulate how both routine and innovative research is carried out. Crucial for his account is the classification of research problems into (perhaps overlapping) problems of explanation, reconciliation and determination. He takes the largely ignored category of determination problems, where the task is to find a determined value of a determinable variable. Determination problems range from answers to well-defined *wh*-questions ("What is the speed of light in vacuo?") to complex questions in which the "restraints" on the answers do not confine to a specific function, structure or species ("What is the structure of the insulin molecule?"). Productive reasoning then proceeds through mental and physical actions falling into the categories of conception, planning, execution and monitoring. The conception of a project in turn contains five ingredients, the setting of the context of the problem, its articulation, the restraints on its solutions, the approach or general method to be used, and "seminal ideas". Planning in turn concerns to the overall strategy of devising a series of subtasks, and execution refers to the carrying out of the experiments and data analyses, to observations and calculations to be made, and to the hypotheses and conclusions to be assessed. Scientific reasoning, then, proceeds in a highly complex manner in which an idea is developed to meet increasingly more specific restraints into articulate hypotheses.

Similarly, Laurie Ann Whitt (1990) suggests, using John Dalton's *A New System of Chemical Philosophy* as her example, that a philosophy of science which focuses entirely on theory acceptance fails to address the way in which promising theories are developed. She submits that whereas the logic of acceptance relies heavily on epistemological criteria, the logic of pursuit contains pragmatic commitments to a line of development. Like Hattiangadi (1978) and Kitcher (1993) Whitt notes that the best overall strategy in a scientific community is not to pursue just one single line. Peter Achinstein (1993) writes, in the same spirit, that we can defend a theory without testing it. Developing a theory by refining its assumptions, or by deducing testable consequences, may aim at the assessment of its theoretical credentials without commitment to its truth or truthlikeness. It is important to know if it is reasonable to invest in the pursuit. Achinstein suggests that such reasoning is weaker than, and takes place prior to, reasoning which attempts to justify a theory as true or truthlike. Nevertheless, it has its "logic" which may also resort to such methodological criteria as simplicity and testability. Achinstein's best-developed example was the way Niels Bohr defended his "speculative" atom model by motivating why such a theory was needed and by pointing out open questions which it could, parhaps, answer. Apart from motivational arguments he tried to deduce more precise answers to these questions as well as to defend more specific assumptions.

5. THE LOGIC OF GENERATION

Although friends of discovery have shown that at least the pursuit stage contains elements which could be subjected to rational analysis, the central question of traditional discovery problem, namely the possibility of the method of generation, is still left open. The proper target of the arguments against the impossibility or epistemological irrelevance of a logic or rationality of discovery is the method of generation. Is it true, as Popper (1968, 31) said (and Feynman echoed), that "the initial stage, the act of conceiving or inventing a theory" neither needs nor admits logical analysis? Are theories and hypotheses guesses of "free creations of the mind"? Can there be any reasoning into the hypotheses which serve as input to the pursuit stage? Is there reasoning in generation contexts? What would the logic of generation be, and what the notion of rationality needed? Are these guesses arational or irrational?

Let us start with guessing. Joseph Agassi (1980) points out that in the complaints that there can be no method of discovery guessing is, without argument, contrasted with rationality. But this is an unwarranted assumption. The best method in solving diffential equations is guessing, indeed students are standardly urged to use the *method* of guessing! Furthermore, although there is an algorithm for carrying out divisions school children are taught to use guessing in problem solving. Why, asks Agassi (1980, 186)? Because guessing is not algorithmic and "so is irrational". But a moment's thought should convince us that exactly the reverse is the case: guessing is routinely used in science and mathematics, and indeed it is difficult to see how it could be eliminated without stifling the process of inquiry (or learning).

As to the philosophically loaded term logic, the classical discovery programme was hardly committed to logic in our sense. Modern mathematical logic is often defined as a study of formal languages or valid inferences, where validity is given a special interpretation. However, logic in earlier centuries referred to the science and art of correct reasoning, and it involved a set of considerations wider than mere logical validity in the modern sense. There is no particular reason to deny that there is a logic of discovery in the sense of logic of inquiry. But is there a logic of generation?

Two obvious candidates for a logic of generation are deductive and inductive logic. But as Peter Achinstein (1980, 120) notes, the methods and principles of good or correct deductive reasoning have not been of much use to scientists. Similarly, Carnap's inductive logic with its rules for assessing the rational support evidence gives to an hypothesis has gained little currency. Achinstein ascribes their weaknesses to excessive generosity: focusing on deductive entailment does not tell the inquiry which inferences are useful for the problem at hand, and the apriori rules of inductive logic cannot tell what the total evidence is or on what level a theory is explanatory. In short, these logics are open to the complaint that Charles Lyell voiced with respect to Darwin's theory of evolution through natural selection. He said he could understand how natural selection acted in the role of two of the members of "hindu trinity", Visnu the sustainer and Siva the destroyer. But he did not understand how it could serve as Brahma the creator – and without its creative power "we cannot conceive the others having any function" ... "nothing new w.^{d} appear if there were not the creative force". (Quoted from Wilson 1970, p. 369). These shortcomings have also inspired, in part, views according to which rationality does not boil down to obeyng the rules of inductive or deductive logic (see e.g. Cherniak 1986 and the entry by Samuels, Stich and Faucher in this volume).

The third possibility, abductive logic, turned out to have similar drawbacks: it might be useful in the context of pursuit but it provides no way of generating explanatory (or any other) hypotheses. Difficulties such as these have shifted the focus from logic in the narrow modern sense to rationality: although it is doubtful if there is any particular logic of generation we may ask if generation nevertheless is rational in some sense. But what does it mean to say that a discovery or an inference to a belief or an initial scientific hypothesis is rational? And if discovery is rational, what is the role of luck and serendipity? Can one intend to generate a new hypothesis?

Scientists themselves often report and sometimes record the way they reasoned into their beliefs and theories.[3] For instance it took Darwin some twenty years to work his way to the theory of evolution through natural selection, and his notebooks and diaries contain detailed comments on the genesis of his ideas. Here one could object that Darwin's reasoning into the 'final' form of his theory belonged to the context of pursuit and not generation, for the hypothesis of natural selection was formulated already in 1837 – long before the full theory. And could it not be objected that the activity of reasoning must, on pain of infinite regress, end in perceptual takings which can no longer be called reasoning and which therefore are beyond rationality, or arational?

Here we could distinguish between several senses of rationality, but the following should suffice here. First, coming to entertain a belief can be rational even if the episode is not a result of judgment or inference at all. Perceptual beliefs are cases in point: I see a car parked on the other side of the street and come to believe that there is a car parked there. Such beliefs count as rational because they result from a reliable perceptual process. There may of course be a naturalistic Mother Nature explanation for the emergence, spreading and survival of such a belief-forming apparatus which refers to its evolutionary advantages for our ancestors. (See Ruse 1985, Ellis 1988, Lycan 1985, Sober 1981).

Secondly, there are actual cases of generation in which both reasoning based on perceptual intake and clearcut deliberation turn out to have been crucial (see Achinstein 1970, 1971, 1980). Galileo peered into his telescope and, on the basis of what he saw, inferred that Mars was not the even-surfaced heavenly body it was claimed to be. An element of inference was incontestable because deeply entreched background knowledge and methodological assumptions made the conclusion initially impossible or unlikely. Achinstein suggests that inference into a hypothesis in a context of generation means that the inquirer comes to believe in the hypothesis for evidential reasons. When these reasons indeed support the hypothesis, in fact and not just in the light of the possibly idiosyncratic opinions of the inquirer, they could be called rational (Achinstein 1980, 118). Again there may be a Mother Nature story to be told about how we come to have such a capacity of inference.[4]

Given that hypothesis generation is rational in one of the senses outlined, what is the role of serendipidity, discovering a result whilst looking for solutions to something else, and luck? Isn't the picture canvassed too rationalist? Several writers have tried to reconcile rationality with serendipidity, e.g., by arguing that goal-directed problem solving usually requires unforeseen developments to succeed. We can here only take two specific proposals for a closer look, Howard Gruber's evolving systems approach and the defence of blind variation and selective retention (BV & SR) by the evolutionary epistemologists and, more recently, Thomas Nickles. Gruber (1980) describes scientific endeavours as goal-directed activities in which the agenda consists of three subsystems, of organization of knowledge, of purpose, and of affect. This evolving systems approach aims at showing that scientific thinking is a series of structural transformations, at all stages of inquiry. It paves the way beyond the simple Aha! to the description and explanation of how creative processes in scientific thought actually proceed and how the agenda of an inquirer grows as a result of transformations. One of his examples is the development of Darwin's thought from the vague idea in 1837 to an articulate presentation of the theory of natural selection over twenty years later. Gruber describes how Darwin started with natural theology and catastrophism but, with novel facts flowing in, came to embrace the uniformitarian view that nature is subject to gradual transformations. Adopting the new background philosophy answered some questions but created new ones: "If organisms are perfectly adapted to the milieu for which they were created, what becomes of this adaptation when the milieu changes?" Such questions create disturbances and disturbances create new questions.

In Gruber's account questions arise from such transformations as well as from perceived similarities and analogies. He warns of hasty conclusions about the impact of analogies, such as Malthus's ideas in Essay on Population on Darwin's thinking, and suggests that they lay dormant until attention was drawn to them again. The Aha! was perhaps triggered by a rereading of Malthus, but it required that the entire structure of Darwin's thought had obtained a shape which made ready for its reception. The role of metaphors and especially analogies in plausibility assessment has also been emphasized by Rom Harré (1960, 1970) and Mary Hesse (1966, 1974): a similarity between a known and an unknown system may suggest neutral analogies, open questions to be explored. In Harré's (1960) account a theory in one science can be used to describe the facts of another, so that one can speak 'more' and less fundamental theories. This proposal – that analogies can be detected between the hierarchically arranged theories of different sciences – has later been generalized within the so-called structuralist theory of science.

The role of models and analogies in concept and theory development is now an important topic in philosophy of science. Thus e.g. Nancy Nercessian (1984, 1987) has drawn attention to their importance in giving concepts meaning, and in making sense of imagery and pictoral representation in the research of, e.g., Faraday, Maxwell and Einstein (for concept formation, see also section 8). Creation through structural transformations does not amount to a logic of generation. Although goal-directed, the outcomes are in part up to luck and contextual contingencies. But isn't Gruber's account still excessively rationalist? Can one plan to make discoveries? One response to this question adapts the Darwinian reply to Lyell into method. Lyell said he did not understand how selection could function in the role of Brahma the creator. The Darwinian response was to show that new forms could arise and in fact have arisen through undirected or blind variation and step-by-step selection of the variants. Gruber's account, as well as Scott Kleiner's (1993, 1995) somewhat similar analyses, are fully compatible with blind variation and selective retention: variations (new ideas) may not be completely blind, for the inquirer is an active agent on the agenda and the choices are in the logical space of reasons, although the end result of the series of transformations cannot be foreseen in detail.[5]

Popper's denial of the logic of discovery was spirited with the view that there is no logical way of having ideas (they thus arise through blind variation) but they are subject to selective pressure. Campbell (1974a, b, See also Hull 1988) turned this into a cornerstone of his evolutionary epistemology, suggesting that blind variation is the key to the discovery process. Inquiry is goal-directed activity in that scientists do their best to heed the established facts and theories concerning a problem. However, these constraints never suffice to deliver a unique theoretical innovation or solution, so that ultimately luck and serendipidity are unavoidable. Thus, the psychologist D.K. Simonton (1988) argues that the process of brooding over a problem results in a largely random reordering of mental elements in the mind of the discoverer.

Aharon Kantorovich (1993, 1994) in turn suggests that serendipity which results from largely unconscious or at least from uncontrolled incubation of an idea is formative of major theoretical innovations. Just as diversity and phenotypic variation is the most pervasive feature of biological nature, so blind variation drives the process of discovery. Kantorovich also adds to this picture something he calls tinkering: epistemic cooperation between scientists working at the same time, and (characteristically inadvertent) borrowing of theories and ideas. As a result actual discovery processes, from the vague initial idea to a recognized result, is messier than is suggested by scientific hero stories (see also section 9 below). The way hypotheses are generated is explainable (though they often are unexplained), but not through the great minds' superior powers to see and aim at an end, and to work around the obstacles, but by help of largely contingent mental and social processes. This does not mean that scientists have no active voice in the discovery process, but it does mean that in many cases cultivating discoveries amounts, as in Gruber's account, to preparing the mind to form associations, e.g. by immersion in background knowledge and by playing with possibilities.[6]

Thomas Nickles (1997) has argued that Campbell and Kantorovich were essentially right in emphasising blind variation. Design models of knowledge, like creationist counterparts in biology, presuppose that only (more) order can beget order: one can acquire knowledge only through someone who knows it, or by help of a method. However, Darwin shows that design and miracle are not the only possible explanations for novelties: chance variation and selective pressure working through thousands of generations can result new forms. Similarly, ideas and hypotheses, even complex ones, can arise through incremental changes and adaptation to local "habitats". To the objection that science is not blind (because scientists are not blind) but committed to, and kept on the road to, truth by method Nickles's short reply is: luck and method are, contrary to what Popper and Campbell thought, allies rather than enemies.

The crucial observation is that BV & SR can "work towards" a solution to a big problem without having the solution within sight (or even behind the closest obstacles). The BV& SR process can be applied in a methodical way, as contemporary advances in computer science testifies. Computers are often said to be uncreative because they cannot solve problems unless the constraints and the algorithmic procedures for solving them are programmed in them. However, results from evolutionary computing, and from genetic algorithms in particular, show that this variant of Lyell's complaint is unfounded. Computers programmed by John Holland (1995) and John Koza (1992) to implement a selection process started with a gamut of programmes for executing simple subroutines ("ideas"), proceeded to generate, randomly, the next generation of programmes which, then, were allowed to breed with programmes of the initial population. In a surprisingly few generations the algorithm was able to produce solutions to a variety of types of problem. The breeding process was blind, but the results were as if designed by a highly intelligent programmer. The entire process of course is not biological – which only goes towards showing that natural selection as a process can be instantiated in non-living stuff.

Two concluding remarks should still be made. First, once the dichotomy to the contexts of discovery and justification was given up, it was possible to take a closer look at what happens *before* generation (and *after* empirical testing Richard Burian (1980, 322) suggested that what is needed is a classification of the factors which are likely to enhance discovery. Although there may be no logic of generation that could guarantee success we can throw partial light on how background knowledge and hard scientific labour might bring results. On their "preparation stage" one tries to isolate a body of reliable information concerning a domain of phenomena, to serve as a preliminary which gives guidance in "the revision of knowledge in the light of new information". It is not clear, even, if the distinction between the "contexts" of preparation and generation can be drawn at all. Secondly, luck and hard work are often contrasted with each other. However, the various types of elements in the background knowledge – the methodological, ontological, and more theory-specific assumptions – needed in the generation of hypotheses are results of hard labour. Indeed, both Michael Polanyi (1958) and Herbert Simon (1977) point to the common wisdom that luck and serendipidity frequent the prepared mind.[7]

6. DISCOVERABILITY

There is an important question yet to be discussed: although there may be a method of generation, it is not clear that it has epistemological relevance. The relevance was taken for granted in the heyday of the scientific revolution and after because the two logics were not opposed to one another: a (or the) method of discovery was simply taken to be the most important (perhaps the only) guarantee of justification. The strongest support would be obtained if a theory or law could be deduced from the phenomena, as Newton claimed to have done. On the received view, however this was a serious confusion, for scientific reasoning always proceeds from predictions deduced from theories (and auxiliary assumptions and initial conditions) to their confirmation by logically weaker evidence. In the current debate Larry Laudan (1980, 1983) especially has challenged the friends of discovery to tell us what the method of discovery adds to tests through consequences: why should the interest in the method be revived? One answer would be to insist that the way ideas are conceived do bear on justification, either because justification cannot be separated from discovery or because discovery methods *can* add something to consequential testing.

Despite the seeming uncontestability of consequentialism ideas reminiscent of deduction-from-the-phenomena have been rehabilitated in recent years. There are several explications of deduction from the phenomena and of weaker forms of generative justification in the philosophical litterature, although they are not always explicitly formulated in these terms. Each account offers a particular interpretation of what such deduction might mean, e.g., demonstrative induction (Dorling 1973), eliminative induction (Norton 1995), or bootstrapping as in Glymour's (1980) theory of confirmation. We shall take a closer look at two particular attempts to rehabilitate something that could be called generative justification, one by Thomas Nickles, the other one by Kevin Kelly. Nickles (1981) starts with the constraint-inclusion model of a problem in which problems are defined by known empirical facts and the accepted theoretical constraints. The latter often go unnoticed but they are needed because they narrow down the set of solutions which square with accepted facts. The central epistemic intuition behind generativism is that theories must somehow be deduced, educed or otherwise derived from what scientists already knows or which is indubitable or uncontroversial enough so that the theory inherits the justification of the evidence. The ideal limit is theoretical proof from accepted constraints (Nickles 1987, and 1985).

But if the aim is justified discovery, an achievement and not just any wild guess, Nickles needs something more. What more is needed to respond to Laudan's challenge? What is the surplus value of generation? Nickles's (1985) first step towards a respond is to insist that meeting prevously set constraints does contain an element of justification. The second step is to distinguish between between actual discovery and potential discovery or discoverability. These steps are cashed out as follows. In generative justification one reasons retrospectively from already justified data and results to a conclusion, and to the extent the premises are warranted the conclusion is likewise warranted – "prior to any independent test of it." Whether logical derivation or something closer to inductive support, the degree of justification of the conclusion also depends on the support the premisses enjoyed to begin with. And to counter the objection that there are multiple paths to a conclusion Nickles notes: "Generative justification usually requires setting out what amounts to a rationally reconstructed discovery path ... it is not discovery in the sense of the initial conception of an idea". (Nickles 1985) The result is a *rational reconstruction* of a potential discovery or discoverability, analogous to explanation as potential predictability. In this characterization of generative justification a hypothesis inherits its warrant from antecedently accepted background knowledge, although the types of arguments used may vary locally. As rational reconstructions discoverability arguments are *post hoc* construals of how the actual discovery *could* have happened, and how the theories and evidence *should* have been arranged and employed in the arguments in order to give good reasons for the initial acceptance of the results. Sometimes the actual discovery episode may coincide with this rational reconstruction, but in general it need not.

There is also a more general answer to Laudan's challenge which involves a reassessment of the central issues in the discovery debate, and the role of novel evidence in theory acceptance. To discern the surplus value of generative paths Nickles (1985) distinguishes between two theses. The *per se* thesis claims that methods of generation, *as methods of generation*, have special "probative weight", while the divorce thesis states that generation is logically distinct from justification (the denial of this is the anti-divorce thesis). The received view does not claim that nothing of interest can be said about generation, but it is committed to the divorce thesis and to the denial of the *per se* thesis: the original invention of a theory or hypothesis has no relevance to the support it gains from empirical tests. Laudan's challenge therefore is in line with the received view: there is no epistemological reason to get excited over generation (or pursuit).

Nickles sides with neither the friends nor the foes of discovery – on grounds that taking these theses as central issues in the discovery controversy would be misleading. Those interested in defending the relevance of discovery should welcome a form of divorce, he writes. Furthermore, the *per se* thesis is not vital to their cause. To make the case Nickles critisizes Robert McLaughlin's (1982a,b) and Marcello Pera's (1981) arguments against the divorce thesis which tried to establish that there is a close connection between discovery and justification, and this connection makes discovery epistemically relevant to justification.

As McLaughlin (1982a,b) puts it, Laudan's and Herbert Simon's problem-solving views entail that "the 'logics' or inference procedures involved in discovery and in justification are different and independent" (1982a, p.198). The underlying rationale for this rendering of the divorce thesis is, according to McLaughlin, an inadequate conception of inductive inference. Especially Laudan ignores the role of plausibility arguments "which provide a crucial link between invention and appraisal, and thence an epistemic rationale for inventionism" (ibid.). This gives justification to the denial of the divorce thesis, the anti-divorce thesis. Pera (1981) in turn argued that the method of discovery consists of inductive inferences from the data. Although we could not always state exactly how such inferences proceed, we can be sure of two things: first, that such inferences from the data in fact occur and, secondly, that these inferences are ampliative and non-deductive in kind. Ampliative inferences could be called inductive even if their exact character is not always clear. Pera also argued that there is no principled distinction between the discovery of empirical generalizations and of theoretical hypotheses. Both involve ampliative inferences and require novel concepts and colligation of facts as William Whewell put it.

Pera further argues that inductive methods of discovery and the consequential testing of hypotheses require each other. In the Bayesian scheme of hypothesis testing, for instance, the positive posterior probability of a hypothesis *h* demands that its prior probability is non-zero, and prior probabilities are determined in the process of discovery. And Pera maintains that Salmon's (1966) suggestion that prior probabilities are adjusted in the context of pursuit in the form of plausibility evaluations will not do either. Even the decision to pursue *h* presupposes that the inquirer believes *h* and has some reasons for her belief. According to Pera these epistemic reasons come from the background knowledge which backs up the inductive methods of generation. So Salmon's suggestion that the generation of hypotheses belongs to the psychology of discovery is not acceptable. Pera concludes that inductive methods of discovery are epistemically relevant, since the epistemic constraints by which we determine the values of prior probabilities arise from the context of generation, not from the context of pursuit. However, Nickles (1985) argues, these responses do not silence Laudan's objection. McLaughlin's plausibility assessments and advancement arguments contribute to the evaluation of their heuristic efficiency. But as Laudan (1983) pointed out in his reply they do not show that discovery methods as such have any special *epistemic* force. The question, then, is if such special epistemic force is needed at all? Is it not enough that discovery methods enhance the efficiency of scientific inquiry?

Part of the answer to these questions is provided by the notion of discoverability. Since the accounts of the discovery of an hypothesis *h* can offer arguments and evidence which differ from the ones used in the tests of *h*, they can, in this way, provide extra justification for it. But, more importantly, the demand for "special epistemic force" becomes simply irrelevant to the discovery programme. Although discovery accounts *per se* contain no special epistemic properties, they need none, since they already have a special contribution to offer. Their contribution is related to heuristic problem solving methods, and to the economy of research. The methods of science have a legitimate and normative concern over finding and designing efficient knowledge strategies. Consequently, although discovery methods are not in any way "special" they nevertheless bear on scientific rationality and therefore can have epistemic force.

This construal of the *per se* thesis gives a new twist to the 19th Century debate over the role of old evidence. In William Whewell (1847) view science progresses through increased consilience in which previously unconnected facts and generalizations are shown to instantiate a common principle. This is how, "in this Tree of Science ... two twigs unite in one branch". But since proposed new theories build on already obtained results it is no wonder that they "predict" these results – they would be ruled out of court at the outset if they did not. However, if we managed to find completely new facts which confirm our expectations we are compelled to accept the new theory. The problem, as John Stuart Mill and later inductive logicians pointed out, is that the degree of confirmation a theory enjoys should not reflect the largely accidental order in which science has revealed facts. Confirmation therefore is a logical notion between sentences expressing a hypothesis and evidence, and questions of who came to know what, when and why are simply irrelevant. Those following Whewell, including Popper and John Worrall (1978), in contrast insist that facts used as constraints cannot be counted in favour of the theory when its degree of confirmation (or corroboration) is assessed. Here confirmation becomes a historical and pragmatic notion, referring to the knowledge inquirers have at any given time.[8]

Now the advocates of the anti-divorce thesis take the historical view of confirmation for granted, thus adopting the heuristic stance on old evidence: discovery arguments accommodate already known evidence. In accordance with the predictivist thesis, they accept that such evidence provides *some* support for *h* (in the form of Bayesian prior probabilities, for instance). Although the main support for *h* comes from successful novel predictions, discovery is not irrelevant for justification. Nickles (1985) argues that this connection is not strong enough if we want to answer Laudan's challenge. His notion of generative justification presupposes the objective (logical, ahistorical) view of confirmation. From this perspective the division between old and new evidence is irrelevant. There is an objective, ahistorical relation between the evidence and the hypothesis (Schlesinger 1987, Howson 1984). Consider the following passage (Nickles 1985, 195):

> [I]t is not discovery in the sense of the initial conception of an idea which is important to justification here. Rather, justifying a claim establishes its "discoverability" in the sense that, regardless of how the claim was discovered or invented historically – regardless of how or why it was first thought of – it could have been discovered in the rationally specified manner had the necessary information and analytical techniques been available....Discoverability need not overlap discovery either temporally or logically.

This position is compatible with consequentialism, since there is nothing which prevents an inquirer, in addition to giving generative arguments, from testing an hypothesis in the usual consequential manner. And indeed Nickles argues that it is preferable to have them both.

Another response to Laudan's challenge comes from abstract computational models. Applying this approach Clark Glymour (1985) and Kevin Kelly (1987) recognized that some earlier models of confirmation could also be employed as models of discovery processes. Especially Kelly has later expanded this approach into an elegant formal model of inquiry. Subsequent results of this formal learning theoretic approach to discovery could be found from Kelly (1996, see also Kelly's article and its references in this volume). On this view there are logics of discovery which are not, however, models of actual human reasoning:

> [I]t is simply false that the logic of discovery is restricted to the study of actual, causal processes underlying actual, human behavior. First, it is not confined to the study of actual, causal processes. Given a programming system, the hypothesis generation procedures specifiable in that system exist abstractly in the same sense that proofs in a given formal system exist. So the logic of discovery is an abstract study whose domain includes all possible procedures. (Kelly 1987, p. 436)

Kelly (1987) counters Laudan's challenge that the logic of discovery is epistemologically irrelevant by showing that generation procedures and *post hoc* test procedures are computationally symmetrical. He shows how hypothesis generators can be built from test procedures and *vice versa* so that "when one sort of procedure squeezes through the door, the other is difficult to exclude" (ibid., p. 441). This symmetry establishes a strong form of the anti-divorce thesis, and can be called the symmetry thesis. Kelly's arguments rely on the supposition that the study of discovery is best conducted on the abstract, normative and computational level. From this perspective the symmetry of various computational tasks follows naturally. He, for instance, addresses Carnap's and Hempel's models of confirmation and counters the arguments these philosophers addressed against the logic of discovery by showing how their models of confirmation in fact depend on generative procedures. However, interesting as these results are, we can still ask whether they are adequate responses to Laudan's sceptical challenge, or whether they provide the key to the understanding of discovery processes. If one aims at describing complex patterns of problem solving in historical cases and at explaining actual decisions, abstract computational studies may not be of much assistance.

7. THE INTERROGATIVE MODEL OF INQUIRY AND ADAPTIVE LOGICS

There are some recent proposals for a logic or method of discovery, apart from the computational and AI models to be discussed in section 8. One of these, the questions-answers or interrogative view, builds on the idea that a good question often advances the aims of inquiry better than a thousand aimless inferences or observations, and that inquiry is a process of searching adequate or conclusive answers to questions. Another proposal, discussed later in this section, resorts to paraconsistent logics which could direct inquiry also in situations where constraints on acceptable answers are inconsistent.

The notion that questions and answers are landmarks in knowledge seeking is in fact one of the earliest insights on inquiry, for Aristotle's four causes can be seen as four types of answers to why-questions, and Plato's view of inquiry as anamnesis proceeds in question-answer terms. Kant (1968, B xiii—xiv) used it as a metaphor, suggesting also that Reason must take an active ("constructive") role: it "must not allow itself to be kept, as it were, in nature's leading-strings". Rather, it should approach nature like an appointed judge who compels the witnesses to answer questions which he has himself formulated." The view was championed by R.G. Collingwood (1939, 30, see also 1940) who observed that Bacon's *Novum Organum* (and Descartes' *Discourse on Method*) expressed the "principle that a body of knowledge consists not of 'propositions', 'statements', 'judgements' or whatever name logicians use in order to denote assertive acts of thought". Rather, knowledge consists of propositions "together with the questions they are meant to answer; and that a logic in which the answers are attended to and the questions neglected is a false logic." More recently, Michel Meyer (1980, 1994) has developed what he calls the problematological view in which all knowledge is an answer to a question. Just as in Kuhn's (1962) account the cognitive virtue of fruitfulness can be cashed out as the ability of a paradigm to give rise to well-defined research question, in Meyer's construction questions have value both through the answers they receive and the new questions they give rise to.[9] The view that questions and questioning, in a linguistic or some prelinguistic form, was a live theme in German philosophy, logic and psychology from Kant on (see Boudier 1988, Gale 1978, and the articles in *Synthese* 1981).

The erotetic view was reinstated analytic philosophy in the theory explanation, where Hempel and Oppenheim (1948) as well as Braithwaite (1959) suggested that explanations are answers to why-questions. But the interrogative view is also an attractive model for discovering new truths. Hanson´s (1958) abductions can be ragarded as infrences to explanatory answers, and the puzzle- or problem-solving models of e.g. Kuhn (1970) and Laudan (1977) can easily be couched in interrogative terms.[10] Dudley Shapere (1977) in turn proposed that the organizing principles in what he called scientific domains enable and suggest certain styles of questions and provide, at the same time, constraints on admissible of intelligible answer. In fact, he maintained, theories can be regarded as answers to questions arising from such domains (for similar but more recent views, see Jardine 1987 and 1991. As regards discovery, e.g. Gary Gutting (1980).

But does the interrogative model illuminate in any precise way the contexts of generation and pursuit? To serve as a logic of discovery the erotetic view should address the questions of how research questions arise, how they are nurtured in the context of pursuit, and how answers are sought. In a series of papers Jaakko Hintikka (1976. 1976, 1981a, 1981b, 1984, 1987, 1999) argues that the logic of questions can be expanded into an interrogative model of inquiry (the I-model) which focuses both on the outcomes of inquiry, the conditions on which answers to questions are deemed adequate and conclusive, and on the process in which answers are derived. Just as logicians have too timidly focused on studying the validity of inferences and left issues of discovery to too little attention, so philosophers of science have failed to capture the scientists' forward-looking strategies of fact-hunting and theory construction.

In the I-model the discovery process is viewed as a game in which an Inquirer attempts to establish a suitable cognitive objective, such as finding out whether a claim is true, which individuals have a certain property, or even why an event of state of affairs or regularity takes place or obtains. This process proceeds by posing an initial research question (yes-no-, whether-, wh-, or how- or why-questions) and by trying to find a conclusive or satisfactory answer to it. The details vary in accordande with the type of inquiry, but the moves fall into two main categories, deductive and interrogative (in more sophisticated games there also a definitory and assertoric moves). Experiments, unsolicited observations, memory consultations and help from friends and colleagues feed in new information and can be therefore be construed as answers to questions put to Nature or other source of information. These steps in the process of inferring and soliciting information are codified in semantical *tableaux* suggested by Beth and Hintikka: explicit tableau-construction rules govern deductive rules, and further rules govern admissible interrogative moves such as: to raise an operational question in a game (in a particular model M and its language) its presupposition must have been established; Nature's cooperates whenever she can; that Her answers are true in M.

There are, in fact, two types of questions in the model. First, there are the initial "big" questions which serve to define the goal of inquiry, expressed as propositional or wh-questions. Secondly, there are "small" or operational questions which serve to bring in information needed in answering the initial question. The main rationale for this distinction is that not all questions can be put to Nature without "begging the question". If one ask Nature why metals expand when they are heated, just as one can ask a teacher, empirical knowledge-seeking would be too easy. The restriction is particularly germane in the case of initial questions which concern explanations and large theory claims: there is no way one could put questions which require conceptual innovations to Nature. What the inquirer must try to do, then, is to find an indirect way by subjecting Nature to a series of small questions, and thus to corner Nature, one step at a time, in accordance with a strategic plan.

The model claims to be a successor to the old art and logic of inquiry (inquiry literally means questioning) by abandoning the sharp dichotomy between disovery and justification. It refines the familiar idea that asking the right questions at the right time is a crucial skill in inquiry. It advices to use a question usually whenever possible, for it is more efficient than a deductive move. The ordering of moves is also important, as is evident in experimentation and systematic observing. Supposing (often contrary to fact) for instance that Nature does give an unambiguous answer to an "experimental question", what should be the next move? The inquirer is invited to consider what information is going to help answer the initial question. Also the notion of scientific rationality get a new twist in the interrogative proposal. Lakatos alrady argued against instant rationality in his methodology of research programmes, but we can now also see that rationality can be more evenly spread also in the discovery process: it does not matter all that much how an idea is initially conceived, for the important thing is the possibility of finding multiple paths into a discovery. Rationality is not a matter of single moves but of strategies of finding independent roads to a result.

There are a number of open questions for the interrogative model so conceived. One has to do with the way initial and operational questions are discovered or invented – an erotetic logic which leaves this question open can hardly serve as a logic of discovery. Another question concerns the possibilities of being more explicit about the strategic principles which govern the search for answers. As to the rising of questions, this is where logic fades into pragmatics. Sylvain Bromberger has developed, in a series of papers (reprinted in his 1992), a full-blown logic of inquiry, including a pragmatic theory of explanation and "a theory of theory", around the erotetic idea. His suggestion for the emergence of initial questions is that knowledge comes from a combination of knowledge and ignorance. Ignorance is "not one big undifferentiated glop, one huge unstructured nothingness" but is rather shaped by background knowledge and other constraints on admissible answers. Thus, a theory specifies how questions arise and paves the way to a heuristics using, in the best of cases, search algorithms.

Kleiner (1970) suggested, more explicitly still than Shapere (1973), that scientific theories provide vocabularies which make some kinds of questions admissible. Thus for instance the languages of classical and relativist mechanics render different questions well-motivated or even illegitimate. Kleiner (1993, 122) elaborates on this theme by suggesting that e.g. the Darwinian research context (the Darwinian paradigm) ruled out, through its built-in presupposition that organisms always have natural parentage, creationist questions. On his view scientific inquiry indeed is a problem solving or question answering process. For him the logic of discovery is not a special form of inference from observation to theory, but rather a theory of the rationality of research which includes, as one of its most crucial objectives, the study of the principles bearing upon "the rational choice of problems, or epistemic objectives, and heuristics, or means to solving the problems". Kleiner critisizes e.g. Kuhn (1970) and Laudan (1977) for embracing notions of a paradigm or research programme which refer to progress as puzzle or problem solving – without giving an account of how problems or questions are chosen. His own proposal is to give principles which grade problems in accordance with their epistemic importance or weight. Thus problems concerning the core or theories or research programmes or about the principles governing the fundamental processes are to be given high priority. This account also leads to the view that some questions are a means of answering another and more important question: thus the question arising from the fudamental problem of evolution, "Do species transmute?", was approached by the subquestion "Are these specimen (or mockingbirds) distinct species or distinct varieties" and the subsubquestion "Do the variations in observable specimen correspond to variations among mockingbird species." Clearly, these subquestions obtain their importance and studying them their motivations from the initial question.

Kleiner's logic of discovery is not a global domain-independent theory but rather an account in which choices are justified locally against a relatively stable background consisting of ontological, conceptual, metascientific and empirical assumptions. He therefore joins those who doubt that the logic of questions could be turned into a logic of inquiry sufficient to deal with actual historical cases. In like manner Sintonen (1984) argues that scientific questions have both logical presuppositions (statements which must be true for the question to have direct answers at all) and pragmatic presuppositions which narrow down the set of admissible answers. The so-called structuralist theory notion, Sintonen (1989, 1990, 1996) suggests, gives a pragmatic account of centrality or importance: a paradigmatic theory which consists of a fundamental theory-element and a set of intended applications is essentially a hierarchically organized structure which both guides and constrains the search for answers to detailed questions within an application as well as for special laws for so-far unconquered applications. The logic of questions, then, has to be augmented with a rich enough notion of a theory to narrow down the set of possible answers.

How about choosing good operational questions, i.e., questions which help to construct answers to initial ones? Kleiner's Darwin-example shows that answering a question may require raising and answering one or more subquestions. More generally, the insight both in problem-solving and interrogative models is to replace initially intractable or woolly research questions by more refined and possibly answerable problems or questions. Andrzej Wisńiewski maintains that the I-model correctly emphasizes the importance of strategic principles in knowledge building, but that it fails to illuminate some of them, viz., the important principles of question-transformation. Nor does it provide an explicit logic for the process of organizing inquiries in hierarchial orders of questions, and like deductivism it is too liberal in that it also sanctions useless questions. What is needed is a way of selecting operative question which enhances the ultimate aim of the inquiry. Wisńiewski (1994, 1995, 1996) provides a logic of questions with an explicit syntax and semantics. This in turn serves as the core of the notion of a search scenarios, as follows. Suppose e.g. that you know that you are looking for a person, and that you know he has gone to Paris, London, Kiev, or Moscow. You can pose the four-fold which-question with these alternatives as possible answers, or raise four yes-no - questions. But if direct answers to these are not available, an indirect strategy might be an option. If you know that the person left for Kiev or Moscow if and only if he departed in the evening, you can check if he departed in the evening. Similarly, there you may know that if the person took a train he did not leave for London or Moscow, and having checked the train connections, you might be able to eliminate one of the initial alternatives. As a consequence, the original query gives rise to auxiliary questions such as when and how the person might have travelled. In Wisniewski's erotetic logic, based on multiple-conclusion logic, a question can, together with declarative sentences, imply erotetically another question. An initial question can then be indirectly answerable through answers tò the more specific questions. In this logic questions are used as premises as well as conclusions, and the result is a logic of question-trasformation. A skilful interrogator can therefore design entire search scenarions in which questions not directly answerable can be corned by strategic small questions.

Wisńiewski's erotetic logic and erotetic search scenarios take us one step towards a logic of discovery, but they do not show that pragmatic considerations are irrelevant. On the contrary, search scenarios can only be canvassed by help of background kowledge: the inquirer must know what information would be relevant and rule out some of the alternatives to itinial question. Interestingly, the logic is the same as in theory testing through experiments where the aim is to pick out one candidate as the most plausible answer. One limitation of the notion of a search scenario is that it is not applicable to ill-defined questions such as most explanation-seeking why- and how-questions. But would it be possible to extend the idea to these? What would a question-transformation for a why-question look like? Sintonen (1993) has argued that scientific theories are devices for turning ill-defined questions into well-defined questions. An example is provided by A.R. Wallace's erotetic reasoning to the theory of natural selection. His background knowledge included the analogy between human and animal population. And since animals breed more rapidly than man, and since evidence shows that they do not increase regularly each year, the magnitude of destruction each year must exceed that in human populations. "Otherwise," he wrote, " the world could long ago have been densely crowded with those that breed most quickly". And in an interesting passage (of discoverabality in Nickles's sense) Wallace (1905, I, 361-363) manages to turn the initially ill-defined why-question into a series of wh-questions and eventually testable yes-no-questions:

> Vaguely thinking over the enormous and constant destruction which this implied, it occurred to me to ask the question, Why do some die and some live? And the answer was clearly, that on the whole the best fitted live. From the effects of disease the most healthy escaped; from enemies, the strongest, the swiftest, or the most cunning; from famine, the best hunters or those with the best digestion; and so on. Then it suddenly flashed upon me that this self-acting process would necessarily improve the race, because in every generation the inferior would inevitably be killed off and the superior would remain – that is, the fittest would survive. ... The more I thought over it the more I became convinced that I had at length found the long-sought-for law of nature that solved the problem of origin of species..."

There are limitation to the I-model which requires separate attention, viz., that its games confine to particular language and a models. Similarly, they leave out ill-defined questions. As a result, erotetic logic is not well-suited to discoveries which require conceptual innovations. To remove this hindrance Diderik Batens and Joke Meheus have proposed a logic which makes it possible to examine the emergence of theories which are incompatible with the knowledge of the inquirer.

Batens (1997, 2000) and Meheus (1993, 1999) agree with Hintikka in thinking that discovery is too important to leave to psychologists and historians, but they depart company in favouring paraconsistent logics over classical logic. Meheus (1999) suggests that some varieties of paraconsistent logics, the so-called adaptive logics, can also be used to evaluate the inferential steps in actual creative processes originating from ill-defined problems. The goal for a logic of methodology of discovery must be to try to account for the way in which a result is derived (deduced, induced, abducted) from what is already known or accepted – and if this could be done the scientific community would be able to assess its epistemic credentials. But if one takes classical logic as the standard for all good, valid or rational reasoning, the most creative types of processes involving abductive analogical inferences must be rejected as non-rational. This would be a serious blow to a methodology of discovery.

What are these troublesome problems, and what is the import of adaptive logics? Batens and Meheus distinguish between problems with inconsistent and incomplete information. Scientists often face problems in which the information is inconsistent and which characterstically require non-monotonic reasoning: when new premises are added formerly accepted ones must be deleted. But non-monotonic reasoning may also be required when information is incomplete and the inferences based on ampliative rules. A result obtained on the basis of an ampliative rule may be deleted when the negation of the result is derived (monotonically) from the premises added. Apart from being non-monotonic such a reasoning is also explicitly dynamical, and Batens and Meheus explicate its features by distinguishing between conditional and unconditional derivation. A sentence derived conditionally at some stage in a proof may at a later stage become underived, namely, when the condition is no longer satisfied. In conditional derivation the application of a rule depends some specified conditions, and when at some stage of inquiry they fail, the conditionally derived sentences are 'marked' as OUT (Meheus 1999, 326).

Classical logic in turn is static, and it does not allow withdrawal of a sentence at a later stage. But whereas, by classical lights, a scientists's reason must, in the face of inconsistent constraints "go on holiday", Batens and Meheus insist that adaptive logics can help. These logics are able to localise or isolate troublesome inconsistences in ways which do not lead to rampant sanctioning of all sentences. Classical logic allows the deduction of any sentence from an inconsistency, thus making a principled course of action impossible. Adaptive logic is not paralysed by specific violations of logical presuppositions, since they can restrict their rules of inference in the face of specific troublesome applications.

As an example of inconsistent information, i.e., a case in which there is too much information to proceed, Meheus (1993) gives Clausius's reasoning to his theory of thermodynamics. There were reasons to believe in the conservation of heat as well as in its not being conserved. In Meheus's reconstruction Clausius nevertheless was able to derive Carnot's theorem from these two incompatible approaches (due to Sadi Carnot and Joule), using similar Reduction ad Absurdum arguments. Both derivations started from the negation of Carnot's theorem and concluded in contradictions, but there was a difference. In one case Clausius only used the premisses and not the hypothesis (that the negation of Carnot's theory is true). In the other argument needed that hypothesis, which is why Clausius only regarded this as a proof of Carnot's theorem. In both cases there is reasoning from an inconsistent set of premisses, but only one is able to say something useful on the hypothesis under examination.

Now the question is, what logic can make sense of this? Batens and Meheus suggest that a logic which would be able to capture such creative reasoning is neither classical logic nor any paraconsistent logic. It should be able to "adapt to its environment by 'oscillating' between a paraconsistent logic and classical logic" (Meheus 1999, 332). In an inconsistent neighbourhoods it should behave like a paraconsistent logic, in consistent neighbourhoods like classical logic, thus 'localizing' the inconsistencies and helping, by purely logical means, to decide whether to accept a hypothesis or not. Apart from these inconsistency -adaptive logics Batens and Meheus also define ampliative-adaptive logics for cases where principled ampliative inferences are needed. The upshot is, in a sense, parallel to what critics of the Duhemian principle of the underdeterminacy have proposed. The principle says that, in a testing situation and in the face of a falsifying observation, logic cannot force the inquirer to give up the hypothesis. But, the critics say, scientists are often able to allocate blame in a reasonable way – after all, one should be able to learn from experience.

8. ARTIFICIAL INTELLIGENCE AND COGNITIVE PROBLEM SOLVING: THE MENTALISTIC AND COMPUTATIONAL TRADITION

By far the most detailed and the most productive model for discovery has been the computational problem-solving paradigm initiated by Herbert Simon and his collaborators (see Simon et al. 1981, Langley et al. 1987). The paradigm applies the concepts and procedures of the computational theory of problem solving, one of the first major achievements of AI research (Simon and Newell 1977).

On the computational view a problem is solved through an algorithmic search in a well-defined problem space. A problem space represents all the relevant variables and their possible values, including an initial stage, a goal stage, and the possible paths the problem solving operators can take while trying to reach the latter from the former. It is easy to see that this kind of situation could lead to a combinatorial explosion. Consequently, heuristic rules are needed to reduce the size of the search space and the number of search paths. The claim is that by this kind of problem presentation one can capture the cognitive processes operating behind most discovery episodes.

To back up their view the AI researchers have constructed a series of existence proofs, starting with computer programmes such as Lenat's (1977) mathematical discovery programme AM and Simon et al.'s BACON which implement powerful enough heuristic search strategies to formulate quantitave laws. Later research has widened the scope of these models and made their picture of the discovery process more sophisticated, but many of the basic assumptions as well as difficulties have remained from these early efforts.

In practice, computational models contain elements which are closely related to the philosophers' and historians' accounts of science. For instance, Langley et al. (1987, 18) describe the scientific enterprise as consisting of four kinds of activities: gathering data, finding parsimonious descriptions of the data, formulating explanatory theories, and testing theories. These activities are all interrelated, and they all start in promissory terms and require ingenuity to be carried out. This is also true of discovery: there simply is no powerful factory method or assembly line for scientific truths in basic inquiry. Yet, they strongly object to the image of the scientific enterprise as relying on intuition and luck. Instead, there are numerous heuristic methods whose range may be limited but which could be very powerful tools in highly specialized domains. Strong heuristics can employ domain-specific information to boost their problem solving powers while weaker ones trade power for scope. A weak heuristics, although stronger than blind trial and error, comes with less guarantee of correct results, but it is characteristically applicable across a wide variety of tasks.

The weak but general heuristics employed by Simon et al. are based on a kind of means-ends -analysis. These heuristics compare the result of an operation to the desired goal and choose a further operation by analysing the deviation between the two. For instance, BACON.4 employs a small set of data-driven heuristic rules which trace invariances and trends in numerical data sets:

> (H1) If the values of two numerical terms increase together, then consider their ratio.
>
> (H2) If the values of one term increase as those of another decrease, then consider their product.

It then uses the heuristic to formulate hypotheses and to define new theoretical terms. With a help of these procedures it has rediscovered Ohm's law, Archimedes' law and many results of chemistry.

The ultimate goal of Simon et al. is philosophically challenging. They try to construct a normative theory for scientific discovery which would contain criteria by which the efficiency of the heuristic discovery methods could be evaluated. To be more specific, the theory should justify the following type of contingent propositions (ibid., p. 45):

> If process *X* is to be efficacious for attaining goal *Y*, then it should have properties *A*, *B*, and *C*.

These propositions define norms which help scientists to attain their goals more efficiently. Historical case studies, then, should specify what the relevant properties *A*, *B*, and *C*, of the most rational discovery process are. Ultimately this calls for the determination of the best heuristic to solve a given discovery problem. Normative conclusions are then achieved by studying historical cases which purportedly simulate the original discovery processes. As such, this is a bold example of a naturalistic perspective on normative epistemology.

But the ambitious goal of using computational models to simulate historical discovery processes makes several controversial assumptions. One is their alleged psychological realism, i.e., the claim that these models simulate, step by step and in an approximately accurate manner, scientist's original thought processes (Langley et al. 1987, Simon 1992). This view, however, is anything but universally accepted. One source of difficulties is Simon's historical methodology, the so called protocol analysis, in which a subject's thought processes are inferred from her verbal reports (Simon and Ericsson 1984). But there are few actual cases with direct verbal protocols, and the researcher has to rely on the subject's correspondence, laboratory notebooks, diaries, published writings etc. It is easy to see that this may not be a sufficient basis for drawing conclusions about actual thought processes (Downes 1990). The data may contain too much retrospective rationalization to be psychologically reliable, thus corrupting the original idea of protocol analysis. And even in contemporary episodes from which enough evidence could in principle be gathered, the protocol analysis may give misleading results (cf. Woolgar 1976).

There are also other problems. For instance, BACON has been critisized for falling short of human problem solving, in scientific contexts and elsewhere, for it requires that the search space which defines the possible solutions is well-defined and indeed given to the programme. The objection claims that typical scientific problems, especially those requiring explanatory theories as solutions, do not fulfil these conditions: they are ill-structured problems with open-ended search spaces (see Belnap and Steel 1976, Meheus 1999, Sintonen 1989). Moreover, the AI - approach seems to trivialize Meno's paradox according to which one can only look for a solution if one knows it. A non-trivial solution to the paradox requires that we specify the goal-state of problem-solving inquiry only partially, not fully, as in the AI-models. These difficulties are admitted by the proponents of AI-models, as is evident in the following passage (Langley et al. 1987, p. 7):

> Indeed, finding problems and formulating them is as integral a part of the inquiry process as is solving them, once they have been found. And setting criteria of goal satisfaction, so that indefinite goals become definite, is an essential part of formulating problems.

Langley et al. admit that these complexities are not accounted for in current models. However, they suggest that the difficulties could be handled by decomposing the complex problems with indefinite goals into sets of simpler problems their models can solve. Problem formulation can, then, be an incremental process involving several intermediate stages. This is a welcome move. Yet, it can be asked if this problem reduction strategy is sufficient. It may help to solve a problem, but not to formulate it more sharply. It can be claimed that AI-models are troubled by difficulties in some basic assumptions. They cannot adequately formulate the goals in complex inquiry, thus failing to solve the Meno paradox.

Another objection is that discovering a quantitative law within such a well-circumscribed settings fails to exhibit what every learner and innovative researcher has to engage in, viz., concept formation. And in fact it has been doubted if computer models could make novel discoveries in the strong sense. In view of the first objection based on Meno's paradox, this seems unlikely. From this point of view the only algoritmic way to learn or discover something is to make explicit the information already implicitly coded into the programme and its data bases. Such programmes could illustrate e.g. computationally demanding theorem proving in mathematics, but they would have serious limitations in scientific discoveries which involve conceptual innovation. The most demanding type of creativity does not respect this limitation, as is testified by serendipitous discoveries. These insights are not easily captured by the existing computational models.

A third complaint is that the AI approach either totally ignores the social nature of the discovery process or cannot deal with them properly because the computational models presuppose cognitive individualism and, consequently, do not mirror the social division of labour within scientific communities (Brannigan 1989, Collins 1989, Downes 1990, cf. section 9). According to the critics, the model of science which consists of sets of isolated computational agents is seriously flawed. The social dimension of science is essential to the whole project of modern science, since its day to day practise is based on large research groups, and computational models can, at best, give a retrospective and highly idealized description of what happened in the group-based discovery process. Given these limitations they can hardly explain, in particular cases, what happened, when and why. And given the difficulties with psychological realism and historical analysis, even their descriptive accuracy could be doubted.

Should we conclude, therefore, that heuristic search models are only good for the retrospectively simulation of relatively routine problem solving? This conclusion would be too hasty. Although it is still doubtful whether machine intelligence now or in the near future matches human intelligence the obstacles need not be insurmountable. As we noted above, BACON-type models are designed for discovering quantitative laws from numerical data. To meet the further requirement of conceptual enrichment Newell and Simon (1976) recognised that many scientific theories involved laws of "qualitative structure." For instance, Pasteur's theory of contageous diseases was initially little more than the inarticulate claim "If an organism is suffering from a disease that appears to be contagious, look for infestation by a micro organism that may be causing it." Koch then enriched the theory by establishing specific connections between particular bacterial diseases, but like many biological or medical theories (the cell theory, the theory of natural selection) the theory remained qualitative for a long time. Yet it contained new concepts which enabled the formulation of a vast number of more specific claims. How does a discovery method account for the emergence of such qualitative but extremely important conceptual innovations?

The response to these queries was a series of problem solving models with increasingly sophisticated and strong built-in heuristics, and programmes which were able to modify their procedures while they ran. By the GLAUBER and STAHL range programmes one could try to discover theories. The former creates taxonomies from the data and then uses these to produce qualitative laws. The latter discovers structural theories of chemical substances which are based on observed reactions. In a similar vein, Darden (1990, 1991) studied heuristics which govern the revision and correction of the theories when they confront anomalies (similar procedures operate in STAHL programmes, too).

Besides these procedural improvements, the AI-approach was enhanced with more sophisticated means for knowledge representation. For instance, Forbus' (1985) work on qualitative processes enables us to represent theories about such processes and to make qualitative predictions. Some models even try to incorporate imagery, thus acknowledging their important role in scientific thinking (Miller

1986). There is also the difficult problem of combining the procedural and representational parts of the model. The goal of these integrated models is to enrich our picture of the computational discovery processes by incorporating different elements into one model. Thus, Kulkarni and Simon's (1988) KEKADA could design and run experiments which test alternative hypotheses while some other models use qualitative generalisations as heuristics to constrain the search of numerical laws.

The AI view of inquiry and discovery in terms of problem solving does, then, recognise the crucial role of concept formation. What about creativity and the objection that the AI models can only capture the relatively routine problem solving processes? The consensus within the problem solving paradigm seems to be that creativity is exploration of a conceptual structure or space. To be creative is to master a set of rules which define admissible and desirable outcomes. And creativity comes in degrees. Exploring an antecedently well-defined structure may of course bring about applications new to the individual or community (Margaret Boden's 1994 P-creativity). But individuals who, having explored a conceptual structure, manage to take a glimpse beyond what the rules specify can be said to be more creative still. Finally, the real giants are persistent individuals who manage to discern or articulate new rules which transgress the existing ones (Sintonen, forthcoming).

The view has helped to demystify creativity by opening the field for tools from artificial intelligence and cognitive science. We can only briefly mention some of the results. For instance, if the structures to be explored are networks of concepts tied together by nodes and links, restructuring and abandoning one differs from mere addition or deletion. The result is a cognitive model for revolutionary change as well as for tinkering (see Thagard 1992). As a result we have conceptual systems in which concepts are organized into hierarchies by help of various types of links, such as kind, instance, rule, property and part links. To the extent scientific conceptual systems consist of networks of nodes and links analogous to other cognitive systems or structures, we have an account of conceptual change. This is to go beyond the Oomph also because it suggests how ideas arise from analogies (and discrepancies) in the structures under exploration. From the point of view of understanding creativity and discovery the most important feature of Thagard's notion of conceptual change is that it uses tools which have been rooted in well-established research programmes in AI, cognitive psychology and cognitive science. Furthermore, it has the crucial advantage that it provides perhaps the most plausible account available for the importance or significance of a conceptual change, whether adding or substracting or altering (which migh be viewed as substraction followed by addition). Finally, it enjoys extra appeal because it gives both tradition and innovation their due. Conceptual revolution involve, by definition as it were, dramatic replacing of major portion of the conceptual systems. Nevertheless continuity prevails because some of the links to other concepts are retained.

The cognitive paradigm builds on the idea that scientific reasoning is on a par with everyday reasoning. If this view is adopted, we can use the results of cognitive studies of ordinary reasoning and problem solving to supplement the available historical evidence, and in this way achieve a more accurate model of scientific problem solving. This kind of model could show in detail how the individual scientists reasoned to their novel results. The result is what Nancy Nersessian (1992) calls cognitive-historical analysis of discovery episodes. It combines the insights of historical case studies with the results of modern psychology and computer science. Nercessian's (1984) own case study is a good example of the cognitive-historical approach. She traces the development of the concept of electro-magnetic field from Faradays's "lines of forces" to Einstein's gravitational field concept. She further suggests that the representational tools of cognitive science could be employed to make sense of the dynamics of conceptual change in science. For instance, the prototype theory of concepts could capture what physicists' various conceptions of field had in common and which parts of it remained and which were revived during its development in the 19th Century. Ryan Tweney and his co-workers have elaborated, in a series of papers the idea that inquiries can be interpreted as moving on different levels, from overall purposes to heuristic scripts and schemes to goals and subgoals, and finally, to states and operations, with specific heuristics for each level. (Simon 1996, Tweney 1989, Duncan and Tweney 1997). More recently, this approach has led into the accommodation of "presymbolic" processes and representations into discovery, thus also illustrating representational borrowing between fields of inquiry (Duncan and Tweney 1997). One of Tweney's claims is that this cognitive framework is applicable also outside the natural sciences.

9. SOCIAL MODELS OF DISCOVERY

The most recent competitor of the mentalist and cognitivist view of discovery comes from social historians of truth. Their claim is that scientific discoveries are not mental episodes at all but rather social negotiation processes. It finds support in studies which show that the identification of discovery events or corresponding knowledge claims is not as straightforward as is usually thought (e.g. Fleck 1979, Kuhn 1957, 1970, Woolgar 1976). It is not always possible, the challenge says, to identify a unique event or a point of time when the discovery occurred. Rather, the claim that such identification is possible depends on various controversial assumptions such as the role of retrospective explanation in history of science.

Consequently, the proponents of social models of discovery have provided detailed examples of how social attribution and negotiation practices function in the creation of scientific discoveries. Although these analyses seem to clash with the corresponding cognitive/individualistic analyses, we will suggest that the differences are not insurmountable.

How does the social dimensions affect our perception of scientific discovery, then? Augustine Brannigan (1981) claims that the status of a scientific discovery always depends on the societal practices by which a series of events is classified as a scientific discovery in the first place. According to him the practices responsible for the identification and announcement of discoveries often go unnoticed. When we think about famous scientific discoveries we usually start to reconstruct their birth process in mental terms, as great intellectual achievements in individual minds. The questions of how their identity is established and how they are attributed to the heroic discoverers do not even arise.

There is a strong linguistic underpinning in Brannigan's analysis of social attribution. He tries to explain discoveries by revealing conceptual resources or methods which form the basis of the speaker's classificatory practice. The purpose is to explain the discovery talk of members of society by postulating that their linguistic practice is based on a model which we may reconstruct as follows (1981, 71-77):

X is a scientific discovery in a society *Y*, if
1) *X* is expected to occur in *Y*.
2) *X* is achieved in the course of action which is classified as scientific research in the society *Y*.
3) *X* is at least a locally true or valid result.
4) *X* is a new result in sense that it was not earlier classified as a discovery in community Y before.

In other words, *X* was a solution to a problem which is based on a scientific tradition accepted in *Y* (conditions 1 and 2). If the result is classified as discovery, it has to be true. But truth is here understood to be relative to society *Y*: *X* is true if it is accepted as true in *Y*. According to Brannigan, it is useful to adopt a kind of methodological relativism and to concentrate on beliefs which were accepted as true in some local research community even if the analyser already knows better. This is a standard historicist or anti-whig position in science studies.

One of Brannigan's examples of social attribution is the alleged rediscovery of Mendel's laws around the year 1900. He claims that Mendel's results were not rediscovered at all. Instead they were announced as discoveries only in 1900 when they could be connected to the problem of inheritance within the Darwinian evolutionary theory. When Mendel originally presented his results in the 1860's they were not ignored (as is usually claimed) but their significance was not understood either. The reason is that Mendel's researches took place within a different social context and scientific tradition. His intellectual home was not the Darwinian theory of evolution by natural selection but rather the competing tradition of the plant hybridists. Mendel's experiments and conclusions were not intended to be general laws of inheritance but only "laws valid for Pisum", which is why they were not perceived to be revolutionary at all.

Simon Schaffer's (1986, 1994) account of discovery is reminiscent of Brannigan's but as a social historian of science he traces the origins of heroic individual stories to the institutionalization of science in the nineteenth century (Schaffer 1991). He shows how historical evidence underdetermines both the cognitive identification of ideas and the dates of their occurrence. In this vein, Schaffer distinguishes four trouble areas for the "mentalist model of discovery" (1986, 391): a) the isolation of discovery in time and space; b) the authorship of discovery; c) the preconditions of work which generates discovery; and d) the process by which a discovery is recognized as a discovery.

According to Schaffer there are no ahistorical and transcendent criteria of discovery. The best we can do is to employ the constructivist methods and to study the debates of experts in which the new ideas were discussed and in which their interpretations were gradually settled. The constructivist approach focuses on the "fixing" of discoveries which are "linked with assent to the matter of fact and to the identity of discovery": a new fact is replicated and it is given an author. Such fixing involves complex negotiations inside the scientific community. The replication of the results and the authorisation of the discoveries are social accomplishments and communal decisions concerning discovery stories, not self-evident facts or discovery events. There is "no event which corresponds to an automatic or instant discovery" (1986, 397) Instead, the facticity of discovery "is closely connected with communal and retrospective decisions about discovery stories".

It is arguable, nevertheless, that many of the criticisms of the cognitive approach focus on simplified models (such as Arthur Koestler's bisociation model, see Brannigan 1981) whose status is suspect also on individualistic terms. We suggest, therefore, that both the individualist, cognitive and the social aspects of discovery are needed.

Brannigan (1981) argues against the mentalistic approach by claiming that it confuses learning and discovery. According to him, these approaches could not make any distinction between the two: both are instances of knowledge acquisition at the psychological level. A student of physics can replicate the mental operations behind a famous discovery while doing laboratory exercises. The learning processes may be identical but only the original result led to the discovery. So, learning is not sufficient for discovery.

If we approach discoveries from the individualistic perspective, there evidently is a close connection between learning and discovery. Could we, then, specify the relationship between learning and discovery in more detail? And how serious is Brannigan's objection? This objection may have some bite against simplified individualistic models but not against its more sophisticated versions. It seems that Brannigan assumes that the individualistic approach is tied to bare mental operations and cannot handle the constraints arising within a scientific and cultural context. But this is not true of problem-solving models in general. In addition to adequate reasoning abilities, a succesful problem solver has to have a great deal of background information about the problem situation.

Moreover, it is a common view that discovery is a special case of learning and their relationship could be specified in many ways. A good way is to argue that discovery processes are learning processes which fulfill certain additional conditions such as: (i) there is no external learning bias in discovery and (ii) in a case of discovery, it is the first establishment of the result that counts (see Zytkow 1993). Condition (i) says that there should not be any instructor who already knows the right answer. And there are also other possibilities. It is also evident that the learning process presupposes as much background information as the discovery process (see Kiikeri 1997).

Brannigan's analysis takes it for granted that the crucial question is that of analysing how events are classified as discoveries. He then explains discoveries by explaining the discovery talk prevalent in the relevant community. But for the individualist there are also other interesting questions, for instance those concerning the efficient organization of one's inquiry (choice of methods and research heuristic etc.).

Mentalistic construals can also be critisized on other grounds. One objection arises from the familiar underdetermination problem. There is never enough evidence to describe a discoverer's original thought processes. – any model of cognitive problem solving applies to any discovery episode if we make suitable idealizations and omissions. If we simplify matters suitably and omit enough details we can, for instance, use the schema of abductive inference to account for any learning process from Kepler's discoveries to animal problem solving. But this argument can be countered by noting that compatibility with historical evidence is not the only constraint. For instance, cognitive studies of reasoning and problem solving may be relevant here. Furthermore all approaches to discovery, individualistic or social, are subject to this underdetermination problem. Is there any reason to suppose that the social attribution model is better in this respect? Hardly. It seems that it relies on the implicit assumption that it is easier to obtain evidence about social interactions than about the inner workings of the mind. But is there such an epistemological barrier between the external social realm and the internal mental realm. Of course written documents (such as letters) tell more about the contacts between scientists and their social roles than about what they really think. But if we adopt a broader view about the relevant evidence it is not obvious that social psychological theories or models of social interaction are on firmer ground than the corresponding cognitive models of reasoning. In fact, the two approaches seem to complement each other. Furthermore, even the assumption that the model of reasoning is psychologically realistic could be abandoned. In this situation, we could still proceed from the individualistic basis and describe parts of the discovery process informatively. Published arguments, for instance, could be enough to render the result an important discovery even if they do not show how the discoverer originally arrived at her ideas. This observation receives an interesting treatment in Nickles' account of discoverability arguments and generative justification (See section 6 above).

Furthermore, the proponents of the individualistic approach need not confine to but one general model (such as the abductive one) which is then adjusted to every historical episode. Heuristic methods, for instance, could vary from case to case even if certain basic assumptions about problem-solving remain constant. The fact that the same model could be adjusted to every case does not imply that it is reasonable always to do so. However, maybe the most interesting objection towards purely individualistic accounts is that reducing discoveries to the psychological level is only *post hoc* explanation with no predictive power and generalizability to other cases. We identify, the argument goes, a learning process as a discovery process because we already know that the result of the research process was a significant discovery. Consider then the situation in which we examine ongoing research from the individualistic point of view. If we do not already know the status of the result, we cannot distinguish significant discoveries from routine results or even failures. And because the status of discovery is achieved only by a social attribution the individualistic story would not explain what makes a result discovery (unless we trivialise the problem by calling any new result a discovery).

But again, what is the role of social attribution if the problem-solving approach is adopted? Consider first the way problems arise in the first place. The quick answer is that they emerge from incompatibilities or gaps in the relevant background knowledge within a research tradition. An expert problem solver usually knows a great deal about the situation in advance. So every initiated expert in a research area knows in advance problem solutions which would constitute significant discoveries. Hence the most interesting questions from the individualistic problem-solving perspective are the importance of problems and the assessment of proposed solutions, i.e. questions corresponding to Brannigan's conditions 1) and 3).

Can these questions be answered without relying on the retrospective, attributional perspective? There are no easy answers to these questions (cf. the discussion of pragmatic commitments in the previous section). Fortunately we do not need definitive answers here, since our analysis reveals that the attributional perspective is not independent of individualistic problem-solving either. Brannigan's model already presupposes that important problems could somehow be specified (recall Brannigan's criterion 1) and that the new results are at least locally accepted as true (criterion 3). Of course, these questions cannot be answered on purely individualistic basis either, since they depend on the social judgment of the research community. But it is clear that importance and acceptance of important parts of the discovery process involve of individualistic problem-solving after all. We therefore propose that the cognitive and the social are not opposed to one another.

Both views, one based on the individualistic problem solving and other on the social attribution, contain a grain of truth. There is a close relation between discovery and learning. Even if historical evidence might be insufficient to reveal the details, learning processes are needed for discovery. However, acknowledging their importance does not mean that they are sufficient for discovery, precisely because the social context of the research community regulates which results are picked but as important and who are those to be honoured as the discoverers.[11]

Matti Sintonen
University of Tampere

Mika Kiikeri
University of Tampere

NOTES

[1] Important recent monographs, articles, anthologies and reviews on scientific discovery include Meyer 1979, Nickles 1980a, 1980b, 1981, 1985, 1988, 1990, Kleiner 1988, 1993, 1997, Schaffner 1985, Giere 1992, Simon 1977, Schaffer 1986, Brannigan 1981, Brzezinski, Coniglione and Marek 1992, Kantorovich 1993, 1994, Langey, Simon, Bradshaw and Zytkow 1987, Boden 1994b, Campbell 1974a,b, Glymour, Scheines, Sprites and Kelly 1987, Jason 1989, Kelly 1987 and 1996, Koertge 1982, Magnani, Nercessian and Thagard 1999, Hintikka and Vandamme 1984, Shrager and Langey 1990, Thagard 1992, Nersessian 1987b, 1992, Meheus 1999. We are particularly much indebted to Thomas Nickles both for his writings and for personal advice and support.

[2] In the traditional 19th and 20th century accounts Newton's concept of induction has been interpreted as the standard Humean one, related to the consequentialist view of justification. Hintikka (1992) has, however, suggested that there are two different concepts of induction. There is the usual Humean (e.g. ordinary enumerative) induction from particulars to generalizations, but also the Newtonian variant in which induction refers to the reconciliation of already available but partial empirical generalizations into more general laws. Hence, the Newtonian sense of induction is a form of inference in which general results can be established on the basis of only one or few pieces of evidence. It is a kind of instance induction, ampliative inference in which the role of evidence is reduced to a minimum. It has

not received much attention although it seems to have an important role in history of science as well as in various contemporary applications. Arguably, Newton used this form of induction both in *Principia* and *Opticks* (Hintikka 1992a, see also Mäenpää 1993).

[3] Scientist's words, e.g., in their memoirs or interviews, cannot of course be taken at face value, for they usually are subject to the same sort of intentional or unintentional omissions and additions as other testimonies based on memory. Methodological caution is both important and difficult in assessing the evidence for any model of discovery. See also the discussion in sections 8 and 9.

[4] It is of course far from clear that this naturalistic strategy for explaining normative principles is successful. Bacon observed that our minds are disposed to accept hypotheses which are pleasing to the mind, and Thomas Reid wrote that "men are often lead into error by the love of simplicity which disposes us to reduce things to few principles and to conceive a greater simplicity in nature than there really is..."
(*The Essay on the Intellectual Powers of Man*, quoted from Laudan 1970, 110). "Our belief in the continuance of nature's laws is not derived from reason, it is an instinctive prescience of the operations of nature, very like to that prescience of human actions which makes us rely upon the testimony of our fellow creatures..." The naturalistic strategy surfaced anew around the turn of the century, e.g., in the writing of the leaders of pragmatism and positivism, Charles Peirce and Moritz Schlick. What is needed for abductive inference to get going is a pool of plausible hypotheses, and indeed Peirce (1931-35, 5.591) accepted that we are innately disposed to form certain types of hypotheses. Thus, although abduction cannot guarantee truth, naturalistic and straightforwardly biological considerations may suffice to explain why belief formation on the basis of perception, as well our inferences, usually result in true (or truthlike etc.) beliefs. Moritz Schlick (1974/1925), finally, gave this strategy an evolutionary grounding: we may find pleasure with some architechtonic features of belief systems (such as unity, harmony and order, because these features have been conducive to survival. Knowledge in general contributes towards "the preservation of the individual and the species", and the drive for knowledge undoubtedly falls under this general principle: "In its origin, thinking is only a tool for the self-maintenance of the individual and the species, like eating and drinking, fighting and courting." And Schlick goes on to say that the mechanism of judging and inferring contributes towards better adaptation to the environment than automatic association which focuses on typical cases. The difficulty is that these prejudices may favour falsities rather than truths. (See Laudan 1970, Sintonen 1990, and Bradie's and Siegel's articles in this volume)

[5] Scott Kleiner (1993) also combines serendipidity with deliberate planning within his interrogative view of inquiry. His example is the same as Gruber's, Darwin's theory of natural selection. Kleiner, however, spells out explicitly Darwin's strategic steps in answering the question "Do species transmute?", suggesting that Darwin approached it through subordinate questions concerning evidence needed to rule out alternative answers, and ways of finding such evidence. Although the details of the answer were results of contingencies and therefore serendipidity, it cannot be denied that Darwin aimed at answering the initial question and in that sense aimed at a discovery.

[6] Note also that the rationality of generation is not incompatible with a retrospective explanations in socio-historic or psychological terms. If the logic and the rational principles used in the interrogation of nature are sufficient, the process could in principle be predicted in advance and explained in retrospect. If, however, logic and nature are insufficient we need to resort to tools available in history and the social sciences. The crucial thing to see is that the internalist and the externalist accounts employ, in part at least, different principles of intelligibility and explanation. Internalist explanations adopt the point of view of the "inner logic" of science (see Hesse 1970, 135), focusing on rational explanation and rational choice by assuming that scientists (at least at their best) use the rational principles of the logic of

inquiry and make the best choices available. Externalists have as an additional source the tools of history and the social sciences: in principle the naturalistic and causal patterns of explanation which are available in explaining people's actual behaviour in non-scientific contexts (for an early view of a partly externalist philosophy of science, see Toulmin (1972).

[7] For a lucid account of background knowledge, see Kleiner 1993.

[8] So, as Worrall observes, Popper's denial of a logic of discovery is coupled by the consequence that constraints used in generating a hypothesis are relevant for justification, but in a negative way! See also Nickles 1985 for discussion.

[9] Michel Meyer in fact suggests that Plato directed the philosophers′ attention away from question-based rationality towards knowledge captured in eternal propositions: "Plato presents us with a propositional view of the logos through his dialectic, which encompasses science". Mayer′s view seems to be that Plato rejected Socratic questioning. While Socrates had grounded his philosophy in interrogation and the unvailability or even impossibility of a previously given answer, Plato gives a view in which questions are only a rhetorical device for bringing to mind a dormant set of propositions

[10] The relationship between interrogative and problem solving views is a relatively little discussed topic. Thomas Nickles (1988) does not wish to make a sharp distinction between problems and questions, but notes that problems have more depth arising from the background theory and other constraints on the solution of the problem. The result is the constraint-inclusion model (CI model) in which problems are defined as sets of constraints and the demand to give a solution which satisfies the constraints. Nickles also argues that Hamblin's (1958) dictum for defining questions, also adopted by Belnap and Steel (1976), in terms of their direct answers is unsatisfactory: one can understand a question without knowing their answers. Sintonen (1984a, 1996) distinguishes between knowing the set of answers and knowing the type of answers which would be adequate (e.g., names or definite descriptions for who-questions). He also argues (1985) that the difference between questions and problems is unimportant if one distinguishes between surface questions and the deep pragmatically enriched questions which actually commit to certain types of "admissible" answers – there is, on such a view, always more to a question than meets the ear.

[11] Like in Kuhnian normal science, scientific community only recognises a small portion of possible problems significant. But we also know from Kuhn's and others' accounts that this is not the whole story.

REFERENCES

Achinstein, P.: 1970, 'Inference to Scientific Laws', in R. Steuwer (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. V, University of Minnesota, Minneapolis, pp. 87 - 111.

Achinstein, P.: 1971, *Law and Explanation*, Oxford University Press, Oxford.

Achinstein, P.: 1980, 'Discovery and Rule-books', in T. Nickles (ed.), 1980a.

Achinstein, P.: 1987, 'Scientific Discovery and Maxwell's Kinetic Theory', *Philosophy of Science* **54**, 409-434.

Achinstein, P.: 1993, 'How to Defend a Theory Without Testing It: Niels Bohr and the 'Logic of Pursuit'', *Midwest Studies in Philosophy* **18**, 90-120.

Agassi, J.: 1980, 'The Rationality of Discovery', in Nickles (ed.), *Scientific Discovery, Logic and Rationality*, 1980a, pp. 185-200.

Bacon, Francis ([1620] 1994): Novum Organum. Translated and edited by Peter Urbach and John Gibson. Chicago and La Salle: Open Court.

Batens, D.: 1997, 'Inconsistencies and beyond. A logical-philosophical discussion', *Revue Internationale de Philosophie* **200**, 257-271.

Batens, D.: 2000, 'A survey of inconsistency-adaptive logics', in D. Batens, C. Mortenson, G. Priest, and J. P. Van Bendegem (eds.): *Frontiers of Paraconsistent Logic*, Research Studies Press, King's College Publications, Baldock.

Belnap, N. D. and T. B. Steel, Jr.: 1976, *The Logic of Questions and Answers*, New Haven and London.

Boden, M.: 1994a. "What is Creativity". In M. Boden 1994b, pp. 75-117.

Boden M. (ed.): 1994b , *Dimensions of Creativity*. Cambridge, MA, and London: The MIT Press.

Braithwaite, R. B.: 1959, *Scientifc Explanation. A Study of the Function of Theory, Probability and Law in Science*, Cambridge University Press, Cambridge.

Brannigan, A.: 1981, *The Social Basis of Scientific Discoveries*, Cambridge University Press, Cambridge.

Brannigan, A.: 1989, 'Artificial Intelligence and the Attributional Model of Scientific Discovery', *Social Studies of Science* **19**, 601-13.

Bromberger, S.: 1992, *On What We Don't Know When We Don't Know Why: Explanation, Theory, Linguistics, and How Questions Shape Them*, Chicago.

Burian, R.: 1980, 'Why Philosophers Should Not Despair of Understanding Discovery', in T. Nickles (ed.), 1980a.

Campbell, D. T.: 1974a, 'Evolutionary Epistemology', in P.A. Schilpp (ed.), *The Philosophy of Karl Popper*, vol. 1, Open Court, La Salle, pp. 413-63.

Campbell, D. T.: 1974b, 'Unjustified Variation and Selective Retention in Scientific Discovery', in F. Ayala and T. Dobzhansky (eds.), *Studies in the Philosophy of Biology*, Macmillan, London, pp. 139-161.

Cherniak, C.: 1986, *Minimal Rationality*, The MIT Press, Cambridge, Massachusetts.

Collingwood, R . G.: 1939, *An Autobiography*, Oxford University Press, Oxford.

Collingwood, R . G.: 1940, *Essay on Metaphysics*, Clarendon Press, Oxford.

Collins, H.: 1989, 'Computers and the Sociology of Scientific Knowledge', *Social Studies of Science* **19**, 613-624.

Curd, M.: 1980, 'The Logic of Discovery" in Scientific Discovery, Logic and Rationality', in T. Nickles (ed.), 1981a, pp. 201-220.

Darden, L.: 1990, 'Diagnosing and Fixing Faults in Theories', in Shrager and Langley 1990, pp. 319-346.

Darden, L.: 1991, *Theory Change in Science: Strategies from Mendelian Genetics*, Oxford University Press, New York.

Dear, P.: 1998, 'Method and the Study of Nature', in D. Garber and M. Ayers (eds.), *The Cambridge History of Seventeeth-Century Philosophy, Vol. 1*, Cambridge University Press, Cambridge.

Descartes, R.: [1637] 1968, *Discourse on Method*, translated by F.E. Sutcliffe, Penguin, Hadmonsworth.

Donovan, A., L. Laudan, and R. Laudan (eds.): 1988, *Scrutinizing Science: Empirical Studies of Scientific Change*, Kluwer, Dordrecht.

Dorling, J.: 1973, 'Demonstrative Induction: Its Significant Role in the History of Physics', *Philosophy of Science* **40**, 360-372.

Downes, S.: 1990, 'Herbert Simon's Computational Models of Scientific Discovery', *PSA 1990, Vol. 1*, 97-108.

Dunbar, K.: 1995, 'How scientists really reason: Scientific reasoning in real-world laboratories', in R. J. Sternberg and J. Davidson (eds.), *Mechanisms of insight*, MIT Press, Cambridge, MA, pp. 365-396.

Duncan, S. C. and R. D. Tweney: 1997, (Abstract). 'The Problem-Behavior Map as cognitive-historical analysis: The example of Michael Faraday', *Proceedings of the Nineteenth*

Annual Conference of the Cognitive Science Society, Lawrence Erlbaum Associates, Hillsdale, NJ, p. 901.

Ellis, B.: 1988, 'Solving the Problem of Induction Using a Values-Based Epistemology', *British Journal for the Philosophy of Science* **39**, 141-160.

Feigenbaum, E. A., B. Buchanan, and J. Lederberg: 1971, 'On generality and problem solving: a case study using the DENDRAL program', *Machine Intelligence* **7**, 165-90.

Feyerabend, P. K.: 1975, *Against Method*, New Left Books, London.

Finocchiaro, M.: 1980, 'Scientific Discoveries as Growth of Understanding', in T. Nickles (ed.) 1981a, pp. 235-256.

Fleck, L.: 1979, *Genesis and Development of a Scientific Fact*, translated by F. Bradley and T.J.Trenn, foreword by T. Kuhn, University of Chicago Press, Chicago.

Forbus, K. D.: 1985, 'Qualitative Process Theory', in D. G. Bobrow (ed.): *Qualitative Reasoning about Physical Systems*, MIT Press, Cambridge, Ma.

Gale, S.: 1978, 'A Prolegomenon to an Interrogative Theory of Scientific Inquiry', in H. Hiz (ed.), *Questions*, Dordrecht, Holland, D. Reidel.

Gamble, T.: 1983, 'The Natural Selection Model of Knowledge: Campbell's Dictum and its Critics', *Cognition and Brain Theory* **6**, 353-63.

Giere, R.: 1996, 'From Wissenschaftliche Philosophie to Philosophy of Science', in R. N. Giere and A. Richardson (eds.), *The Origins of Logical Empiricism*, Minnesota Studies in the Philosophy of Science, Vol. 16, University of Minnesota Press, Minneapolis.

Gingerich, O. (ed.): 1975, *The Nature of Scientific Discovery*, Smithsonian Institution Press, Washington.

Glymour, C.: 1980, *Theory and Evidence*, Princeton University Pess, Princeton, NJ.

Glymour, C.: 1985, 'Inductive Inference in the Limit', *Erkenntnis* **22**, 23-31

Glymour, C., R. Scheines, P. Spirtes, and K. Kelly: 1987, *Discovering Causal Structure: Artificial Intelligence, Philosophy of Science, and Statistical Modeling*, Academic Press, Orlando.

Gower, B.: 1997, *Scientific Method: An Historical and Philosophical Introduction*, Routledge, London and New York.

Griffiths, P. E. and R. D. Gray, 'Developmental Systems and Evolutionary Explanaton', *Journal of Philosophy* **XCI**, 277-304.

Gruber, H.: 1980, 'The Evolving Systems Approach to Creative Scientific Work: Charles Darwin's Early Thought', in T. Nickles (ed.), 1980b, pp. 113 - 130.

Gutting, G.: 1980a, 'The Logic of Invention', in T. Nickles (ed.) 1980a, pp. 221- 234.

Gutting, G.: 1980b, 'Science as Discovery', *Revue Internationale de Philosophie* **131** 32: 26-48.

Hamblin, C. L.: 1958, 'Questions', *Australasian Journal of Philosophy* **36**, 159-168.

Hanson, N. R.: 1958, *Patterns of Discovery*, Cambridge University Press, Cambridge.

Hanson, N. R.: 1961, 'Is there a logic of scientific discovery', in H. Feigl and G. Maxwell (eds.), *Current Issues in the Philosophy of Science*, Holt, Rinehart and Winston, Inc., New York.

Harré, R.: 1960, *An Introduction to the Logic of the Sciences*, St. Martins, New York.

Harré, R.: 1970, *The Principles of Scientific Thinking*, The University of Chicago Press, Chicago.

Hattiangadi, J. N.: 1978, 'The Structure of Problems', *Philosophy of the Social Sciences* **8**, 345-365, and **9**, 49-76.

Hattiangadi, J. N.: 1980, 'The Vanishing Context of Discovery', in Nickles, T (ed.) 1980a, pp. 257-266.

Hempel, C. G.: 1966, *Philosophy of Natural Science*, Prentice-Hall Englewood Cliffs, New Jersey.

Hempel, C. and P. Oppenheim: 1948, 'Studies in the Logic of Explanation', reprinted in C. Hempel, *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, The Free Press, New York, 1965.

Hesse, M.: 1963, *Models and Analogies in Science*, 2nd enlarged edition, Notre Dame University Press, Notre Dame.

Hesse, M.: 1964, 'Francis Bacon´s Philosophy of Science', in D. J. O´Connor (ed.), *A Critical History of Western Philosophy*, Free Press, New York.

Hesse, M.: 1970, "Hermeticism and Historiography", in R. Stuewer (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. 5, University of Minnesota, Minneapolis Press, 1970.

Hesse, M.: 1974, *The Structure of Scientific Inference*, The University of Minnesota Press, Minneapolis.

Hintikka, J.: 1976, *The Semantics of Questions and the Questions of Semantics*, Acta Philosophica Fennica, Vol. 28, No 4, North Holland.

Hintikka, J.: 1981a, 'On the Logic of an Interrogative Model of Scientific Inquiry', *Synthese* **47**, 69-83.

Hintikka, J.: 1981b 'The Logic of Information-Seeking Dialogues: A Model', in W. Becker and K. Essler (eds.), *Konzepte der Dialektik*, Frankfurt-am-Main, 1981, p. 212—231.

Hintikka, J.: 1984, 'The Logic of Science As a Model-Oriented Logic' in P.D. Asquith and P. Kitcher (eds.) *PSA 1984* **1** The Philosophy of Science Association, East Lansing, Michigan, pp. 177-185.

Hintikka, J.: 1985, 'True and False Logics of Scientific Discovery', *Communication and Cognition* **18** (1/2), 3-14.

Hintikka, J.: 1987, 'The Interrogative Approach to Inquiry and Probabilistic Inference', *Erkenntnis* **26**, 429-42.

Hintikka, J.: 1992, 'The Concept of Induction in the Light of the Interrogative Approach to Inquiry', in J. Earman (ed.), *Inference, Explanation and Other Frustrations*; University of California Press, Berkeley.

Hintikka, J.: 1999, *Inquiry as Inquiry: A Logic of Scientific Discovery*, Selected Papers 5, Kluwer Academic Publishers.

Hintikka, J. and U. Remes: 1974, *The Method of Analysis: Its Geometrical Origin and Its General Significance*, D. Reidel, Dordrecht.

Hintikka, J. and F. Vandamme (eds.): 1985, *The Logic of Discovery and the Logic of Discourse*, Plenum Press, New York.

Holland, J.: 1992, 'Genetic Algorithms', *Scientific American*, July issue, p. 66-72.

Holland, J.: 1995, *Hidden Order: How Adaptation Builds Complexity*, Addison-Wesley, Reading, Mass.

Holton, G.: 1974, 'Mainsprings of Scientific Discovery', in Gingerich (ed.), 1975.

Howson, C.: 1984, 'Bayesianism and Support by Novel Facts', *British Journal for the Philosophy of Science* **24**, 245-251.

Hoyningen-Huene, P.: 1987, 'Context of Discovery and Context of Justification', *Studies in History and Philosophy of Science* **18**, 501-515.

Hull, D. L.: 1988, *Science as a Process*, Chicago University Press, Chicago.

Ippolito, M. F. and R. D. Tweney: 1995, *The inception of insight*, in R.J. Sternberg & J.E. Davidson (eds.), *The Nature of Insight*, The MIT Press, Cambridge, MA, pp. 433-462.

Jacob, F.: 1977, 'Evolution and Tinkering', *Science* **196**, 1161-66.

Jardine, L.: 1974, *Francis Bacon: Discovery and the Art Discourse*, Cambridge University Press, New York.

Jardine, N.: 1991, *The Scenes of Inquiry: On the Reality of Questions in the Sciences*, Clarendon Press, Oxford.

Jason, G.: 1989, *The Logic of Scientific Discovery*, NY Lang.

Kant, I.: 1968, *Critique of Pure Reason*, translated by N. Kemp Smith, St Martin's Press, New York.

Kantorovitch, A. and Y. Ne'eman: 1989, 'Serendipity as a Source of Evolutionar Progress in Science.' *Studies in History and Philosophy of Science* **20**, 505-530

Kantorovich, A.: 1993, *Scientific Discovery: Logic and Tinkering*, State University of New York Press, Albany.

Kantorovich, A.: 1994, 'Scientific Discovery: A Philosophical Survey', *Philosophia* **1-4**, 3-23..

Kelly, K.: 1987, 'The Logic of Discovery', *Philosophy of Science* **54**, 435-452.

Kelly, K.: 1994, *Out of Control: The New Biology of Machines, Social Systems, and the Economic World*, Addison-Wesley, Reading, Mass.

Kelly, K.: 1996, *The Logic of Reliable Inquiry*, Oxford University Press, New York.

Kiikeri, M.: 1997, 'On the Logical Structure of Learning Models', *Poznan Studies in the Philosophy of the Sciences and the Humanities* **51**, 287-307.

Kitcher, P.: 1993, *The Advancement of Science*. New York, Oxford University Press, Oxford.

Kleiner, S. A.: 1970, 'Erotetic Logic and the Structure of Scientific Revolution', *British Journal for Philosophy of Science* **21**,149-165,

Kleiner, S. A.: 1983, 'A New Look at Kepler and Abductive Argument', *Studies in History and Philosophy of Science* **14**, 279-313.

Kleiner, S. A.: 1988a, 'The Logic of Discovery and Darwin's Pre-Malthusian Researches', *Biology and Philosophy* **3**, 293-315.

Kleiner, S. A.: 1988b, 'Erotetic Logic and Scientific Inquiry', *Synthese* **74**, 19-46.

Kleiner, S. A.: 1993, *The Logic of Discovery: A Theory of the Rationality of Scientific Research*, Kluwer Adademic Press, Boston MA.

Kleiner, S. A.: 1997, 'The Structure of Inquiry in Developmental Biology', in M. Sintonen (ed.), *Knowledge and Inquiry. Poznan Studies in the Philosophy of Science and the Humanities* **51**, Rodopi, Amsterdam-Atlanta.

Koertge, N.: 1980, 'Analysis as a Method of Discovery During the Scientific Revolution', in T. Nickles (ed.), 1980, pp.139-158.

Koertge, N.: 1982, 'Explaining Scientific Discovery', in *PSA 1982* **1**, 14-28.

Koestler, A.: 1960, *The Watershed*, Anchor Books, New York NY.

Kordig, C.: 1978, 'Discovery and Justification', *Philosophy of Science* **45**, 110-117.

Koza, J.: 1992, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, vol. 1 MIT Press, Cambridge, Mass..

Koza, J.: 1994, *Genetic Programming II: Automatic Discovery of Reusable Programs*, MIT Press, Cambridge, Mass.

Kuhn, T. S.: 1957, *The Copernican Revolution*, Random House, New York.

Kuhn, T. S.: 1970 *The Structure of Scientific Revolutions*, 2nd edition, Chicago.

Kulkarni, D. and H. A. Simon: 1988, 'The Processes of Scientific Discovery: The Strategy of Experimentation', *Cognitive Science* **12**, 139-175.

Lakatos, I.: 1970, 'Falsification and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (eds.), *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge.

Lakatos, I.: 1976, *Proofs and Refutations*, Cambridge University Press, Cambridge.

Langley, P., H. A. Simon, G. Bradshaw, and J. Zytkow: 1987, *Scientific Discovery*, MIT Press, Cambridge, Mass.

Laudan, L.: 1970, 'Thomas Reid and the Newtonian Turn of British Methodological Thought', in R. E. Butts and J. W. Davis (eds.), *The Methodological Heritage of Newton*, Basil Blackwell, Oxford, pp. 103-131.

Laudan, L.: 1977, *Progress and Its Problems: Towards a Theory of Scientific Growth*, London and Henley.

Laudan, L.: 1980, 'Why Was the Logic of Scientific Discovery Abandoned?', in T. Nickles (ed.) 1980a, pp. 173-184.

Laudan, L.: 1981, *Science and Hypothesis*, D. Reidel, Dordrecht.

Laudan, L.: 1983, 'Invention and Appraisal' (reply to McLaughlin), *Philosophy of Science* **50**, 320-322.

Laudan, L.: 1984, *Science and Values*, University of California Press, Berkeley.

Laymon, R.: 1978, 'Newton's Experimentum Crucis and the Logic of Idealization Theory Refutation', *Studies in History and Philosophy of Science* **9**, 51-77.

Laymon, R.: 1994, 'Demonstrative Induction, Old and New Evidence and the Accuracy of the Electrostatic Inverse Square Law', *Synthese* **99**, 23-58.

Lenat, D.: 1977, 'The Ubiquity of Discovery', *Artificial Intelligence* **9**, 257-285.

Leplin, J.: 1980, 'The Role of Models in Theory Construction' in T. Nickles (ed.) 1980a, pp. 267-284.

Lycan, W. G.: 1985, 'Epistemic Value', *Synthese* **64**, 137-164.

Magnani, L., N. Nercessian, and P. Thagard (eds.): 1999, *Model-Based Reasoning in Scientific Discovery*, Kluwer, Dordrecht.

McLaughlin, R.: 1982a, 'Invention and Appraisal', in R. McLaughlin (ed.), *What? Where? When? Why?*, D. Reidel, Dordrecht.

McLaughlin, R.: 1982b, 'Invention and Induction: Laudan, Simon, and the Logic of Discovery', *Philosophy of Science* **49**, 198-211.

Meheus, J.: 1993, 'Adaptive Logic in Scientific Discovery: The Case of Clausius', *Logique et Analyse* **143-144**, 359-391.

Meheus, J.: 1999, 'Deductive and Ampliative Adaptive Logics as Tools in the Study of Creativity', *Foundations of Science* **4**, 325-336.

Meyer, M.: 1979, *Dicouverte et Justification en Science*, Editions Klincksieck, Paris.

Meyer, M.: 1980, 'Science as a Questioning Process: A Prospect for a New Type of Rationality', *Revue Internationale de Philosophie*, 1980, 49 - 89.

Meyer, M.: 1994, *Rhetoric, Language and Reason*. University Park, Pennsylvania: The Pennsylvania State University Press.

Miller, A. I.: 1986, *Imagery in Scientific Thought*, MIT Press, Cambridge, MA.

Monk, R.: 1977, 'The Logic of Discovery', *Philosophy Research Archives* **3**, 1-51.

Monk, R.: 1980, 'Productive Reasoning and the Structure of Scientific Research', in T. Nickles (ed.), 1980a, 337-354.

Musgrave, A.: 1974, 'Logical versus Historical Theories of Confirmation', *British Journal for the Philosophy of Science* **25**, 1-23.

Mäenpää, P.: 1993, *The Art of Analysis: Logic and History of Problem Solving*, Academic Dissertation, Limes, Helsinki.

Nersessian, N. J.: 1985,'Faraday's field concept,' in D. Gooding and F. James (eds.), *Faraday Rediscovered: Essays on the Life and World of Michael Faraday, 1791-1867*, Macmillan, London, pp. 175-18.

Nersessian, N. J.: 1984, *Faraday to Einstein: Constructing the Meaning in Scientific Theorie*, Martinus Nijhoff, Dordrecht.

Nercessian, N.: 1987a, 'A Cognitive-Historical Approach to Meaning in Scientific Theories', in N. J. Nercessian 1987b.

Nercessian, N. (ed.): 1987b, *The Process of Science*, Martinus Nijhoff Publishers, Dordrecht.

Nersessian, N.: 1992, 'How Do Scientists Think? Capturing the Dynamics of Conceptual Change in Science', in R. Giere (ed.): *Minnesota Studies in the Philosophy of Science,* *Vol. XV: Cognitive Models of Science*, University of Minneapolis Press, Minneapolis, pp. 3-44.

Nersessian, N.: 1993, 'Opening the black box: Cognitive science and history of science', *Osiris* **10**, 194-214.

Newell, A. and H. A. Simon: 1972, *Human Problem Solving*, Prentice Hall, Englewood Cliffs, NJ.

Newell, A. and H. A. Simon: 1976: 'Computer Science as Empirical Enquiry: Symbols and Search', *Communications of the ACM* **19**.

Nickles, T. (ed.): 1980a, *Scientific Discovery, Logic, and Rationality*, D. Reidel, Dordrecht.

Nickles, T.: 1980b, *Scientific Discovery: Case Studies*, D. Reidel, Dordrecht.

Nickles, T.: 1980c, 'Scientific Discovery and the Future of Philosophy of Science', in T. Nickles (ed.), 1980a, pp. 1-62.

Nickles, T.: 1981, 'What is a Problem That We May Solve It?', *Synthese* **47**, pp. 85-118.

Nickles, T.: 1984, 'Positive Science and Discoverability', in *PSA 1984, Vol. 1*, Philosophy of Science Association, East Lansing, pp. 13-27.

Nickles, T.: 1985, 'Beyond Divorce: Current Status of the Discovery Debate', *Philosophy of Science* **52**, 177-206.

Nickles, T.: 1987, 'Lakatosian Heuristics and Epistemic Support', *British Journal for the Philosophy of Science* **38**, 181-205.

Nickles, T.: 1988, 'Truth or Consequences? Generative Versus Consequential Justification in Science', in A. Fine and J. Leplin (eds.), *PSA 1988, Vol. 2.*, Philosophy of Science Association, Lansing, pp. 393-405.

Nickles, T.: 1990, 'Discovery logics', *Philosophica* **45**, 7-32.

Nickles, T.: 1992a, 'Good Science as Bad History: From Order of Knowing to Order of Being', in E. McMullin (ed.), *The Social Dimensions of Science*, University of Notre Dame Press, Notre Dame, pp. 85-129.

Nickles, T.: 1992, *Epistemic Amplification: Toward a Bootstrap Methodology of Science*, in J. Brzezinski, F. Coniglione, and T. Marek (eds.), *Science: Between Algorithm and Creativity*, Eburon, Delft.

Nickles, T.: 1995, 'History of Science and Philosophy of Science', *Osiris* **10**, 139-163.

Nickles, T.: 1997 'Methods of Discovery' *Biology and Philosophy* **12**, 127-140.

Niiniluoto, I.: 1980, 'The Growth of Theories: Comments on the Structuralist Approach', in *Proceedings of the Second International Congress for History and Philosophy of Science, Pisa, 1978*, Dordrecht, pp. 3-47.

Niiniluoto, I.: 1980b, 'Scientific Progress', *Synthese* **45**, 427-462.

Norton, J.: 1994, 'Science and Certainty', *Synthese* **99**, 3-22.

Norton, J.: 1995, 'Eliminative Induction as a Method of Discovery: How Einstein Discovered General Relativity', in Leplin 1995, pp. 29-70.

Peirce, C.: 1931-35, *Collected Papers of Charles Sanders Peirce*, vols. I-VI, reprinted by the Belnap Press of Harvard University, Cambridge, MA, 1965.

Peirce, C.: 1931-58, *Collected Papers*, Harvard University Press, Cambridge, MA.

Pera, M.: 1981, 'Inductive Method and Scientific Discovery', in M. Grmek, R. S. Cohen, and G. Cimino (eds.), *On Scientific Discovery*, D. Reidel, Dordrecht.

Polanyi, M.: 1958, *Personal Knowledge*, University of Chicago Press, Chicago.

Polya, G.: 1945, *How to Solve It*, Princeton University Press, Princeton.

Popper, K. R.: 1959, *The Logic of Scientific Discovery*, Basic Books, New York; trans. with revisions of *Logik der Forschung*, 1934.

Popper, K. R.: 1972, *Objective Knowledge*, Oxford.

Reichenbach, H.: 1938, *Experience and Prediction*, The University of Chicago Press, Chicago.

Reitman, W.: 1964, 'Heuristics, Decision Procedures, Open Constraints, and the Structure of Ill-defined Problems', in M. W. Shelly and G. L. Bryan (eds.), *Human Judgments and Optimality*, John Wiley, New York, pp. 282-315.

Rescher, N.: 1990, 'Luck', *Proceedings of the American Philosophical Association* **64**, 5-19.

Rosenkrantz, R. D.: 1977, *Inference, Method and Decision: Towards a Bayesian Philosophy of Science*, D. Reidel, Dordrecht.

Ruse, M.: 1985, 'Evolutionary Epistemology: Can Sociobiology Help?', in J. H. Fetzer (ed.), *Sociobiology and Epistemology*, D. Reidel, Dordrecth, pp. 249-265.

Salmon, W.: 1966, *The Foundations of Scientific Inference*, Pittsburgh University Press, Pittsburgh.

Salmon, W.: 1970, 'Bayes's Theorem and the History of Science', in R. Stuewer (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. 5, University of Minnesota, Minneapolis Press, 1970, pp. 68-86.

Schaffer, S.: 1986, 'Scientific Discoveries and the End of Natural Philosophy', *Social Studies of Science* **16**, 387-420.

Schaffer, S.: 1990, 'Genius in Romantic Natural Philosophy', in A. Cunningham and N. Jardine (eds.): *Romanticism and the Sciences*, Cambridge University Press, Cambridge.

Schaffer, S.: 1994, 'Making Up Discovery', in M. Boden (ed.), *Dimensions of Creativity*, MIT Press, Cambridge, Mass.

Schaffner, K.: 1985, *Logic of Discovery and Diagnosis in Medicine*, University of California Press, Berkeley.

Schlesinger, G. N.: 1987, 'Accommodation and Prediction', *Australasian Journal of Philosophy* **65**, 33-42.

Schlick, M.: 1974, *General Theory of Knowledge*, second edition, translated by A. E. Blumberg. Springer-Verlag, Wien, New York; originally published in 1925 as *Allgemeine Erkenntnislehre.*

Shrager, J. and P. Langley: 1990, *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufmann, San Mateo.

Shapere, D.: 1977, 'Scientific Theories and Their Domains', in F. Suppe (ed.), *The Structure of Scientific Theories*, 2nd Edition, University of Illinois Press, Urbana, Chicago, London.

Shapere, D.: 1980, 'The Character of Scientific Change" in Scientific Discovery, in T. Nickles, 1980a, pp. 63-101.

Shapere, D.: 1982, 'The Concept of Observation in Science and Philosophy', *Philosophy of Science* **49**, 485-525.

Shapin, S.: 1992, 'Discipline and Bounding: The History and Sociology of Science as Seen through the Externalism-Internalism Debate', *History of Science*, 333-369.

Shrager, J. and P. Langley: 1990, *Computational Models of Scientific Discovery and Theory Formation*, Morgan Kaufmann, San Mateo.

Siegel, H.: 1980, 'Justification, Discovery, and the Naturalizing of Epistemology', *Philosophy of Science* **47**, 297-321.

Simon, H. A, 1973, 'Does Scientific Discovery Have a Logic?', *Philosophy of Science* **40**, 471-480.

Simon, H. A.: 1977, *Models of Discovery*, D. Reidel, Dordrecht.

Simon, H. A.: 1996, *The Sciences of the Artificial*, 3rd ed., MIT Press, Cambridge; originally published in 1969.

Simon, H. A.: 1992, 'Scientific Discovery as Problem Solving', *International Studies in the Philosophy of Science* **6**, 3-14.

Simon, H. A. and A. Ericsson: 1984, *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, MA.

Simon, H. A., P. W. Langley, and G. L. Bradshaw: 1981, 'Scientific Discovery as Problem Solving', *Synthese* **47**, 1-28.

Simonton, D. K.: 1988, *Scientific genius: A psychology of science*, Cambridge University Press, Cambridge.

Sintonen, M.: 1984, 'On the Logic of Why-Questions', in P. D. Asquith and P. Kitcher (eds.), *PSA 1984, Volume One*, The Philosophy of Science Association, East Lansing, Michigan, pp. 168-176.

Sintonen, M.: 1985, 'Separating Problems from their Backgrounds: A Question-Theoretic Proposal', *Communication and Cognition* **18**, No ½, pp. 25-49.

Sintonen, M.: 1989, 'Explanation: In Search of the Rationale', in P. Kitcher and W. C. Salmon (eds.), *Scientific Explanation*, Minnesota Studies in the Philosophy of Science, University of Minnesota Press, Minneapolis.

Sintonen, M.: 1990, 'How to Put Questions to Nature', in D. Knowles (ed.), *Explanation and Its Limits*, Cambridge University Press, Cambridge, pp. 267-284

Sintonen, M.: 1993, 'In Search of Explanations: From Why-Questions to Shakespearean Questions', *Philosophica* **51**, 55-81.

Sintonen, M.: 1996, '"Structuralism and the Interrogative Model of Inquiry', in W. Balzer and C. Ulises-Moulines (eds.), *Structuralist Theory of Science*, Walter de Gruyter, Berlin, New York, pp. 45-74.

Snyder, L.: 1997, 'Discoverers' Induction', *Philosophy of Science* **64**, 580-604.

Sober, E.: 1981, 'The Evolution of Rationality', *Synthese* **46**, 95-120.

Solomon, R.: 1981, *Introducing the German Idealists*, Hackett, Indianapolis.

Stein, H.: 1991, '"From the Phenomena of Motions to the Forces of Nature": Hypothesis or Deduction?', in *PSA 1990*, Vol. 2, pp. 209-222.

Sternberg, R. J. and J. Davidson (eds.): 1995, Mechanisms of Insight, MIT Press, Cambridge, MA.

Szabo, A. K.: 1974, 'Working Backwards and Proving by Synthesis', Appendix I, Hintikka and Remes 1974.

Thagard, P.: 1992, *Conceptual Revolutions*, Princeton University Press, Princeton.

Toulmin, S.: 1972, *Human Understanding*, Princeton University Press, Princeton.

Tursman, R.: 1987, *Peirce's Theory of Scientific Discovery. A System of Logic Conceived as Semiotic*, Indiana University Press, Bloomingston and Indiana.

Tweney, R. D.: 1989, 'A framework for the cognitive psychology of science', in B. Gholson, A. Houts, R. A. Neimeyer, and W. Shadish (eds.), *Psychology of Science and Metascience*, Cambridge University Press, Cambridge, pp. 342-366.

Tweney, R. D.: 1996, 'Presymbolic processes in scientific creativity', *Creativity Research Journal* **9**, 163-172.

Urbach, P.: 1987, *Francis Bacon's Philosophy of Science: An Account and a Reappraisal*, Open Court, La Salle, Illinois.

Wallace, A. R.: 1905, *My Life, A Record of Events and Opinions*, London.

Whewell, W.: 1847, *Philosophy of the Inductive Sciences*, 2nd ed., 2 vols., London.

Whitt, L. A.: 1990, 'Theory Pursuit: Between Discovery and Acceptance', *PSA 1990*, Vol. 1, The Philosophy of Science Association, East Lansing, Michigan, pp. 467-483.

Williams, B.: 1976, 'Moral Luck', reprinted in his *Moral Luck*, Cambridge University Press, Cambridge, 1981.

Winston, P. H.: 1992. *Artificial Intelligence*, 3rd ed., Addison-Wesley Reading, Mass.

Wilson, L. G. (ed.): 1970, *Sir Charles Lyell's Scientific Journals on the Species Question*, New Haven, Connecticutt.

Wisńiewski, A.: 1994, 'Erotetic Implications', *Journal of Philosophical Logic* **23**, 173-195.

Wisńiewski, A.: 1995, *The Posing of Questions: Logical Foundations of erotetic Inferences*, Kluwer Academic Publishers, Dordrechtl/ Boston/ London.

Wisńiewski, A., 'The logic of questions as a theory of erotetic arguments', *Synthese* **109**, 1-25.

Woolgar, S.: 1976, 'Writing an Intellectual History of Scientific Development: The Use of Discovery Accounts', *Social Studies of Science* **6**, 395-422.

Woolgar, S.: 1988, *Science: The Very Idea*, Tavistock, London.

Worrall, J.: 2000, 'The Scope, Limits, and Distinctiveness of the Method of "Deduction from the Phenomena": Some Lessons from Newton's "Demonstrations" in *Opticks*', *British Journal for the Philosophy of Science* **51**, 45-80.

Zahar, E.: 1983, 'Logic of Discovery or Psychology of Invention?', *British Journal for the Philosophy of Science* **34**, 243-262.

Zytkow, J.: 1993, 'Cognitive Autonomy in Machine Discovery', *Machine Learning* **12**, 7-12.

SVEN OVE HANSSON

BELIEF REVISION FROM AN EPISTEMOLOGICAL POINT OF VIEW

I. INTRODUCTION

Beginning in the 1970's, a more focused discussion of the requirements of rational belief change has taken place in the philosophical community. Work by Isaac Levi (1977, 1980) and William Harper (1977) has been particularly important. In 1985, Carlos Alchourrón, Peter Gärdenfors, and David Makinson presented a formal model of belief change that is now known as the AGM model of belief change. Their work set a new standard for formal precision in this area. A rapid development of formal models of belief change has taken place in the 1990's, but it has not been matched by a sufficient amount of philosophical reflection on these models or on their relationship to issues in non-formalized epistemology.

The purpose of this review is to highlight the epistemological issues that arise in the interpretation of formal models of belief change, and to begin a discussion of how such models can be used as tools for epistemological analysis. This will be done with a bare minimum of formal apparatus. (The reader interested in formal matters is referred to Gärdenfors 1988, Gärdenfors and Rott 1995, Hansson 1998 or Hansson 1999*a* for an overview.)

To begin with, we are going to have a close look at the idealizations that are conventionally made in the representation of belief states (Sections 2–3) and of operators of change (Sections 4–5). Section 6 provides a brief introduction to the AGM model and Section 7 a discussion of the most contested property of that model, namely its recovery postulate. After that, two generalizations are discussed, namely belief bases (Sections 8–9) and non-prioritized belief revision (Section 10). The final Section 11 is devoted to a discussion of how belief change theory can be made useful in epistemological analysis.

2. REPRESENTING THE BELIEF STATE

Actual processes of belief change are multifarious and often quite complex. In order to obtain a model that is at all manageable, substantial simplifications are necessary. In other words, a useful formal model of belief change has to be idealized in the sense that certain aspects of real-world belief changes have been omitted in order to make others come out more clearly. Furthermore, the models available in the literature are also idealized in another sense of that word: they represent standards of rationality that are higher than what actual persons live up to.

It has not been sufficiently clarified in what ways the ideal rational agents of belief change theory differ from ordinary human beings. Some researchers seem to conceive these ideal agents as having unlimited cognitive capacity. It is then fairly

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 255-279.
© 2004 *Kluwer Academic Publishers.*

unproblematic to construct formal models in which these agents have to process infinite entities (such as infinite sets of sentences). An alternative view is that the ideal agents of belief change theory have limited cognitive capacity, which they make rational use of. On that view, finiteness and computability are important *desiderata* of formal models.

In a model of change there must be something that changes. In belief change theory, that object of change is the *belief state*. The very idea of a belief state is in itself an idealization. It artificially isolates beliefs from other constituents of a state of mind such as emotions and preferences.

The available belief change models are *sentential*: beliefs are represented by sentences. This, too, is clearly an idealization. Actual beliefs do not necessarily have the structure of sentences in a language. However, although sentences do not capture all aspects of beliefs, they provide the best available general-purpose representation of beliefs.

In most belief change models, the belief-representing sentences are assumed to be elements of a simple, truth-functional propositional language. There seems to be a consensus among workers in this field that the addition of quantifiers to the belief-representing language would not provide new insights in proportion to the complications that would ensue. The inclusion of modal or conditional sentences in the language gives rise to interesting puzzles, but also seems to make further formal developments difficult. (Gärdenfors 1986; Hansson 1992 and 1995*b*; Levi 1988; Lindström and Rabinowicz 1998.)

The notion of belief can be conceived as an all-or-nothing concept: either you believe something, or you do not. Alternatively, it may be thought of as admitting of degrees: you may believe something to various degrees. Mainstream belief change models are *dichotomous*: they divide the sentences of the language into two distinct categories: those representing beliefs and those not doing so. (On non-dichotomous models that allow for degrees of belief, see Spohn 1988, Dubois et al. 1998, and Smets 1998.)

This dichotomy is not uncontroversial. According to the Bayesian ideal of rationality, a rational subject should assign a definite probability value to each statement about the world. Only logically true sentences are assigned probability one. Non-logical propositions can, at most, be assigned high probabilities that are marginally lower than 1. The resulting belief system is a complex web of interconnected probability statements. (Jeffrey 1956) In practice, however, such a belief system would be unmanageable for human subjects. (McLaughlin 1970) Our cognitive limitations are so severe that massive reductions from high probability to full belief (certainty) are inevitable in order to make us capable of reaching conclusions and making decisions. This reduction to full belief, or 'fixation of belief' (Peirce 1934) helps us to achieve a cognitively manageable representation of the world.

The prevalence of this reduction (fixation) process is one of the reasons why dichotomous belief models represent some features of doxastic behaviour (notably those related to logic) more realistically than probabilistic models. Clearly, there are other features that can be more realistically represented in the latter models. Note, however, that this argument for dichotomous models refers to their relevance for agents with limited cognitive capacities. The use of dichotomous models in a discussion of (ideal) agents with unlimited cognitive capacities does not seem to be equally well motivated.

In what follows we will be concerned with dichotomous, sentential models of belief. In such models, the relation between belief states and believed sentences can be expressed with a *support function* that sorts out the sentences that are supported by the belief state. (Hansson 1992) Let $\mathcal{K}$ be a belief state. Then $s(\mathcal{K})$ is the set of all sentences that are beliefs in $\mathcal{K}$.

It is commonly assumed in belief change theory that logical consequences of beliefs are themselves beliefs, i.e. that $s(\mathcal{K})$ is *closed under logical consequence* ($s(\mathcal{K}) = \mathrm{Cn}(s(\mathcal{K}))$), where Cn is an operator of logical consequence). This is not a realistic assumption, but it has turned out to be extremely helpful as a means to obtain a manageable formal structure. An interesting argument in its favour was put forward by Isaac Levi (1977 and 1991); according to him, $s(\mathcal{K})$ should be interpreted as consisting of the sentences that someone is *committed to believe*, not those that she actually believes in.

A set, such as $s(\mathcal{K})$, that is closed under logical consequence is called a "theory" in logical parlance. In belief change theory, it is called a "corpus", "knowledge set", or (more commonly) "belief set".

The simplest and most obvious representation of belief states is to identify each belief state with its respective belief set, so that $s(\mathcal{K}) = \mathcal{K}$. Then operations of change are performed on the belief set, rather than on some underlying belief state from which it can be derived. Belief sets are usually denoted by a boldface **K**. (**K** originally stands for 'knowledge', thus bearing witness to the unfortunate habit of many authors in this field to use 'knowledge' as a synonym for 'belief'.)

In the terminology introduced by the Artificial Intelligence researcher Allen Newell (1982), belief sets have their place on the so-called *knowledge level*. ("Belief level" would have been a more accurate term.) In the traditional hierarchy of system levels, beginning with the device level and the circuit level, the knowledge level is positioned immediately above the symbol level (program level). The knowledge level is specified entirely in terms of the contents of the knowledge (beliefs). There is no distinction on this level between information that is explicitly available and information that is implied by available information. (Brachman 1986) It should be possible to predict and understand what an agent does on the knowledge level, without referring to the symbol level, in much the same way as the symbol level should allow for prediction and understanding without reference to the lower levels of the system. The existence of a knowledge level is closely related to Mukesh Dalal's principle of *irrelevance of syntax* for databases, according to which the outcome of an operation that changes a database should be independent of the syntax (representation) of either the old or the new information. (Dalal 1988; Katsuno and Mendelzon 1989) Irrelevance of syntax is a useful tool to describe an important form of idealization in a model of belief change. However, irrelevance of syntax clearly cannot be a reasonable criterion of rationality; that would amount to requiring that the language not be used to convey as much information as possible.

3. DYNAMIC INFORMATION IN THE BELIEF STATE

For dynamic purposes, a belief set is not a sufficient description of a belief state. We need to know not only the beliefs presently endorsed, but also what will be the fate of these beliefs after various operations of change have been performed. New information often requires that we give up previous beliefs, and we then have to choose how to curtail the previous belief set in order to accommodate the new information.

There are two major ways to provide this dynamic information. One of these is to introduce a representation of the *vulnerability* of the elements of the original belief set, indicating how easily different beliefs are given up. The idea is, of course, that when choosing which previous beliefs to give up, less vulnerable ones are retained as far as possible. Vulnerability is independent of the particular operations to be performed (and hence of the input sentence). Suppose that there are two operations O_1 and O_2, and two sentences α and β that are candidates for being retracted in both of these operations. Then, according to a vulnerability approach, α is more vulnerable than β in operation O_1 if and only if it is so in operation O_2. (For concreteness, we may think of O_1 as the retraction of the belief $\alpha\&\beta$ and O_2 as the retraction of the belief $\alpha\&\beta\&\delta$.)

Vulnerability is the predominant approach in AGM-style belief revision. Various formal representations of vulnerability have been constructed, some of which will be introduced in Section 6.

The other type of dynamic information relates to the *justificatory structure* of the belief set. Some beliefs have no independent standing, but are held only because they are justified by some other belief(s). When the justification of a belief has been lost, that belief should arguably also be deleted. There are two major ways to express information about justificatory structure. The simplest of these employs *belief bases*. A belief base is a set of sentences that is not (except as a limiting case) closed under logical consequence. Its elements represent beliefs that are held independently of any other belief or set of beliefs. The logical closure of a belief base is a belief set. Those elements of the belief set that are not in the belief base are "merely derived", i.e., they have no independent standing. (Hansson 1994) Changes are performed on the belief base, and derived beliefs are changed only as a result of changes of the base. (See Sections 8–9.)

More precise justificatory information is contained in what has been called *track-keeping* representations. (Hansson 1991*a*) Here, to each sentence is appended a list of its justifications or origins. This approach has been much explored by computer scientists, beginning with the "truth maintenance systems" (reason maintenance systems) developed by Jon Doyle (1979).

The relation between vulnerability and justificatory structure remains to investigate. It is not clear, either on a conceptual or a technical level, to what degree the justificatory structure can be expressed in terms of vulnerability, or vice versa.

4. MODELLING CHANGE

Given a formal representation of the belief state, let us now consider how changes in that state can be expressed in the formal framework. In what may be called *time-indexed* models, a (discrete or continuous) variable is employed to represent time. The object of change (such as a state of affairs, state of the world, or belief state) can then be represented as a function of this time variable. (This framework can also be made indeterministic by allowing for a bundle of functions, typically structured as a branching tree.)

A quite different mode of representation is that of *input-assimilating* models. In such models, the object of change is exposed to an input, and is changed as a result of this. No explicit representation of time is included. Instead, the characteristic mathematical constituent is a function that, to each pair of a state and an input, assigns a new state. In a well-constructed input-assimilating model of belief change, the representation of a belief state after a change has taken place should have the same format as the representation of the belief state before the change. This has been called the principle of *categorical matching*. (Gärdenfors and Rott 1995) As an example, if we begin with a belief base, the outcome of a change should be a new belief base, not a belief set. Similarly, if the original belief state is a belief set combined with information about the vulnerability of its element, the new belief state after change should contain the some two constituents (and not, e.g., be a belief set with no accompanying vulnerability information).

Input-assimilating models have the advantage of focusing on the causes and mechanisms of change. They exhibit the effects of external causes on systems that change only in response to such external influences ("inputs") and are otherwise stable. This makes them tolerably well suited to represent important aspects of changes in human states of mind, and of compartments of mind such as states of belief. At least for some purposes, it is a reasonable idealization to disregard such changes in a person's beliefs that have no direct external causes, in order to focus better on the mechanisms of externally caused changes.

The major models of belief change are all input-assimilating. Furthermore, they are *deterministic* in the sense that given a belief state and an input, the next belief state is well-determined. (On *indeterministic* belief change see Lindström and Rabinowicz 1989, Doyle 1991 and Gallier 1992.)

In the presence of conflicting information, selections are necessary. We have a choice between (1) making these selections as part of the operations of change when new information is received, and (2) letting operations of change leave conflicts unresolved, and instead make the necessary selections when information is recovered from the system. (Rott 1999) There is a trade-off in simplicity between retrieval and change. In the AGM model, the retrieval operation is as simple as possible – it is just the identity operation. The change operations of AGM are more complex. In other models, with a more complex retrieval operation, simpler operations of change may be sufficient. The relations between retrieval and change remain to be investigated, both from a formal and a more philosophical point of view. We do not know whether these approaches are fundamentally different or one of them can in some way be reduced to the other.

5. WHAT TYPES OF BELIEF CHANGES ARE THERE?

In the AGM framework, there are three types of belief change. (The basic ideas derive from earlier work by Isaac Levi, 1977 and 1980.) In *contraction,* a specified (belief-representing) sentence is removed from the belief set, without anything else being added to it. Hence, the outcome of contracting a belief set by a non-tautologous sentence α is a new belief set, which does not contain α. By *expansion* is meant that a specified sentence is set-theoretically added to the belief set (without anything else being excluded), and this expanded set is then closed under logical consequence. By *revision* is meant that a specified sentence is added to the belief set, under the condition that the new belief set be consistent and closed under logical consequence. (Alchourrón et al. 1985) (The word 'revision' is commonly used both for this specific type of operation, and as a synonym of change; this whole field of research is often called 'theory revision' or 'belief revision'.)

It should be noted that these three operations are all sentential: The inputs of contraction, expansion, and revision are taken to be sentences in the belief-representing formal language. This is by no means unproblematic. Actual epistemic agents are moved to change their beliefs largely by non-linguistic inputs, such as sensory impressions. Sentential models of belief change (tacitly) assume that all inputs can, in terms of their effects on belief states, be adequately represented by sentences. When I see a hen on the roof (a sensory input), I am assumed to adjust my belief state *as if* I modified it to include the sentence 'There is a hen on the roof' (a linguistic input). (Hansson 1995*a*)

It is a fundamental assumption in belief dynamics – introduced by Isaac Levi (1977) – that complex changes can be analyzed as sequences of changes of these three simple types:

> *Decomposition principle* (Fuhrmann 1989)
> Every legitimate belief change is decomposable into a sequence of contractions, expansions, and revisions.

(As we will see in Section 6, revision is in its turn constructed out of expansion and contraction; hence the decomposition principle can be reformulated without mention of revisions.)

The decomposition principle need not be read as a requirement that you actually change your beliefs in this stepwise fashion: one expansion, revision, or contraction at a time. All that is required is that the outcomes of complex changes are the same *as if* you had performed them in this way.

Even if we accept the basic underlying idealizations, this typology of change operations is open to criticism of at least two kinds. First, the realism and relevance of the three proposed types of operations can be questioned. Secondly, it may be argued that these three types of operations (or their combinations in sequences) do not cover all the types of belief changes that there are.

The first type of criticism has been directed primarily at the operation of contraction. (The realism of expansion or revision does not seem to have been seriously contested.) In contraction, as conventionally defined, the outcome is a subset of the original belief set, that does not contain the input sentence. Hence, this is an operation in which old beliefs are deleted but no new beliefs are added. It is difficult, however, to find examples of such *pure* contraction, in which no new belief is added. When we give up a belief, this is typically because we have learnt something new that forces the old belief out. For concreteness, suppose that I previously believed that the dinosaurs died out due to sudden climatic change (α). Then a geologist told me that this is only one out of several competing hypotheses. This makes me give up my belief in α (without starting to believe in its negation). Strictly speaking, this is not a case of (pure) contraction, since a new belief was acquired to the effect that there are several competing scientific hypotheses on the extinction of the dinosaurs. In the literature on belief dynamics, examples such as this are often interpreted as referring to (pure) contraction. The new belief that gave rise to contraction is neglected, and is not included in the new belief set. This is an imprecise but convenient convention, that makes it much easier to find examples of contraction.

We sometimes hypothetically give up a belief in order to give a contradictory belief a hearing. Such hypothetical contractions, or contractions for the sake of argument have sometimes been taken to be pure contractions. (Levi 1991; Fuhrmann 1991; Fuhrmann and Hansson 1994) However, their value as examples is controversial, since these contractions are not seriously undertaken by the agent.

The evasiveness of pure contraction should not lead us to believe that contraction is unimportant. Contraction is an essential element of rational belief change. It typically occurs as a part of more complex changes that involve both losses and acquisitions of information. For the formal analysis, it is useful to develop models of pure contraction, i.e., contraction that is not accompanied by any incorporation of new beliefs. In order to guide this development, intuitive examples of pure contraction may be helpful. Since the contractive parts of complex belief changes cannot be perfectly isolated, a considerable amount of idealization is necessary.

The other line of attack is that other types of change than the three mentioned should be allowed for. Several additional categories of operations have been proposed. In *multiple contraction* and *multiple revision*, the input consists of sets of sentences rather than single sentences. (Fuhrmann, 1988; Fuhrmann and Hansson 1994; Hansson 1989; Niederée 1991; Li 1998) In *updating*, the change takes part in the real world, rather than in the agent's beliefs about an unchanging world.(Katsuno and Mendelzon 1992; Keller and Winslett 1985) Abhaya Nayak (1994) has proposed a variant of revision in which the input contains not only a sentence to be incorporated into the belief set, but also its degree of priority in the resulting new belief set. (Hence, in his model there are different ways to revise one and the same belief state by one and the same sentence. See also Spohn 1988.) Operations of *non-prioritized revision* differ from those of conventional revision in that the input sentence is not always accepted. (See Section 10.) Operations of *consolidation* enhance the integrity of the belief state by making it consistent (Hansson 1994 and 1997) or coherent (Olsson 1997*a* and 1997*b*).

6. THE AGM MODEL

The purpose of this section is to briefly introduce the formal structure of AGM theory.

Sentences are denoted by lower-case Greek letters and sets of sentences by capital letters. To express the logic, a Tarskian consequence operator is used:

Definition 1 (Tarski 1956): *A consequence operation* on a language $\mathcal{L}$ is a function Cn that takes each subset of $\mathcal{L}$ to another subset of $\mathcal{L}$, such that:

(i) $A \subseteq \mathrm{Cn}(A)$ (*inclusion*)

(ii) If $A \subseteq B$, then $\mathrm{Cn}(A) \subseteq \mathrm{Cn}(B)$ (*monotony*)

(iii) $\mathrm{Cn}(A) = \mathrm{Cn}(\mathrm{Cn}(A))$ (*iteration*)

Postulate 2: Cn satisfies the following three properties:

(iv) If α can be derived from A by classical truth-functional logic, then $\alpha \in \mathrm{Cn}(A)$. (*supraclassicality*)

(v) $\beta \in \mathrm{Cn}(A \cup \{\alpha\})$ if and only if $(\alpha \rightarrow \beta) \in \mathrm{Cn}(A)$. (*deduction*)

(vi) If $\alpha \in \mathrm{Cn}(A)$, then $\alpha \in \mathrm{Cn}(A')$ for some finite subset $A' \subseteq A$. (*compactness*)

$X \vdash \alpha$ is an alternative notation for $\alpha \in \mathrm{Cn}(X)$, and X α for $\alpha \notin \mathrm{Cn}(X)$. The language is assumed to contain the usual truth-functional connectives: negation ($\neg$), conjunction (&), disjunction ($\vee$), implication ($\rightarrow$), and equivalence ($\leftrightarrow$). $\perp$ denotes an arbitrary contradiction ("falsum") and T an arbitrary tautology. $\mathrm{Cn}(\varnothing)$ is the set of tautologies.

The *expansion* of the belief set **K** by a sentence α is denoted $\mathbf{K}+\alpha$, and is defined as follows:

$$\mathbf{K}+\alpha = \mathrm{Cn}(\mathbf{K} \cup \{\alpha\})$$

In AGM theory, *contraction* is assumed to be *minimal*, i.e. changes on the belief set are as small as is compatible with the requirement that the input sentence be removed. If the principle of minimality is applied uncompromisingly, then the contracted belief set $\mathbf{K} \div \alpha$ will have to be as large a subset of **K** as it can be without implying α. In other words, it should be an element of the set $\mathbf{K} \perp \alpha$ of inclusion-maximal subsets of **K** that do not imply α. More precisely:

Definition 3 (Alchourrón and Makinson 1981): Let A be a set of sentences and α a sentence. The set $A \perp \alpha$ ("*A* remainder alpha") is the set of sets such that $B \in A \perp \alpha$ if and only if:

(1) $B \subseteq A$

(2) $\alpha \notin \mathrm{Cn}(B)$

(3) There is no set B' such that $B \subset B' \subseteq A$ and $\alpha \notin \mathrm{Cn}(B')$

An operation $\div$ such that $\mathbf{K}\div\alpha \in \mathbf{K}\perp\alpha$ is a *maxichoice contraction* (originally called "choice contraction"). (Alchourrón and Makinson 1982) Maxichoice contraction was soon found to be unsatisfactory since it does not allow us to contract cautiously. When you find out that two of your beliefs, α and β, cannot both be retained, and you have no reason to prefer one over the other, it may be a good idea to give up both of them to be on the safe side. This type of reasoning led up to the construction of *partial meet contraction*, the major innovation in the classic 1985 paper by Carlos Alchourrón, Peter Gärdenfors and David Makinson. An operator of partial meet contraction employs a *selection function* that selects the "best" elements of $\mathbf{K}\perp\alpha$. The outcome of the contraction is equal to the intersection of the set of selected elements of $\mathbf{K}\perp\alpha$.

> **Definition 4** (Alchourrón et al 1985): Let **K** be a belief set. A *selection function* for **K** is a function γ such that for all sentences α:
>
> (1) If $\mathbf{K}\perp\alpha$ is non-empty, then $\gamma(\mathbf{K}\perp\alpha)$ is a non-empty subset of $\mathbf{K}\perp\alpha$, and
>
> (2) If $\mathbf{K}\perp\alpha$ is empty, then $\gamma(\mathbf{K}\perp\alpha) = \{\mathbf{K}\}$.

> **Definition 5** (Alchourrón et al 1985): Let **K** be a belief set and γ a selection function for **K**. The *partial meet contraction* on **K** that is generated by γ is the operation $\sim_\gamma$ such that for all sentences α:
>
> $$\mathbf{K}\sim_\gamma\alpha = \cap\gamma(\mathbf{K}\perp\alpha)$$
>
> An operation $\div$ on **K** is a partial meet contraction if and only if there is a selection function γ such that for all sentences α: $\mathbf{K}\div\alpha = \mathbf{K}\sim_\gamma\alpha$.

The following representation theorem is one of the central results of the AGM model. The six postulates referred to in the theorem are commonly called the *basic Gärdenfors postulates* (or basic AGM postulates).

> **Theorem 6** (Alchourrón et al 1985): The operator $\div$ is an operator of partial meet contraction for a belief set **K** if and only if it satisfies the following postulates:
>
> If **K** is logically closed, then so is $\mathbf{K}\div\alpha$ for all α. (*closure*)
>
> $\mathbf{K}\div\alpha \subseteq \mathbf{K}$ (*inclusion*)
>
> If $\alpha \notin \mathrm{Cn}(\mathbf{K})$, then $\mathbf{K}\div\alpha = \mathbf{K}$. (*vacuity*)
>
> If $\alpha \notin \mathrm{Cn}(\varnothing)$, then $\alpha \notin \mathrm{Cn}(\mathbf{K}\div\alpha)$ (*success*)
>
> If $\alpha\leftrightarrow\beta \in \mathrm{Cn}(\varnothing)$, then $\mathbf{K}\div\alpha = \mathbf{K}\div\beta$. (*extensionality*)
>
> $\mathbf{K} \subseteq (\mathbf{K}\div\alpha)+\alpha$ (*recovery*)

Partial meet contraction is based on what Tor Sandqvist (1995) has called the "meet of the best strategy": when there are several equally choice-worthy belief sets, their intersection is chosen. As Sandqvist has shown, the choice of this strategy is far from self-evident. If A_1 and A_2 are the best elements of $\mathbf{K}\perp\alpha$ (i.e., $\gamma(\mathbf{K}\perp\alpha) =$

$\{A_1,A_2\}$), then the intersection $A_1 \cap A_2$ may nevertheless be of very little value since it does not contain any of the sentences in A_1, respectively A_2, that made each of them valuable. There may very well be two belief sets A_3 and A_4 that are elements of $(\mathbf{K}\perp\alpha)$, but not of $\gamma(\mathbf{K}\perp\alpha)$, such that $A_3 \cap A_4$ is more valuable than $A_1 \cap A_2$. Sandqvist was able to construct a formal interpretation of choiceworthiness that makes sense of partial meet contraction (in terms of the possible worlds ruled out by belief sets), but this interpretation is not obviously compatible with basing epistemic choice on epistemic value, pragmatic value or any combination thereof.

A selection function for a belief set **K** should, for all sentences α, select those elements of $\mathbf{K}\perp\alpha$ that are "best", or most worth retaining. However, the definition of a selection function is very general, and allows for quite disorderly selection patterns. An orderly selection function should choose the best element(s) of the remainder set according to some well-behaved preference relation.

Definition 7 (Alchourrón et al 1985): A selection function γ for a belief set **K** is *relational* if and only if there is a binary relation $\sqsubseteq$ such that for all sentences α, if $\mathbf{K}\perp\alpha$ is non-empty, then
$\gamma(\mathbf{K}\perp\alpha) = \{B \in \mathbf{K}\perp\alpha \mid C \sqsubseteq B \text{ for all } C \in \mathbf{K}\perp\alpha\}$.
Furthermore, it is *transitively relational* if and only if it is based in this way on a transitive relation.
An operator of partial meet contraction is relational (transitively relational) if and only if it can be constructed from a selection function that is relational (transitively relational).

Theorem 8 (Alchourrón et al 1985): Let **K** be a belief set and $\div$ an operation for **K**. Then $\div$ is a transitively relational partial meet contraction if and only if it satisfies *closure*, *inclusion*, *vacuity*, *success*, *extensionality*, *recovery*, and:
$(\mathbf{K}\div\alpha) \cap (\mathbf{K}\div\beta) \subseteq \mathbf{K}\div(\alpha\&\beta)$ (*conjunctive overlap*)
If $\alpha \notin \mathbf{K}\div(\alpha\&\beta)$, then $\mathbf{K}\div(\alpha\&\beta) \subseteq \mathbf{K}\div\alpha$ (*conjunctive inclusion*)

The last two postulates of this theorem are called the supplementary Gärdenfors postulates.

Substantial refinements of these results have been obtained by Hans Rott (1993 and 1999), who brought to light close relationships between the properties of AGM-type selection functions and the properties of choice functions of the type studied in preference logic and social choice theory. As he has himself pointed out, these results contribute to uniting practical rationality (rational choice theory) with theoretical rationality (epistemic choices in belief change theory).

Several other constructions have been shown to coincide with transitively relational partial meet contraction. One of the more important is entrenchment-based contraction, that was proposed by Peter Gärdenfors (1988). The basic idea is that contraction of beliefs should be ruled by a binary relation of *epistemic entrenchment*. To say of two elements α and β of the belief set that "β is more entrenched than α" means that β is more useful in inquiry or deliberation, or has more "epistemic value" than α. At least ideally, it should be possible to determine the comparative degree of entrenchment of various sentences prior to (and without reference to) the operator of contraction or any other operator of change. When we perform belief contraction, the beliefs with the lowest entrenchment should turn out to be the ones that are given up.

Gärdenfors has proposed five postulates for epistemic entrenchment:

> **Definition 9** (Gärdenfors 1988): A *standard entrenchment ordering* is a binary relation $\leq$ on the language such that:
> If $\alpha \leq \beta$ and $\beta \leq \delta$, then $\alpha \leq \delta$ (*transitivity*)
> If $\alpha \vdash \beta$, then $\alpha \leq \beta$ (*dominance*)
> Either $\alpha \leq \alpha \& \beta$ or $\beta \leq \alpha \& \beta$ (*conjunctiveness*)
> If the belief set **K** is consistent, then $\alpha \notin$ **K** if and only if $\alpha \leq \beta$ for all β. (*minimality*)
> If $\beta \leq \alpha$ for all β, then k α (*maximality*)
> $<$ denotes the strict part of $\leq$ and $\equiv$ its symmetric part ($\alpha < \beta$ if and only if $(\alpha \leq \beta) \& \neg(\beta \leq \alpha)$; $\alpha \equiv \beta$ if and only if $(\alpha \leq \beta) \& (\beta \leq \alpha)$)

According to Gärdenfors, contraction can be defined in terms of entrenchment as follows:

$$(\leq \Rightarrow \div)\ \beta \in \mathbf{K} \div \alpha \text{ if and only if } \beta \in \mathbf{K} \text{ and either } \alpha < (\alpha \vee \beta) \text{ or } \alpha \in \mathrm{Cn}(\varnothing).$$

This construction is equivalent with transitively relational partial meet contraction:

> **Theorem 10** (Gärdenfors and Makinson 1988; Gärdenfors 1988): (1) Let $\leq$ be a standard entrenchment ordering on the consistent belief set **K**. Furthermore, let $\div$ be Gärdenfors's entrenchment-based contraction on **K**, based on $\leq$ according to $(\leq \Rightarrow \div)$. Then $\div$ satisfies the six basic and two supplementary Gärdenfors postulates.
> (2) Let $\div$ be an operation on the consistent belief set **K** that satisfies the six basic and two supplementary Gärdenfors postulates. Then the relation $\leq$, defined as follows:
> $\alpha \leq \beta$ if and only if $\alpha \notin A \div (\alpha \& \beta)$ or $\alpha \& \beta \in \mathrm{Cn}(\varnothing)$
> is a standard entrenchment ordering, and the contraction it gives rise to through $(\leq \Rightarrow \div)$ coincides with $\div$.

The two major tasks of a revision operator $*$ are to add the new belief α to the belief set **K**, and to ensure that the resulting belief set $\mathbf{K} * \alpha$ is consistent (unless α is inconsistent). The first task can be accomplished by expansion by α. The second task can be accomplished by prior contraction by its negation $\neg\alpha$. If a belief set does not imply $\neg\alpha$, then α can be added to it without loss of consistency.

An operator of revision can therefore be constructed out of two suboperations. The recipe is as follows:

(1) Contract by $\neg\alpha$

(2) Expand by α

Note that the two operations cannot meaningfully be performed in reverse order. If we expand a belief set by a sentence that contradicts it, then the outcome will be equal to the whole language, so that all distinctions are lost.

More succinctly, this composition of suboperations is expressed by the *Levi identity* (Gärdenfors 1981; Alchourrón and Makinson 1982; Levi 1977):

$$\mathbf{K}*\alpha = (\mathbf{K}\div\neg\alpha)+\alpha.$$

If $\div$ is an operator of partial meet contraction, then the corresponding revision operator (obtained via the Levi identity) is called an operator of *partial meet revision.* It has been axiomatically characterized as follows:

Theorem 11 (Alchourrón et al 1985; Gärdenfors 1988)**:** Let **K** be a belief set. The operator $*$ is an operator of partial meet revision for **K** if and only if it satisfies:
$\mathbf{K}*\alpha$ is a belief set (*closure*)
$\alpha \in \mathbf{K}*\alpha$ (*success*)
$\mathbf{K}*\alpha \subseteq \mathbf{K}+\alpha$ (*inclusion*)
If $\neg\alpha \notin \mathbf{K}$, then $\mathbf{K}+\alpha = \mathbf{K}*\alpha$. (*vacuity*)
$\mathbf{K}*\alpha$ is consistent if α is consistent. (*consistency*)
If $(\alpha\leftrightarrow\beta) \in \mathrm{Cn}(\varnothing)$, then $\mathbf{K}*\alpha = \mathbf{K}*\beta$. (*extensionality*)

The six postulates of this theorem are commonly called the *basic Gärdenfors postulates for revision.*

The Levi identity takes us from contraction operators to revision operators. The reverse direction is taken care of by the Harper identity. (Gärdenfors 1981)

Theorem 12 (Alchourrón et al. 1985)**:** Let **K** be a belief set, $\div$ a partial meet contraction on **K** and $*$ the partial meet revision operator derived from $\div$ via the Levi identity. Then:

$$\mathbf{K}\div\alpha = \mathbf{K}\cap(\mathbf{K}*\neg\alpha).\ (\textit{the Harper identity})$$

The AGM model, as presented above, is only concerned with changes of one and the same belief set. When we contract or revise **K** by α, partial meet revision provides us with a new belief set ($\mathbf{K}\div\alpha$ respectively $\mathbf{K}*\alpha$), but not with a new selection function that can be used for further changes of this new belief set. Clearly, a realistic model of belief change should allow for repeated (iterated) changes, such as $\mathbf{K}\div\alpha\div\beta*\delta*\varepsilon\div\zeta$. In other words, we need operators that can contract or revise any belief set by any sentence. Such operators are called *global,* in contrast to *local* operators that are defined only for a single specified belief set. In the 1990's, much of the formal work in this area was devoted to attempts to construct global operators. (Williams 1993; Levi 1988; Boutilier 1996; Darwiche and Pearl 1994. For a critical review, see Friedman and Halpern 1997.) This is no easy task, due to the paucity of the information provided by AGM inputs. The specification "revise by α" does not tell us how vulnerable to future changes the new belief α should be.

One solution to this problem is to allow for several ways to revise by one and the same sentence α. As shown by Abhaya Nayak, this can be done by letting the inputs be binary relations (that satisfy the standard entrenchment postulates except minimality). Such inputs may be seen as "fragments" of belief states, to be incorporated into the previous belief state. (Nayak 1994; Nayak et al. 1996. See also Spohn 1988 and Rott 1999.) It must be emphasized that this approach involves a rather radical departure from the simple structure of the original AGM framework. Arguably, the same may be true of all other constructions that offer a plausible solution to the problem of iterated change.

7. RECOVERY AND JUSTIFICATORY STRUCTURE

A rejected desire often leaves behind itself a residue in the form of regret for that which was given up. Defeated moral principles give rise to moral residues, e.g. in the form of duties of compensation. (Williams 1973) In analogy with this, we can suppose that rejected beliefs should, at least on some occasions, leave behind themselves *epistemic residues*, in the form of beliefs or doxastic propensities.

Epistemic residues can be expressed in the AGM framework by considering the operation of first adding a sentence α to a belief set and then contracting by it. If there are epistemic residues, then there should be at least some sentence α that leaves something behind itself that was not present in the original belief set **K**, or in other words:

> There is some belief set **K** and some sentence α such that $(\mathbf{K}+\alpha) \div \alpha \not\subseteq \mathbf{K}$ (*residuality*) (Hansson 1999*c*)

The following observation provides strong support for the residuality postulate:

> **Observation 13** (Hansson 1999*c*): Let $\div$ be a global operator such that $\mathbf{K}\div\alpha \not\subseteq \mathrm{Cn}(\{\neg\alpha\})$ holds for some belief set **K** and some sentence $\alpha \in \mathbf{K}$. Then $\div$ satisfies residuality.

Although not logically related in a straight-forward way to the residuality postulate, the recovery postulate of AGM contraction ($\mathbf{K} \subseteq (\mathbf{K}\div\alpha)+\alpha$) can be seen as expressing the existence of a special type of epistemic residue. This postulate says that after contraction by α there are sufficient sentences left to allow us to recover the original belief set if the contracted sentence α is reinstated.

Recovery is the most debated postulate of belief change. One member of the AGM trio, David Makinson (1987), has conceded that it is "open to query from the point of view of acceptability under its intended reading". Several authors have argued against it as a general principle of belief contraction. (Fuhrmann 1991; Niederée 1991; Lindström and Rabinowicz 1991; Levi 1991; Hansson 1991*b*) The following example has been offered to show that recovery does not hold in general:

Example (Hansson 1993*c*)**:** I previously entertained the two beliefs "George is a criminal" (α) and "George is a mass murderer" (β). When I received information that induced me to give up the first of these beliefs (α), the second (β) had to go as well (since α would otherwise follow from β).

I then I received new information that made me accept the belief "George is a shoplifter" (δ). The resulting new belief set is the expansion of $\mathbf{K}\div\alpha$ by δ, $(\mathbf{K}\div\alpha)+\delta$. Since α follows from δ, $(\mathbf{K}\div\alpha)+\alpha$ is a subset of $(\mathbf{K}\div\alpha)+\delta$. By recovery, $(\mathbf{K}\div\alpha)+\alpha$ includes β, from which follows that $(\mathbf{K}\div\alpha)+\delta$ includes β.

Thus, since I previously believed George to be a mass murderer, I cannot any longer believe him to be a shoplifter without believing him to be a mass murderer.

David Makinson (1997) has defended recovery against this and similar counterexamples. The apparent problems with recovery, he says, arise when we make use of a justificatory structure that is not represented in the belief set. In this example, says Makinson, we tend to take it for granted that $\beta\vee\neg\delta$ is in the belief set only because β is there. "The example is thus presented with an implicit assumption of a particular pattern of justification among the beliefs held. The [belief set > is not 'naked'."

Makinson summarizes his major conclusion as follows:

> "As soon as contraction makes use of the notion '*y* is believed only because of *x*', we run into counterexamples to recovery.... But when a theory is taken as 'naked', i.e. as a bare set $A = \mathrm{Cn}(A)$ of statements closed under consequence, then recovery appears to be free of intuitive counterexamples." (Makinson 1997)

In (Hansson 1999*c*) it was argued that Makinson was right in pointing out that this and similar counterexamples depend on the justificatory structure, that is not and cannot be reflected in the belief set. However, actual human beliefs always have such a justificatory structure; at least it does not seem possible to find a case in which they do not. It is difficult if not impossible to find examples about which we can have intuitions, and in which the belief set is not associated with a justificatory structure that guides our intuitions. Therefore, the recovery postulate is a consequence of the abstraction of a belief set from its associated justificatory structure. The recovery postulate can be defended, but only as a postulate for contraction of belief sets that have been artificially isolated in this way. The absence of justificatory structure does not seem to be a rationality requirement, or otherwise a reasonable normative requirement, on a system of beliefs.

8. BELIEF BASES

A belief set is a very large entity. For any two sentences α and β, if α is in my belief set, then so are both $\alpha\vee\beta$ and $\alpha\vee\neg\beta$, even if both they and β are sentences that I have never thought or heard of. If the language is infinite, then the belief set will contain an infinite number of sentences. It seems unnatural for changes to be performed on such large entities as belief sets, that contain all kinds of irrelevant and never-thought-of sentences. It may be more natural to think of the belief state as represented by a limited number of sentences that correspond (roughly) to the explicit beliefs. Changes can operate on this smaller set, rather than directly on the belief set. This will bring us closer to the workings of actual human minds (and actual computers).

We are thus led to represent belief states by sets of sentences that are *not* closed under logical consequence. Such sets are called *belief bases*. They are not required by definition to be finite, but in all realistic applications they will be so.

In a belief base approach, the criterion for a sentence α to be believed is that it is a consequence of the belief base B, $\alpha \in \mathrm{Cn}(B)$. The elements of the belief base are the *basic beliefs*, and the elements of its logical closure that are not elements of the belief base itself are the (merely) *derived beliefs*. In set-theoretical language:

α is a belief if and only if $\alpha \in \mathrm{Cn}(B)$
α is a basic belief if and only if $\alpha \in B$
α is a (merely) derived belief if and only if $\alpha \in \mathrm{Cn}(B)$ and $\alpha \notin B$.

Although we (are committed to) believe the logical consequences of our basic beliefs, these consequences are subject only to exactly those changes that follow from changes of the basic beliefs. If one of the merely derived beliefs loses the support that it had in basic beliefs, then it will be automatically discarded. (This process has been called 'disbelief propagation'. (Martins and Shapiro 1988))

> **Example** (Hansson 1994): I believe that Paris is the capital of France (α). I also believe that there is milk in my fridge (β). Therefore, I believe that Paris is the capital of France if and only if there is milk in my fridge ($\alpha \leftrightarrow \beta$). I open the fridge and find it necessary to replace my belief in β with belief in $\neg\beta$. I cannot then, on pain of inconsistency, retain both my belief in α and my belief in $\alpha \leftrightarrow \beta$.

In a belief set approach, both α and $\alpha \leftrightarrow \beta$ are elements of the belief set. When I open my fridge and find no milk, I make a choice between retaining α and retaining $\alpha \leftrightarrow \beta$. The retraction of $\alpha \leftrightarrow \beta$ does not follow automatically. It has to be ensured by a selection mechanism (such as a selection function) that chooses between α and $\alpha \leftrightarrow \beta$. (Gärdenfors 1990) In the belief base approach, on the other hand, β in our example is a basic belief, whereas $\alpha \leftrightarrow \beta$ is a merely derived belief. When β is removed, $\alpha \leftrightarrow \beta$ disappears automatically. The option of retaining it will not even arise.

For every belief base B, there is a belief set $\mathrm{Cn}(B)$ that represents the beliefs held according to B. On the other hand, one and the same belief set can be represented by different belief bases. In this sense, belief bases have more expressive power than belief sets. As an example, the two belief bases $\{\alpha, \beta\}$ and $\{\alpha, \alpha \leftrightarrow \beta\}$ have the same logical closure, since $\mathrm{Cn}(\{\alpha, \beta\}) = \mathrm{Cn}(\{\alpha, \alpha \leftrightarrow \beta\})$. Nevertheless, these belief bases are not identical. They are *statically equivalent*, in the sense of representing the same beliefs. On the other hand, the following example shows that they are not *dynamically equivalent* in the sense of behaving in the same way under operations of change.

> **Example:** Let α denote that the Liberal Party will support the proposal to subsidize the steel industry, and let β denote that Ms. Smith, who is a liberal MP, will vote in favour of that proposal.
> Abe has the basic beliefs α and β, whereas Bob has the basic beliefs α and $\alpha \leftrightarrow \beta$. Thus, their beliefs (on the belief set level) with respect to α and β are the same.
> Both Abe and Bob receive and accept the information that α is false, and they both revise their belief states to include the new belief that $\neg\alpha$. After that, Abe has the basic beliefs $\neg\alpha$ and β, whereas Bob has the basic beliefs $\neg\alpha$ and $\alpha \leftrightarrow \beta$. Now, their belief sets are no longer the same. Abe believes that β whereas Bob believes that $\neg\beta$.

Since belief sets are logically closed, there is only one inconsistent belief set. In other words, if two belief sets are both inconsistent, then they are identical. The corresponding property does not hold for belief bases. The following two belief bases:

$B_1 = \{p, \neg p, q_1, q_2, q_3, q_4\}$ and
$B_2 = \{p, \neg p, \neg q_1, \neg q_2, \neg q_3, \neg q_4\}$

are both inconsistent, but they are not identical. They are statically equivalent, since $\mathrm{Cn}(B_1) = \mathrm{Cn}(B_2)$. However, they are not dynamically equivalent since, by any reasonable operator of contraction:

$B_1 \div p = \{\neg p, q_1, q_2, q_3, q_4\}$ and
$B_2 \div p = \{\neg p, \neg q_1, \neg q_2, \neg q_3, \neg q_4\}$

so that $\mathrm{Cn}(B_1 \div p) \neq \mathrm{Cn}(B_2 \div p)$.

Belief bases have the advantage of allowing for more distinctions than belief sets, but they also give rise to troublesome questions on how these distinctions should be drawn.

As was indicated above, the ultimate criterion for a belief to be an element of the belief base is that it is "self-sustained", i.e., worth retaining for its own sake (even if it is not implied by some other belief that is worth retaining). In a sense, however, this is a reformulation of the question rather than an answer. The next question is: Which beliefs are self-sustained in this sense?

Belief bases have often been taken to consist of the beliefs that have independent justification. This is only a very rough approximation, since beliefs may be self-sustained without being independently justified:

> **Example** (Hansson 1989): I originally believed, for good and independent reasons, both that Andy is the mayor's son (α) and that Bob is the mayor's son

> (β). Then I hear the mayor say in a public speech: "I certainly have nothing against our youth studying abroad. My only son did it for three years."
> Upon hearing this, I contract my belief state by $\alpha\&\beta$. As a result of this I lose both my belief in α and my belief in β. However, I retain my belief that $\alpha\vee\beta$, i.e., that either Andy or Bob is the son of the mayor.

In this case, we may assume that $\alpha\vee\beta$ had no independent justification. It was believed only as a consequence of my beliefs in α and β. If the belief base was $\{\alpha,\beta\}$, then $\alpha\vee\beta$ cannot be an element of the contracted belief base. It seems reasonable, however, in this and many other cases, to retain belief in the disjunction of two independently justified beliefs, when they can no longer coexist and one cannot choose between them.

The difference between belief bases and belief sets has often been related to the distinction between foundationalist and coherentist epistemology. (Doyle 1992; Gärdenfors 1990) Belief bases have been taken to represent the foundations of a foundationalist belief system. Belief sets, on the other hand, are said to represent a coherentist structure. Recently, it has been argued that this account of foundationalism and coherentism is misleading. (Hansson and Olsson 1999) The simple deductive relationship between a belief base and the corresponding belief set does no justice at all to the complex relations of justification in a reasonable version of foundationalism. The deductive relationships of belief sets do not either fit in exactly with the coherentist view. Although coherentists typically claim that *all* beliefs contribute to the justification of other beliefs, they hardly mean this to apply to merely derived beliefs such as "either Paris or Nice is the capital of France", that I believe only because I believe Paris to be the capital of France. To the extent that belief sets represent coherentism, and belief bases foundationalism, they do so in a sense of the two terms that is not the same as that of traditional, non-formal epistemology.

9. OPERATIONS ON BELIEF BASES

It is a fairly easy matter to transfer the AGM operations to belief bases, i.e., sets of sentences that are not closed under logical consequence. The same three main types of change, *viz.* expansion, contraction, and revision, have been applied to belief bases as well as to belief sets.

Due to the principle of categorical matching, the outcome of *expansion* on a belief base should be a belief base, and thus not logically closed. Therefore, the expansion operation of AGM has to be adjusted to be suitable for belief bases:

> **Definition 14:** Let B be a belief base and α a sentence. B+α, the (*non-closing*) *expansion* of B by α, is defined as follows:
> $B+\alpha = B\cup\{\alpha\}$.

Partial meet contraction, as defined in Section 6, can be applied, unmodified, to belief bases, but the axiomatic characterization will have to be different. (Hansson

1993*a* and 1993*b*) In particular, the recovery postulate is not satisfied for belief bases.

Just like the corresponding operators for belief sets, revision operators for belief bases can be constructed out of two suboperations: expansion by α and contraction by $\neg\alpha$. According to the Levi identity ($B*\alpha = (B\div\neg\alpha)+\alpha$), we should first contract, and then (non-closingly) expand:

(1) Contract by $\neg\alpha$
(2) Expand by α

Alternatively, the two suboperations may take place in reverse order:

(1) Expand by α
(2) Contract by $\neg\alpha$

More compactly, this is expressed by the *reversed Levi identity* (Hansson 1993*a*):

$$B*\alpha = (B+\alpha)\div\neg\alpha.$$

As was indicated in Section 6, this latter possibility does not exist for belief sets. If $\mathbf{K} \cup \{\alpha\}$ is inconsistent, then $\mathbf{K}+\alpha$ is always the same (namely identical to the whole language), so that all distinctions are lost. For belief bases this limitation is not present, and we have two distinct ways of basing revision on expansion and partial meet contraction:

Definition 15: Let $\div$ be a global contraction operator on belief bases. Then:
(1) the operator of *internal revision*, based on $\div$, is the operator $\mp$ such that:
$$B\mp\alpha \;=\; (B\div\neg\alpha)+\alpha$$
(2) the operator of *external revision*, based on $\div$, is the operator $\pm$ such that:
$$B\pm\alpha \;=\; (B+\alpha)\div\neg\alpha$$

The names "internal" and "external" revision indicate that in internal revision, the suboperation of contraction takes place inside the original belief base, whereas in external revision it takes place outside of the original set. The symbols $\pm$ and $\mp$ should be read top-down: in external revision ($\pm$) expansion (+) takes place first, and is followed by contraction (–).

The two revision operators have been shown to differ in their formal properties (Hansson 1993*a*). They correspond to different intuitions about how belief-contravening information should be accommodated by a rational doxastic agent. Consistency is preserved in every step of internal revision, but there is an *intermediate non-committed state* in which neither the input sentence α nor its negation $\neg\alpha$ is believed. In (belief-contravening) external revision, there is instead an *intermediate inconsistent state* in which both α and $\neg\alpha$ are believed, Which of the two operations is the more plausible? Our intuitions about this seem to differ between different cases:

Examples: (1) Anthony and Beatrice are a married couple. I used to think that they were both Roman Catholics. Then I heard Beatrice say: "In our marriage, it was never a problem that we belong to different denominations." When I heard this, I gave up my belief that Beatrice was a Roman Catholic, but I retained my belief that Anthony was so (since I have seen him enter the local Catholic Church several times).
(2) When Joseph Black learned of the results of Lavoisier's new experiments, he gave up his previous belief in the phlogiston theory of combustion, and accepted Lavoisier's oxygen theory.
(3) I believed that John was dead. Then I met him in the street.

In case 1, the "external" account seems to be the most plausible one. More generally: if it is obvious that the new information must be accepted, but less obvious which previous beliefs it should push out, then external revision seems to be closest to the actual psychological process. In case 2, there was a phase of hesitation in which neither the new belief nor its negation was accepted. Internal revision is closer than the external variant to this kind of process. In case 3, it is difficult to determine if internal or external revision is the most adequate model. Intuitively, the two operations seem to be simultaneous – a feature that is not easy to capture in logical representation.

10. NON-PRIORITIZED BELIEF CHANGE

According to the success postulate of belief revision ($\alpha \in \mathbf{K}*\alpha$, or for belief bases $\alpha \in B*\alpha$), the input sentence is always accepted. In actual epistemic and doxastic processes, this is certainly not true. To determine whether or not to accept a new piece of information is no less an essential process than to determine, in the former case, which old sentences to throw out in order to preserve consistency. In models of *non-prioritized* belief revision, the success postulate has been relaxed, and new information is not given the absolute priority that it has in the AGM framework.

One way to construct non-prioritized belief revision is to base it on the following two-step process: First we decide whether to accept or reject the input. After that, if the input was accepted, then it is incorporated into the belief state through (conventional) revision.

Decision+revision:
(1) Decide whether the input α should be accepted or rejected.
(2) If α was accepted, revise by α.

The decision+revision model is foreshadowed in some of Isaac Levi's work, but the first fully formalized model of it seems to be David Makinson's (1997) *screened revision*. This operation makes use of a set A of potential core beliefs that are immune to revision. The belief set $\mathbf{K}$ should be revised by the input sentence α if α is consistent with the set $A \cap \mathbf{K}$ of actual core beliefs; otherwise it remains unchanged. A series of other revision models based on the decision+revision recipe is introduced in (Hansson et al. 2000).

Both steps in a decision-revision model involve a choice among beliefs. In the first step, a choice is made between the input sentence α and beliefs already held. The second step involves a choice among previous beliefs, some of which may have to be given up in order to retain consistency when α is added. The relationship between these two choices is an interesting and largely unanalyzed issue. In (Makinson 1997) they are independent of each other, whereas in (Hansson et al. 2000) they are based on the same selection mechanism. It can be argued that the truth must be sought somewhere these two extremes.

Another approach to non-prioritized revision is to provisionally accept the new information and, if this led to inconsistency, afterwards regain consistency by throwing out either the input or some of the previous beliefs.

Expansion+consolidation (Hansson 1994 and 1997):
(1) Expand by α.
(2) Consolidate the belief state.

where consolidation is a procedure that makes the belief state consistent. This approach has been developed only for belief bases. Consolidation can be defined as contraction by a contradictory sentence. Erik Olsson (1997*a* and 1997*b*) has developed another variant, in which the consistency requirement is replaced by a requirement of coherence. Another interesting development is to use a "localized" consolidation operator that consolidates only a compartment of the belief base. Contrary to full consolidation, this process will not eradicate all inconsistencies. (Wassermann and Hansson 1999) This is a realistic feature, since in real life inconsistencies are often tolerated, and do not propagate to make the whole belief state degenerate.

Non-prioritized belief revision is a relatively new research area. (For an overview, see Hansson 1999*b*.) The relationship between the two recipes, decision+revision and expansion+consolidation, remains to investigate. We do not yet know if one of them can be reduced to the other. It also remains to investigate how the first step in decision+revision can be brought to relate to the models of rational decision-making that have been developed in decision theory.

11. HOW USEFUL IS BELIEF CHANGE THEORY?

In some areas of philosophy, the use of formal logic has made it possible to treat philosophical problems in a much more precise and clarifying manner. (Studies of the relation between truth and language provide excellent examples of this.) However, formalization is not always useful. Unfortunately, there are also examples of formalizations that have given rise to more confusion than clarity.

The philosophical and interpretational discussions that a formal treatment gives rise to can be divided into three categories:

(1) New aspects on issues already discussed in informal philosophy.

(2) Issues not previously discussed in informal philosophy, but with a clear philosophical interest.
(3) Issues that are peculiar to the chosen formalism and have no bearing on philosophical issues that can be expressed without the formalism.

(Some of the paradoxes discussed in deontic logic exemplify the third category.) A formal treatment is more successful, the more it gives rise to discussions of the first two types and the less it gives rise to discussions of the third type. As compared to other branches of philosophical logic, belief change theory has been fairly successful, due to a reasonable number of discussions of the second type that it has given rise to. Some of these have been mentioned above, including the following:

- the relationship between choice and retrieval
- the reducibility of complex belief changes into sequences of simple changes (the decomposition principle)
- the relationship between rationality in choice and rationality in belief change
- the nature of the epistemic residues left behind by rejected beliefs
- the nature of the justificatory structure in terms of which the distinction between foundationalist and coherentist epistemology can be expressed
- the role of merely derived beliefs in coherentist belief systems
- the role of intermediate non-committed states and intermediate inconsistent states in revision (internal respectively external revision)

On the other hand, belief change theory has not given rise to many discussions of the first type, i.e. formalized clarifications of issues already discussed in informal epistemology. In my view, this depends on three barriers inherent in the predominant models of belief change, that make them unsuitable for most epistemological applications.

One of these barriers is the success postulate: the input sentence is always accepted. In epistemology, the crucial issue is typically whether or not to accept a new piece of information, rather than exactly how to incorporate it when this is done. The recent development of non-prioritized models of belief revision will break down this first barrier, as is already evident from Erik Olsson's pioneering analysis of Keith Lehrer's coherence theory in terms of concepts from non-prioritized belief revision. (Olsson 1997*a*)

The second barrier is that traditional belief revision models represent minimal change, whereas important belief change processes discussed in epistemology are non-minimal. Induction and explanation are obvious examples. To add a generalization or an explanation is, in a logical sense, a non-minimal extension of the original belief set. Therefore, in order to capture some types of belief change we need models of non-minimal change. This is essentially an unexplored field (but there are interesting beginnings, such as Pagnucco 1996).

The third barrier is that belief change theory has focused on the *internal* workings of doxastic or epistemic agents. The relations between states of belief and the objects that these states refer to remain an essentially unexplored issue. Clearly, major theories of truth cannot be accounted for in such a framework. More sophisticated models that distinguish between knowledge and belief need to be developed in order to make belief change theory more directly useful in epistemological studies.

Belief change theory has great promise for providing a more precise account of central issues in epistemology. However, although the foundations have been laid, further conceptual and formal developments are needed before that promise can be fulfilled.

Sven Ove Hansson
Royal Institute of Technology

REFERENCES

Alchourrón, C. E., P. Gärdenfors and D. Makinson: 1985, 'On the Logic of Theory Change: Partial Meet Contraction and Revision Functions', *Journal of Symbolic Logic* **50**, 510-530.

Alchourrón, C. E. and D. Makinson: 1981, 'Hierarchies of Regulation and Their Logic', in R. Hilpinen (ed.), *New Studies in Deontic Logic*, Reider Publishing Company, Dordrecht, pp. 125-148.

Alchourrón, C. E., and D. Makinson: 1982, 'On the logic of theory change: Contraction functions and their associated revision functions', *Theoria* **48**, 14-37.

Boutilier, C.: 1996, 'Iterated revision and minimal change of conditional beliefs', *Journal of Philosophical Logic* **25**, 262-305.

Brachman, R. J. and H. J. Levesque: 1986, 'What Makes a Knowledge Base Knowledgeable? A View of Databases from the Knowledge Level', in L. Kerschberg (ed.), *Expert Database Systems, Proceeding from the first international workshop*, The Benjamin/Cummings Publishing Company, pp. 69-78.

Dalal, M.: 1988, *Investigations Into a Theory of Knowledge Base Revision: Preliminary Report*, Morgan Kaufmann, pp. 475-479.

Darwiche, A. and J. Pearl: 1994, 'On the logic of iterated belief revision', in R. Fagin (ed.), *Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann, San Francisco, California, pp. 5-23.

Doyle, J.: 1979, 'A Truth Maintenance System', *Artificial Intelligence* **12**, 231-272.

Doyle, J.: 'Rational belief revision (preliminary report)', in J. Allen, R. Fikes, R., and E. Sandewall (eds.), *Principles of Knowledge Representation and Reasoning*, Morgon Kaufmann, Los Altos, CA, pp. 163-174.

Doyle, J.: 1992, 'Reason maintenance and belief revision: Foundations versus coherence theories', in P. Gärdenfors (ed.), *Belief Revision*. Cambridge University Press, Cambridge, pp. 29-51.

Dubois, D., S. Moral and H. Prade: 1998, 'Belief change rules in ordinal and numerical uncertainty theories', in D. M. Gabbay and P. Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3, D. Dubois and H. Prade (eds.), *Belief Change*, Kluwer, Dordrecht, pp. 311-392.

Friedman, N. and J. Y. Halpern: 1997, *Belief Revision: A Critique*, manuscript.

Fuhrmann, A.: 1988, *Relevant Logics, Modal Logics, and Theory Change*, Doctoral Thesis, Australian National University, Canberra.

Fuhrmann, A.: 1989, 'Reflective Modalities and Theory Change', *Synthese* **81**, 115-134.

Fuhrmann, A.: 'Theory Contraction Through Base Contraction', *Journal of Philosophical Logic* **20**, 175-203.

Fuhrmann, A. and S. O. Hansson: 1994, 'A Survey of Multiple Contraction', *Journal of Logic, Language and Information* **3**, 39-75.

Gallier, J. R.: 1992, 'Autonomous belief revision and communication', in P. Gärdenfors (ed.), *Belief Revision*, Cambridge University Press, Cambridge, pp. 220-246.

Gärdenfors, P.: 1981, 'An Epistemic Approach to Conditionals', *American Philosophical Quarterly* **18**, 203-211.

Gärdenfors, P.: 1986, 'Belief Revision and the Ramsey Test for Conditionals' *Philosophical Review* **95**, 81-93.

Gärdenfors, P.: 1988, *Knowledge in Flux. Modeling the Dynamics of Epistemic States.* The MIT Press, Cambridge, Massachusetts.

Gärdenfors, P.: 1990, 'The Dynamics of Belief Systems: Foundations vs. Coherence Theories', *Revue Internationale de Philosophie* **44**(172), 24-46.

Gärdenfors, P. and D. Makinson: 1988, 'Revisions of Knowledge Systems Using Epistemic Entrenchment', *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, Morgan Kaufmann Publ., pp. 83-95.

Gärdenfors, P. and H. Rott: 1995, 'Belief Revision', in Gabbay, Hogger and Robinson (eds.), *Handbook of Logic in AI and Logic Programming*, Oxford University Press, Oxford, pp. 35-132.

Hansson, S. O.: 1989, 'New Operators for Theory Change', *Theoria* **55**, 114-132.

Hansson, S. O.: 1991a, *Belief Base Dynamics*, Doctoral Thesis, Uppsala University.

Hansson, S. O.: 1991b, 'Belief Contraction Without Recovery', *Studia Logica* **50**, 251-260.

Hansson, S. O.: 1992, 'In Defense of the Ramsey Test', *Journal of Philosophy* **89**(10), 522-540.

Hansson, S. O.: 1993a, 'Reversing the Levi Identity', *Journal of Philosophical Logic* **22**, 637-669.

Hansson, S. O.: 1993b, 'Theory Contraction and Base Contraction Unified', *Journal of Symbolic Logic* **58**, 602-625.

Hansson, S. O.: 1993c, 'Changes on Disjunctively Closed Bases', *Journal of Logic, Language and Information* **2**, 255-284.

Hansson, S. O.: 1995a, 'Taking Belief Bases Seriously', in D. Prawitz and D. Westerståhl (eds.), *Logic and Philosophy of Science in Uppsala*, Kluwer Academic Publishers, Uppsala, pp. 13-28.

Hansson, S. O.: 1995a, 'Changes in Preference', *Theory and Decision* **38**, 1-28.

Hansson, S. O.: 1995b, 'The Emperor's New Clothes. Some recurring problems in the formal analysis of counterfactuals', in G. Crocco, L. Fariñas del Cerro and A. Herzig (eds.), *Conditionals: from Philosophy to Computer Science*, Clarendon Press, Oxford, pp. 13-31.

Hansson, S. O.: 1997, 'Semi-Revision', *Journal of Applied Non-Classical Logic* **7**, 151-175.

Hansson, S. O.: 1998, 'Revision of Belief Sets and Belief Bases', in D. M. Gabbay and P. Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3, D. Dubois and H. Prade (eds.), *Belief Change*, Kluwer, Dordrecht.

Hansson, S. O.: 1999a, *A Textbook of Belief Dynamics*, Kluwer.

Hansson, S. O.: 1999b, 'A Survey of Non-Prioritized Belief Revision', *Erkenntnis* **50**, 413-427.

Hansson, S. O.: 1999c, 'Recovery and epistemic residues', *Journal of Logic, Language and Information* **8**, 421-428.

Hansson, S. O., E. Fermé, J. Cantwell and M. Falappa: 2000, 'Credibility-limited Revision', *Journal of Symbolic Logic*, in press.

Hansson, S. O. and E. Olsson: 1999, 'Providing Foundations for Coherentism', *Erkenntnis* **51**, 243-265..

Hansson, S. O. and R. Wassermann: 1999, 'Local change', manuscript.

Harper, W.: 1977, 'Rational Conceptual Change', in *PSA 1976*, 462-494.

Jeffrey, R. C.: 1956, 'Valuation and Acceptance of Scientific Hypotheses', *Philosophy of Science* **23**, 237–249.

Katsuno, H., A. O. Mendelzon: 1989, 'A Unified View of Propositional Knowledge Base Updates', in *11th International Joint Conference on Artificial Intelligence*, Morgan Kaufmann, pp. 269-276.

Katsuno, H., A. O. Mendelzon: 1992, 'On the Difference between Updating a Knowledge Base and Revising it', in P. Gärdenfors (ed.), *Belief Revision*, Cambridge University Press, Cambridge, pp. 183-203.

Keller, A. and M. Winslett: 1985, 'On the use of an extended relational model to handle changing incomplete information', *IEEE Transactions on Software Engineering* **SE-11:7**, 620-633.

Levi, I.: 1977, 'Subjunctives, Dispositions and Chances', *Synthese* **34**, 423-455.

Levi, I.: 1980, *The Enterprise of Knowledge*, The MIT Press, Cambridge, Massachusetts.

Levi, I.: 1988, 'Iteration of Conditionals and the Ramsey Test', *Synthese* **76**, 49-81.

Levi, I.: 1991, *The Fixation of Belief and Its Undoing*, Cambridge University Press, Cambridge, Mass.

Li, J.: 1998, 'A Note on Partial Meet Package Contraction', *Journal of Logic, Language and Information* **7**, 139-142.

Lindström, S. and W. Rabinowicz: 1989, 'On probabilistic representation of non-probabilistic belief revision', *Journal of Philosophical Logic* **18**, 69-101.

Lindström, S. and W. Rabinowicz: 1991, *Epistemic entrenchment with incomparabilities and relational belief revision*, in A. Fuhrmann and M. Morreau (ed.), *The Logic of Theory Change*, Springer-Verlag, Berlin, pp. 93-126.

Lindström, S. and W. Rabinowicz: 1998, 'Conditionals and the Ramsey test', in D. M. Gabbay and P. Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3, D. Dubois and H. (eds.), *Belief Change*, Kluwer, Dordrecht, pp. 147-188

Makinson, D.: 1987, 'On the Status of the Postulate of Recovery in the Logic of Theory Change', *Journal of Philosophical Logic* **16**, 383-394.

Makinson, D.: 1997, 'On the force of some apparent counterexamples to recovery', in E. G. Valdés (ed.), *Normative Systems in Legal and Moral Theory. Festschrift for Carlos E. Alchourrón and Eugenio Bulygin,* Duncker & Humblot, Berlin, pp. 475-481.

Makinson, D.: 1997, 'Screened Revision', *Theoria* **63**, 14-23.

Martins, J. P. and S. C. Shapiro: 1988, 'A Model for Belief Revision', *Artificial Intelligence* **35**, 25-79.

McLaughlin, A.: 1970, 'Science, Reason and Value', *Theory and Decision* **1**, 121–137.

Nayak, A.: 1994, 'Iterated Belief Change Based on Epistemic Entrenchment', *Erkenntnis* **41**, 353-390.

Nayak, A, P. Nelson and H. Polansky: 1996, 'Belief Change as Change in Epistemic Entrenchment', *Synthese* **109**, 143-174.

Newell, A.: 1982, 'The Knowledge Level', *Artificial Intelligence* **18**, 87-127.

Niederée, R.: 1991, 'Multiple contraction. A further case against Gärdenfors principle of recovery', in A. Fuhrmann and M. Morreau (ed.), *The Logic of Theory Change. Berlin: Springer-Verlag*, pp. 322-334.

Olsson, E. J.: 1997a, *Coherence. Studies in Epistemology and Belief Revision*, Ph.D. Thesis, Uppsala University, Uppsala.

Olsson, E.J.: 1997b, 'A Coherence Interpretation of Semi-Revision'. *Theoria* **63**, 105-134.

Pagnucco, M.: 1996, *The Role of Abductive Reasoning within the Process of Belief Revision*, pH Thesis, University of Sydney, Australia.

Peirce, C.: 1934, 'The fixation of belief', in C. Hartshorne and P. Weiss (ed.), *Collected Papers of Charles Sanders Peirce*, Harvard University Press, Cambridge, pp. 223–247.

Rott, H.: 1993, 'Belief contraction in the context of the general theory of rational choice', *Journal of Symbolic Logic* **58**, 1426-1450.

Rott, H.: 1999, *Change, Choice and Inference*, Oxford University Press, in press.

Sandqvist, T.: 1995, 'Why should the best always meet?. On the Intuitive Basis of Some Contraction Operations', in S. O. Hansson and W. Rabinowicz (eds.),. *Logic for a Change. Essays dedicated to Sten Lindström on the occasion of his fiftieth birthday*, Uppsala Prints and Reprints, Uppsala, pp. 125-135.

Smets, P.: 1998, 'Numerical representations of uncertainty', in Gabbay and P. Smets (eds.), *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. 3, D. Dubois and H Prade (eds.), *Belief Change*, Kluwer, Dordrecht, pp. 265-309.

Spohn, W.: 1988, 'Ordinal conditional functions: A dynamic theory of epistemic states', in W. Harper and B. Skyrms (ed.), *Causation in Decision, Belief Change and Statistics*, pp. 105-134.

Tarski, A.: 1956, 'On some fundamental concepts of metamathematics', in A. Tarski (ed.), *Logic, Semantics, Metamathematics*, Oxford University Press, Oxford, pp. 30-36.

Williams, B.: 1973, 'Ethical Consistency', in B. Williams (ed.), *Problems of the Self: Philosophical Papers 1956–1972*, Cambridge University Press, Cambridge, pp. 166–186.

Williams, M.-A.: 1993, *Transmutations of Knowledge Systems*, University of Sidney, Australia, 1993.

PART III: TRUTH AND JUSTIFICATION

ROBERT K. SHOPE

THE ANALYSIS OF KNOWING

1 TYPES OF ANALYSIS

1.1 Conceptual Analysis vs. Alternative Forms of Analysis

When philosophers speak of concepts, they are seldom concerned with an everyday focus on a given person's 'conception' of something, which might include what the person thinks is important about a topic or ought to happen regarding it. Nor are they typically discussing what psychologists call 'concept acquisition' in the sense of a person's coming to be able to make judgments about a given topic. Instead, philosophers are quite often relating to a tradition illustrated by Kant and having roots in Plato's theory of Forms, which speaks of concepts as particulars that figure in judgments, or in propositions specifying the contents of judgments, and treats concepts as applicable to or true of various items. Philosophers in that tradition presume that such particulars can be described in what they call an analysis of a concept.

Many philosophers regard Plato's *Euthyphro* as the beginning of one current within this tradition, according to which analyzing a concept begins by articulating some condition implied by the application of the concept, even within fiction, called a 'conceptually (or logically) necessary' condition forming part of the content of the concept. The analysis is completed by listing enough such conditions that their joint satisfaction involves applicability of the concept. Then the conditions are 'jointly conceptually (or logically) sufficient' for such applicability and exhaust the content of the concept.

Other currents in the tradition allow that at least some concepts might have their contents described in a different fashion, e.g. by being treated as 'family resemblance' concepts, or as 'cluster' concepts. But much philosophical research in the last half century has involved reactions to concerns about conceptual analyses of a subject, S, knowing that P (where a complete declarative sentence is to be instantiated for 'P'), which contain three types of conditions purported to be individually conceptually necessary and jointly conceptually sufficient for the application of the concept of knowing that P to S.

1.2 Three Standard Conditions of Knowing That

Let us follow the common practice of calling an analysis of S's knowing that P a justified, true belief analysis (JTB analysis, for short) when it has the following structure or a close variation on it:

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 283-329.
© 2004 *Kluwer Academic Publishers.*

(JTB) S knows that P if and only if
(1) it is true that P,
(2) S is justified in believing that P, and
(3) S believes that P.

The first condition is commonly called the truth condition, although it is sometimes phrased without the term 'true' by simply requiring that P. The second condition can be viewed as not implying the third, since Watson and Holmes might each possess excellent evidence that P whose force Watson nonetheless does not appreciate, so that he is not led to the belief that P, which Holmes does form in light of the evidence. The second condition is commonly called the justification condition, although variants concern other epistemic states of S, for instance, S's having adequate evidence that P; or its being evident to S that P; or S's having a right to be sure that P; or its being certain for S that P. The third condition, called the belief condition, has some variants that require a special degree of firmness of the belief that P or confidence accompanying it, or may speak not of belief but of a different kind of acceptance that P, or of being sure or certain that P. There are indications of JTB analyses in Plato's *Theatetus* and *Meno*, in the works of Kant, and during this century in the writings of A. J. Ayer and Roderick Chisholm.

1.3 Linguistic Analyses

Philosophers who take the linguistic turn are less prone to speak of analyzing concepts than to speak of analyzing/defining/explaining linguistic expressions, and may concentrate upon expressions of the form, 'S knows that P'. Such philosophers must be careful not to confuse a description of conditions required for the linguistically appropriate use of such an expression with conditions for the truth of the statement made by means of the expression. Moreover, some will treat as an aspect of the meaning of an expression what H. P. Grice (1961) called its contextual implications.[1] Such philosophers will not regard a philosophical analysis of *knowing that* to be complete unless it spells out any contextual implications of expressions of the form, 'S knows that P', and discusses which, if any, are cancellable.

1.4 Analytic versus Synthetic

Many philosophers will not regard explaining contextual implications as an aspect of analyzing a concept. They will restrict analysis to the specification of conditions 'analytic upon' the concept. This terminology hearkens back to Kant's distinction between analytic truths and synthetic truths. But the clarity and existence of such a distinction has come under serious attack, notably by W. V. Quine, who is not even comfortable with regarding it as a philosophically respectable aim to articulate the meaning of sentences, taken one at a time, such as those of the above form.

1.5 A Project in Common

In light of such controversy, we can broaden the present discussion by remaining noncommittal as to whether philosophers who have contributed to the debate over the proper analysis of knowledge–in the sense of the analysis of a person's or animal's knowing that something is so–have been working within the tradition of conceptual analysis or instead have been linguistic philosophers concerned with meaning. For there is one project that even philosophers of the Quinean persuasion recognize as of philosophical interest and as prima facie possessing potential scientific significance, to which a conceptual analysis of knowing or an explanation of the meaning of expressions of the form, 'S knows that P', might be relevant. It is the project of articulating lawlike statements concerning knowing. Keith Lehrer (1974) is willing to call such a project an analysis of knowledge, and has expressed the hope that such an analysis might eventually be 'poured out into' a scientific theory of knowledge.

This perspective can be illustrated by reference to a JTB analysis, even when the analysis is presented as a conceptual analysis or a meaning analysis. The presence of each condition in the analysans corresponds to some schema for generalizations that a defender of the analysis accepts. For instance, corresponding to the truth condition is the schema, 'S knows that P only if it is true that P'. And the analysis as a whole endorses the schema, 'S knows that P if and only if S has justified, true belief that P'.

Philosophers who reject the *Euthyphro* model for characterizing conditions of knowing will fill in something of a different, possibly much more complex, form to the right of the above biconditional connective, 'if and only if'. Lehrer's point is that the resulting schema will be a lawlike biconditional, with implications not just for actual cases of knowing but also for hypothetical, counterfactual cases. The latter implications are expressible most briefly by subjunctive conditionals of the form, 'If S were to know that P then S would have justified, true belief that P', and of the form, 'If S were to have justified, true belief that P then S would know that P'.

1.6 Constitutive Analyses

A special variety of analysis of this form is what I have elsewhere (1999) called a 'constitutive analysis'. Roughly, the biconditional in such an analysis is both true and counterfactual-sustaining because any state of affairs labelled by a gerundive (e.g., 'S's knowing that P') corresponding to a sentence having the form of the analysandum is constituted by a state of affairs labelled by a gerundive (e.g., 'S's having justified, true belief that P') corresponding to a sentence having the form of the analysans.

1.7 Ambiguity Concerning Analysis

An implication of the above considerations is that merely from the fact that a philosopher speaks of providing an analysis of knowledge and offers it in a schema of the form, (ø) 'S knows that P, if and only if Q', we cannot tell which of the above types of analysis is intended. And many articles on knowing have indeed left that ambiguous. Yet suppose that there are arguments aiming to show that some statements of form (ø) are false because a statement of the form, 'S knows that P, if Q', is false regarding some actual or hypothetical case, i.e., aiming to show that the proposed analysis is too weak to cover all cases of *knowing that.* Or suppose that there are arguments aiming to show that there are cases rendering false some statements of the form, 'S knows that P, only if Q', i.e., showing that the analysis is too strong and rules out genuine cases of knowledge. Such arguments challenge the analysis no matter what type it is. Perhaps this is why so many philosophers leave ambiguous which of the above types of analysis they are pursuing.

The upshot is that we shall also be able to leave this issue vague when focusing upon the extensive literature that has emerged from the efforts of philosophers to describe real or hypothetical examples challenging or defending proposed analyses of form (ø) and from the efforts to develop amended versions of such analyses in response to those challenges.

2 Dangers in Narrowing the Focus of Discussion

Much of this literature has developed from a line of debate stemming from an attack on JTB analyses presented in a famous, brief paper by Edmund Gettier. Recent discussions of the analysis of knowing, even when presented as general inquiries into the topic, typically launch early on into a consideration of the 'Gettier Problem', and responses to it. And indeed we shall do so shortly. But several possible dangers in this manner of proceeding should be noted.

2.1 First Person vs. Third Person Uses

One danger is that the tradition's focus on analyses of form (ø) may conceal something important concerning the meaning of 'know' that concerns a contrast between first and third person uses of the term. D. S. Clarke Jr. (1989) cites J. L. Austin's suggestion that when I assert that I know something to be so, the sentence functions to give my guarantee or strong endorsement that it is so. Of course, those philosophers who regard the tradition within which JTB analyses were developed may admit that such an analysis only captures something in common to first and third person usages. They might allow that first person usages perform the further job mentioned by Clarke or might treat the performance of that job as a contextual implication of the assertion.

But Clarke takes a more radical view of Austin's insight, namely, that the *entire* linguistic function of first person instances of 'I know that' is to indicate that the utterance is to be taken as a guarantee, so that the only descriptive function of first person instances of the schema is what is performed via the expression instantiated for 'P', and there is no description of oneself as believing such a thing or being justified in believing it.

Clarke overlooks the fact that a first person sentence of the given form can have its truth challenged without suggesting the falsity of what I assert that I know. If I am a medical researcher and assert that at last I know that the procedure over which I have labored so long is safe, someone might respond, 'No you don't. You're forgetting that the results of the final major trial haven't yet been announced'.

Again, a person declining slowly into mental illness could point to a sentence in his diary from many years ago employing the first person usage and comment, 'I'm sure that was true when I wrote it, but my competence is so poor these days, I no longer claim to know very much at all and wouldn't write that now'.

Furthermore, a layperson hearing for the first time some potted philosophical presentation of an extreme form of scepticism might respond with irritation, 'Oh come on–I know that I'm not a brain in a vat!' It is strained to suppose that, to employ Clarke's words, the person is staking his reputation for veracity on his not being a brain in a vat.

Moreover, Austin has warned that we should expect that during their development natural languages have increased their usefulness by assigning different jobs to different locutions, and thus presently contain few short synonyms. Yet today English contains the expression, 'I guarantee that', which Clarke seems to treat as synonymous with 'I know that'.

Clarke cites an additional consideration that has been in the philosophical air for some time, namely, that 'I believe' is used to express hesitancy of the kind ruled out by strong endorsement (cf. 1989, p. 25). But such a proposal raises an objection to a JTB analysis only when turned into the stronger claim that this usage is the *only* first person usage of the expression. It is far from obvious that every religious believer who says, 'I believe that God exists', or every juror who says, prior to voting, 'Well, I've listened carefully and I believe that he is guilty', are expressing hesitancy. Roderick Firth once suggested to me that a psychologist might ask subjects at the end of an experiment to list new things that they have come to believe in the course of the experiment, and that subjects who list sentences of the form, 'I believe that P', are not necessarily using them to express hesitancy.

2.2 Beyond Knowing That

A second risk in focusing too quickly and steadily on the discussion swirling about JTB analyses is that one may thereby overlook valuable insights that would have been gained by seeking a unified account extending to analysanda expressed by other forms of sentences employing 'know'. Philosophers have traditionally been willing to take the risk, and to presume that analysanda, for instance, about knowing people and places, about knowing who, what, when, where, as well as about knowing how to do things, can be handled by a mopping-up operation, perhaps in some cases analyzing them partly in terms of knowing that various things are so (see section 11).

2.3 Beyond Adult Humans

A third danger in turning quickly toward discussions of JTB analyses is the risk of adopting an overly narrow, intellectual focus. The process of philosophy involves debate by highly articulate, intelligent persons who aspire to know things in a manner that permits them to grapple with potential challenges and to acquire information and ideas through the testimony of others. Philosophers such as Descartes and Lehrer, who are inspired by the accomplishments of science, will also be much concerned with knowing things in a manner that permits probing for potential weaknesses in people's opinions by means of sustained inquiry. Yet many philosophers (*pace* Descartes) think that an animal such as a dog can know things, e.g., can know that someone is about to appear. Even more people regard young children or infants as knowing some things to be so before being in any position to enter into debate concerning them.

Philosophers sometimes regard such knowledge attributions as metaphorical or as employing a different sense of 'knows that', occasionally labeling the phenomenon involved as 'animal knowledge' in contrast to 'human knowledge'. Describing young children and infants as only having animal knowledge and not human knowledge may seem distasteful, but it is even less appropriate when dealing with certain victims of Altzheimer's or other mental diseases whose intellectual facility has been markedly damaged but whom we speak of as still knowing, for example, that various things happened to them in the past, or that their name is such-and-such. The more intellectually complex that the application of the justification condition in a JTB analysis becomes, the more defenders of the tradition may be pushed toward having to explain why, when speaking about such patients, no conversational warnings seem required that one is making the supposed slide to a second sense or to a metaphorical usage.

2.4 Communicating with Adults

Indeed, Clarke thinks that even the way in which we employ third person usages of 'knows that' regarding healthy human adults fails to carry with it any concern about the justified status of their beliefs. Clarke suggests that in standard communication situations we mainly ask about knowledge for one of the following reasons: (1) We want to find out whether a person is a potential source of information about a state of affairs; (2) We want to learn whether some information or command that we wish to convey could be capably received by a person (cf. 1989, p. 25). Clarke maintains that neither interest leads us to care about the state of belief of the person in question, nor about the rational status of potential states of belief.

It is not clear how these reasons could cover third person questions about dogs, infants, or young children. Moreover, a whiff of concern with justification may creep in when Clarke adds regarding a case illustrating reason (1) that the person "can be said to 'know' in the sense of being a reliable source of information . . .". If we only want accurate information, it will not matter whether the source, in addition, is reliable. The case discussed by Clarke concerns the expert chicken sexer who supposedly discriminates male from female chickens without forming relevant beliefs based on evidence. Related examples in the literature, to be considered later, are most likely purely hypothetical cases, such as telepaths or the boy who is a 'seer' or 'psychic' forecaster in D. H. Lawrence's story, "The Rocking Horse Winner".[2]

A philosopher who is inclined to grant that the preceding examples show that an analysis of knowing should not include the belief condition and perhaps not even the justification condition may at least consider whether a JTB analysis comes close to describing a *species* of knowing possessing special importance in areas where people are expected to participate in cooperative inquiry by submitting assertions to mutual scrutiny and potential debate.

3 GETTIER-TYPE CASES

3.1 Early Gettier-Type Cases

Gettier showed (1963) that with respect to the following examples a JTB analysis is too weak to rule out S's knowing that P1 and too weak to rule out S's knowing that P2:

Coins in the Pocket S justifiably believes about another person, Jones, the unsuspectedly false proposition that F1: 'Jones will get the job, and Jones has ten coins in his pocket'. S recognizes that this proposition entails that P1: 'The man who will get the job has ten coins in his pocket', which S then believes on the grounds of the proposition that F1. Unsuspectedly, not only does S have ten coins in S's pocket, but it is S who is going to get the job.

Brown in Barcelona S has strong evidence for a proposition, which S does not realize is false, namely, that F2: 'Jones owns a Ford'. S picks at random a city name, 'Barcelona', and recognizes that the proposition that F2 entails that P2: 'Either Jones owns a Ford or Brown is in Barcelona'. Not having any idea of Brown's whereabouts, S proceeds to accept that P2 on the grounds of the proposition that F2.

When reacting to Gettier's concerns, Keith Lehrer (1965) formulated an example that is close to Gettier's second one and has provoked many variants in the course of the ensuing debate, so that philosophers have come to speak of a category of 'Gettier-type' examples:

Mr. Nogot Somebody in S office, Mr. Nogot, has given S evidence, E, that completely justifies S in believing that F3: 'Mr. Nogot, who is in the office, owns a Ford'. Evidence E consists in such things as Nogot's having been reliable in dealings with S in the past, having just said to S that he owns a Ford, and having just shown S legal documents affirming it. From the proposition that F3, S deduces and thereby comes to believe that P3: 'Somebody in the office owns a Ford'. Unsuspectedly, Nogot has been shamming and it is someone else in the office who happens to own a Ford.

An even earlier example rather like the coins in the pocket was offered by Bertrand Russell (1948), but provoked little discussion: S has true, justified belief as to the time by looking at a clock that, unsuspectedly, stopped twelve hours earlier.

3.2 The Gettier Problem

Typically, when the flood of literature that emerged from the impact of Gettier's brief paper mentions 'the Gettier problem', this has been taken to be either the problem of finding a fourth condition of knowing which can be added to the analysans in a JTB analysis so as to obtain a satisfactory analysis, or else the problem of finding a correct analysis by some modification of a JTB analysans. Attempts at the latter differ in how radical the proposed changes are, for instance, whether some standard condition needs to be totally replaced by a new requirement.

There is presently no widespread agreement that a solution of either problem has been found.

4 INITIAL RESPONSES TO GETTIER CONCERNING FALSEHOODS

Gettier's having described the coins in the pocket example as a case where S accepts that Pl on the grounds of the falsehood that Fl initially led philosophers to think that the first of the above problems can be solved by adding a fourth condition to a JTB analysis requiring that no false beliefs are formed by S as crucial parts of S's reasoning to P. But Lehrer (1974) described a variant on the Nogot case where S does not care who it is that owns the car but only cares whether someone in the room/group does, and so S reasons that because it is possible that someone else there owns such a make car, less risk of error will be involved in S's accepting the more general proposition that P3 than accepting that F3. Lehrer also proposed a variant in which S directly infers that P3 from the proposition that F1.

Ernest Sosa (1991) has rejected Lehrer's examples by saying that S must see a connection between E and the proposition that P3 in order to be justified in believing that P3, and that the only apparent way to connect them is by way of a statement saying that a specific person or persons are involved in the situation and own a Ford. But even if that proves that such a statement must be thought of by S, it does not show that it must also be believed by S.

Nonetheless, Sosa would still be viewing E as connected with the proposition that P3 by 'reasoning'–in a broad sense–through a false statement, and Gilbert Harman (1973) has proposed to solve the Gettier problem by requiring that no lemmas crucial to S's reasoning to the conclusion that P be false.

Many regard Richard Feldman (1974) as having providing a Gettier-type example free from such false beliefs or lemmas when he described S as reasoning to the proposition that P3 merely from an existential generalization, G, of evidence E. But Sosa might object that because G includes such details as the supposition that someone in the office/group told S that he owns a Ford, S will indeed be seeing a connection between E and the proposition that P3 when S regards the former as justifying the (false) statement that the person who said that he owned a certain Ford is someone who does own that very one.

The following variant may avoid such an objection:

Secondhand Feldmanizing This resembles Feldman's case except that S testifies to S' that E': 'There is someone in S's office who has given S strong evidence in support of the proposition that P3'. Since S' realizes that S has been reliable in the past, S' comes to believe that P3.

Here S´ has no idea who provided the evidence, nor what it was, and so neither believes nor reasons that it was a person evidencing that same person's ownership. Nor need S´ believe that S believes that F3.

Suppose that it is replied that S´ must at least have the false belief that S has not arrived at belief that E´ crucially by means of false beliefs or reasoning involving false lemmas. That response incurs the risk mentioned earlier of an overly intellectual view of knowing. Young children can know things via testimony before reaching such a level of sophistication as to worry over whether the testimony rests in some inferential way on some false belief of the testifier. So let us add to the above example the detail that S is such a young child.

Alvin Goldman (1976) credits Carl Ginet with the following frequently discussed, non-Gettier-type example, in which it indeed would require too much to suppose that S must have an opinion about certain relevant factors:

The Barn Facsimiles S believes that P4: 'Here is a barn', because S sees a barn from the front while driving through an unfamiliar countryside, unaware that people there who wish to appear quite affluent have erected many paper-mâché constructions that look just like the barns in the area from the road.

We do not expect people to reason to perceptual beliefs partly on the basis of beliefs or statements about the incidence of fakes in the neighborhood, and so no falsehoods of the latter sort infect S's coming to believe that P4 as an explanation of S's lacking knowledge that P4.

Risto Hilpinen (1988) argues that it is too strong to require that knowledge is not crucially based upon any false belief. He cites the example of the scientist, Millikan, who believed on the basis of careful research that the electron's charge has the value that he assigned (at least within a certain margin of error), whereas today we regard it as a value close to that one (but falling outside that margin of error). Hilpinen proposes that Millikan nonetheless could on the basis of his false belief come to know some other things to be so (say, that H).

5 DEFEASIBILITY THEORIES

When there is some false belief or lemma crucially involved in a Gettier-type case, S's realization of that fact would have an epistemic impact on S to the extent of making S no longer justified in believing that P. Defeasibility theories might, by way of introduction, be characterized as taking a more abstract perspective on this point and as considering what impact is made on a certain aspect, A, of S's epistemic situation bearing on whether S knows that P by bringing A into a certain relation, R, to some proposition, D, which, unsuspected by S, is true and is of a specific type, T. If the impact alters A in a fashion that prevents S from satisfying a requirement for knowing that P, then D is technically said to 'defeat' the proposition that P or to be a 'defeater' of it, and the possible existence of a truth with such an impact is called the 'defeasibility' of A. More broadly one could say (cf. Shope 1983) that a defeasibility theory offers a fourth condition of knowing that requires a particular truth value for some subjunctive conditional(s) about what would obtain concerning the justification of S's believing or accepting that P if certain hypothetical circumstances were to obtain (and Plantinga (1996) points out other candidates for A, such as S's believing that P being reliably produced, or being produced by properly functioning cognitive faculties, or being appropriately 'coherent').

5.1 Early Defeasibility Theories

The earliest defeasibility theory was offered by Lehrer (1965), who presumed that in Gettier-type examples S does come to believe the relevant falsehood, such as the proposition that F1, or the proposition that F3. On Lehrer's proposal, aspect A is S's being justified in believing that P, type T is being a proposition denying the relevant falsehood, and R is the relation of being supposed to be true by S. Somewhat simplified, his requirement amounts to the following:

> (4a) If S is justified in believing any false statement, that F, which entails that P, then S would be justified in believing that P even if S were to suppose that ~F and to suppose nothing else in addition unless entailed by the proposition that ~F.

This subjunctive conditional mentions a change within the mind of S, and because of this psychologistic phrasing, the resulting analysis is rendered too strong to permit S to know that P5: 'S is not supposing that ~G', in a case where the proposition that G is justifiably believed by S yet unsuspectedly is false. S's supposing that ~G would be something that S would be aware of, and this would stop S from being justified in believing that P5 (cf. Shope 1983). The conditional also rules out the knowledge possessed by Millikan that H in Hilpinen's example.

Moreover, in the case of the barn facsimiles, the troublesome falsehood that the area does not contain a large number of fakes fails to be one that S is justified in believing to be true.

The latter case does not run counter to a subsequent fourth condition proposed by Lehrer (1970), which is close to the following:

> (4b) For any false proposition, that F, if S were to suppose for the sake of argument that ~F, then S would still be justified in believing that P.

Yet the resulting analysis is not only subject to the counterexample concerning S's knowing that P5 (and perhaps to the one concerning Millikan's knowledge), but was revealed by Lehrer and Thomas Paxson (1969) to be too strong to cover the following case:

Demented Mrs. Grabit S knows concerning his acquaintance, Tom Grabit, that P6: 'Tom stole a book from the library' since S saw Tom do it. But, unsuspected by S, Tom's mother has vowed that V: 'Tom's identical twin was in the library at the time of the theft and Tom was thousands of miles away'. Nonetheless, the twin is merely a figment of the demented imagination of Mrs. Grabit.

It is controversial whether the example constitutes a genuine counterexample (cf. Shope 1983, p. 50). Nonetheless, the case does share one feature with the genuine counterexample concerning the case of the barn facsimiles, namely, the detail that the relevant false statement is so far from actually crossing the mind of S that it is strained to speak of S as believing it.

5.2 Barker's Defeasibility Theory

Some subsequent defeasibility theories took account of the latter concern in the course of attempting to avoid treating knowledge as being precluded by a merely 'misleading defeater', such as the proposition that V. One such defeasibility condition was proposed by John A. Barker (1976):

> (4c) There is some way that any other true proposition besides the proposition that P could come to be justifiably believed by S without destruction of S's original justification for believing that P.

On this view, a defeater need not contradict what is believed by S, but when it is misleading, the defeater could come to be believed in a way that accompanies it with belief in true information revealing why it is defeating, e.g., S could come to believe that V together with believing that the twin is only a figment of the imagination. In that sense, knowledge is purportedly 'indefinitely extendible'.

Barker's condition is phrased so as to deal with a further limitation of previous defeasibility conditions, namely, that they permitted the consequent clause in their subjunctive conditional to be satisfied merely because the satisfaction of the antecedent would provide S with a new basis for being justified in believing that P. For instance, consider a variant of the original Nogot case where S, after coming to believe that someone in the office owns a Ford, infers, a là the Brown in Barcelona example, that P7: 'Either someone in the office owns a Ford or Mr. Nogot does not own the Ford in question'. A new justification for believing that P7 would be provided by S's coming justifiably to believe that Nogot does not own that Ford.

This advantage, however, may prevent Barker from allowing Millikan's knowledge that H, since we seem to fix on a new justification for his believing that H when we imagine Millikan as coming to share our present opinion that the electron's charge is quite close to what he concluded, and to have that as a crucial part of his basis for believing that H. For our opinion contradicts his original false belief, and he might have lacked even the disposition to share it.

Moreover, because condition (4c) is still psychologistic, it faces counterexamples such as the following:

> *The Introspective Inventory* Reflecting on what S's own opinions are, S knows that P8: 'S does not believe that R', where unsuspectedly, it happens to be true that R.

We may presume that if the truth that ~R were to be justifiably believed by S, then S would be aware of believing that ~R and would no longer have any justification for believing that P8.

A further concern is Peter Klein's point (1976) that the addition of some truths to S's evidence would nullify the proper effect of a genuine defeater, e.g., the truth in the barn facsimile case that S is not looking at any of the paper-mâché replicas.

5.3 Klein's Treatment of Defeasibility

A hint of moving in a certain direction away from psychologistic defeasibility conditions was provided by Marshall Swain (cf. 1974, p. 22), and explicitly developed by Klein (1976). We no longer regard aspect A of S's epistemic situation as S's actually believing various things, but instead treat A as S's being justified in believing various things, which remains neutral as to whether S actually believes them. The challenge is then to find a way to characterize the type, T, of truth whose impact is to be considered on such justification as a means of distinguishing between misleading and genuine defeaters.

Klein's initial efforts ran into difficulties (cf. Shope 1983, pp. 70–72). But his most careful attempt, which has been as fully developed as any defeasibility theory, appears in his book, *Certainty* (1981), and has recently been updated (cf. Klein 1996).

In the book, Klein asks us to consider not states of believing but what he calls confirmation relationships between propositions. If the proposition that Q confirms the proposition that N, this is to be symbolized as 'QCN' and failure to confirm is symbolized as '$Q\not{C}N$'. Klein construes the justification condition in a JTB analysis as entailing a confirmation relationship, e_nCP, where the proposition that e_n is believed by S and states evidence that confirms that P. Klein asks us to consider 'e-chains', defined as sequences of propositions ending in the proposition that e_n, in which any member, e_i, is believed by S and confirmed by the prior member, e_{i-1}. Klein presumably wishes all these links, as well as the link between the proposition that e_n and the proposition that P, to involve direct confirmation, i.e., the beginning of the link is not to confirm the end only through confirming another proposition which then confirms the end.

Klein treats a truth that T as being a 'direct' defeater when its conjunction with some member, e_i, of the sequence fails to confirm that next member, for instance, when $(T\&e_{n-3})\not{C}e_{n-2}$, or when $(T\&e_n)\not{C}P$. A truth, T, will be an 'indirect' defeater when it is the first member of a sequence of propositions, which Klein calls a 'd-chain', each member of which (directly) renders the next member something that S would be reasonable in believing, that is, each member of which is a good reason for thinking the next member to be so, at least in conjunction with other beliefs–whether or not they are true–actually held by S, and the last member of which is a direct defeater. Klein proposes that a misleading defeater will be one that is indirect and such that some falsehood in the d-chain is rendered reasonable for S to believe independently of the involvement of any actual false beliefs of S as conjuncts of parts of the chain prior to the falsehood. Given this background, the fourth condition of knowing becomes the following:

(4d) If anything is a defeater of P (for S) then it is a misleading defeater.

For example, Mrs. Grabit's demented avowal that V is an indirect defeater generating a d-chain leading to the direct defeater that D: 'Someone within the library at the time of the theft was such that S could not discriminate between that person's committing the theft and Tom's committing it'. For suppose that the proposition that e_n states S's visual evidence and other relevant evidence which S possesses confirming that P6. Then $(D\&e_n)\not{C}P6$. Yet any d-chain leading from the proposition that V to the proposition that D contains some falsehood (for example, that Tom has an identical twin) not resting on any falsehoods believed by S.

In contrast, in a case where the following is true: 'Tom has an identical twin who was in the library at the time of the theft whom S could not have discriminated from Tom', this defeater is not misleading on the above account. For a similar reason, Klein's defeasibility condition gives the correct verdict concerning the barn facsimile example. And there will be obvious nonmisleading defeaters in the other Gettier-type cases mentioned earlier, such as the proposition that Jones will not get the job, or the proposition that the person who provided the evidence about Ford ownership is pretending to own one.

Klein's approach is also independent of whether the conjunction of a defeater with evidence possessed by S generates a new justification (or better, a new e-chain) leading to the proposition that P. I had overlooked this point when offering (1998) as a counterexample a variant of the Nogot case in which (1) after coming to believe that N: 'Nogot owns a Ford', S sees, as luck would have it, a true sentence expressing a disjunction, D, which, unsuspected by S, is such that any defeater of N entails one or another disjunct of D that S is not presently justified in believing; (2) on the basis of believing that N, S comes to believe that P9: 'N or D'; (3) it is true that N because Nogot is lucky enough to win a Ford in a raffle while in the company of S. I suggested that in this case of Lucky Nogot the defeaters of the proposition that N fail to be defeaters of the proposition that P9, because they support its second disjunct, while there need be no other truths that defeat P9.

Yet since Klein does not express his defeasibility condition in subjunctive terminology, there should have been no temptation to regard the above example as conflicting with his account. All that matters for defeat is that with respect to each actual e-chain, the conjunction of some proposition that is a non-misleading, effective defeater with some member of the e-chain creates a conjunction that does not confirm the next member of the e-chain.

5.3.1 Plantinga's Objections to Klein

A different flaw occurs regarding one of Alvin Plantinga's attempts (1996) to provide a counterexample to Klein's account. The general form of the example involves S's mistakenly believing that ~N, the denial of some truth, on the testimony of an excellent authority, and where we consider anything whatsoever that S does know to be so, say, that 2+1=3. Plantinga claims that a true statement, K, of the following form is counted by Klein's definition as a genuine defeater, where the arrow expresses material implication: '(~N –> (N –> ~(2+1=3)))&N'. Plantinga maintains that K is the initiating defeater and '~(2+1=3)' the effective defeater, and that one d-chain begins with the two conjuncts of K taken together with the proposition that ~N. But Plantinga has obtained an inconsistent conjunction by conjoining the proposition that ~N with the two conjuncts of K. And since an inconsistency is not a reason for thinking anything to be so, no d-chain begins in the manner that Plantinga maintains.

Nonetheless, a second example offered by Plantinga does create trouble for Klein's account. In order to sketch its general form, let us consider any proposition that A, mistakenly believed by S, and anything that S knows to be so, say, that P. Suppose that S deduces from the proposition that A, and comes to believe, that C: '~A –> ~P'. Plantinga charges that the proposition that ~A is both an initiating and effective defeater of P. He is envisioning formation of a conjunction of the proposition that ~A with the proposition that C so as to directly defeat the proposition that P. The case does create trouble for Klein because S already believed the falsehood that C, and so the latter proposition is available as a conjunct in a d-chain.

But perhaps we should take it as a tacit point in Klein's explanation of his views that when considering whether S knows that P, we are to be concerned not only with the history of confirmation in e-chains running backwards from the link which ends in the proposition that P, but also with history of e-chains, if any, ending in other things that S believes when the latter are part of a d-chain. Since in the proposed counterexample, the proposition that ~A is inconsistent with the very evidence that had led S to believe that C, why should we suppose that any reason for S's thinking something to be so is created by conjoining the proposition that ~A with the proposition that C?

Even if this consideration blunts the force of Plantinga's example, a variant of it is waiting in the wings, where S believes that C on the basis of nothing at all. This trouble might be avoided by modifying Klein's defeasibility condition so that in forming d-chains from an initiating defeater, we can only draw upon the assistance of other present beliefs of S whose content is something that S is presently justified in believing. But even this will not block another variant of the example in which what earlier justified S in believing that C was the testimony of some expert that C.

One might consider dealing with Plantinga's example by changing Klein's analysis so that d-chains are no longer permitted to employ any false beliefs of S. The only reason that Klein had granted such permission in *Certainty* was in order to see how many sceptical claims about knowing could remain uncontested without making the possession of at least some empirical knowledge impossible. Yet no actual sceptic has pressed the demand that every false belief of S is a candidate for membership in d-chains, let alone has provided a rationale for that demand.

5.3.2 Further Problems Concerning Klein

Whether the case of the introspective inventory shows that Klein's analysis of knowing is too strong depends upon whether his focus on confirmation relations among propositions avoids troublesome psychologism.[3] The latter concern returns by the back door when Klein analyzes 'x confirms y' as follows: according to the rules of confirmation or evidence, it is permissible to infer y from x (cf. 1981, pp. 25, 34, 74). Presumably, inferring is an activity of coming to believe or to accept one or more things on the basis of having already come to believe or to accept various things. In the example at hand, we are asked to consider how rules of confirmation relate to inferring something at the present time, t_j, from the conjunction of some proposition, that R, with e, whatever evidence there may have been at t_{j-n} for S's having believed that P8: 'S does not believe that R'. So in order to apply the present analysis of confirmation, consider beginning an inference from believing or accepting that R&e, and moving to believing or accepting that ~P8. When the premise of this inference is true, the conclusion is bound to be true. Why is this not enough to show that (R&e)C~P8, which would entail that (R&e)ЄP8, and thus make Klein's analysis too strong to allow that S knows that P8?

Klein may reply that he has also specified that whether xCy "depends upon whether x, if true, is a sufficiently good guarantee of y" (1981, p. 74). In the above situation what constitutes a guarantee of its being true that ~P8 is not the truth of the proposition that R&e, but the fact that the believing or accepting of that proposition is the start of an inference. When Klein speaks of degrees of goodness of a guarantee he may be concerned with what philosophers sometimes call varying degrees of the likelihood of the truth of a conclusion.

So let us go on to consider a variant of the above example that might be said to involve an introspective inventory of beliefs about beliefs. Let the proposition that R be the true proposition, 'S does believe that K', where S happens not to realize that the belief that K is present in his mind, perhaps because of the distasteful content of the proposition that K. Then the truth of the proposition that R fails to make quite likely the truth of the proposition that P8, so that R∈P8. In that case, is it at all obvious that (R&e)CP8? A correct analysis of knowing should not produce the verdict that S's knowing that P8 is not a clear case of knowledge.

It might appear that one can defend Klein by admitting that (R&e)∈P8 while maintaining that this concerns an *indirect* relationship, so that the direct defeater of P8 is the proposition that ~R, which is false and thus will be a misleading defeater. But the truth of the proposition that R directly makes it unlikely that P8. Consider certain times when one has difficulty remembering some detail, and at least holds back from judging that one's memory of the detail has been lost, even though one does not go so far as to believe that the memory is still possessed. Given one's introspective dispositions, the continued possession of a memory makes it unlikely that one will believe it to be lost, independently of making it likely that one will believe it still to be present.

More recently Klein has allowed that sometimes it is epistemically permissible for S to believe that P, and even that S epistemically ought to believe that P, without there being any proposition believed by S confirming that P. This could happen for a 'basic' belief, e.g., that 2+1=3; e.g., that it looks to S exactly as if something pink is present. In order to accommodate this point, Klein incorporates an element from some reliability theories (to be surveyed below) and requires that such beliefs be "reliably produced" without there being a genuine defeater–in a now extended sense–such that adding the defeater to S's present beliefs moves the belief that P "too far" from being justified (1996 , p. 127).

In a way, this broader perspective was anticipated in *Certainty* when Klein spoke of confirmation as concerning what is permissible to infer. Speaking of what *ought* to be believed, of what *ought not* to be believed, and of what it is *permissible* to believe when various states of affairs obtain is speaking of relationships among propositions. Since philosophers sometimes call such evaluative expressions 'deontological', let us simply speak of a 'D-relationship' and let 'xDQ' symbolize that the occurrence of state of affairs x makes it something that one ought to believe that Q. And let us allow that part of some relatum entering into a D-relationship may be the occurrence of a state of affairs consisting in some propositions' standing in a C-relationship to other propositions, while other relata in a D-relationship might be the occurrence of other types of states of affairs, e.g., its looking to S exactly as if something pink is present; e.g., S's belief that P being reliably produced. Klein could then find an expanded definition of a defeater to cover how the combination of x with the occurrence of the total additional state of affairs relevant to reliability might fail to make it something that S ought to believe that Q.

But this procedure would carry over from the earlier definitions concerning defeaters the fact that a true proposition that d counts as a genuine defeater independently of many other truths that hold about S which do not form any part of e-chains leading to the proposition that P and are not part of the d-chain through which the proposition that d defeats.[4] Thus the amended account may appear vulnerable to Plantinga's objection (1996) that in hypothetical cases, where S is not designed like the typical human, there are additional truths thanks to which S knows that P even though some true proposition that d genuinely defeats. For instance, Plantinga imagines a hypothetical S who is designed so as to be unable to look at the barn facsimiles because paper-mâché makes S's eyes water when S turns them in its direction, or because a guardian angel is assigned to prevent S from looking at such fakes. But Klein did not exclude the possibility that the rules of confirmation/justification are relativized to types of inquirers, and perhaps would maintain that the difficulty of articulating them for every imaginable type of knower is not a flaw in his analysis of knowing but only a limitation in our ability to articulate a complete theory of confirmation/justification.

Since Klein's analysis of knowing places no causal requirements on the presence of S's belief that P (cf. 1981, pp. 45 ff., 150), it allows Millikan knowledge that H insofar as a undefective confirmation chain runs from Millikan's true beliefs concerning the evidence he had (which led him to his false belief) through the true proposition (which Millikan did not believe) that the electron's value is close to the one that he assigned, and thence to the proposition that H. But a number of philosophers complain that there must be some restriction in an analysis of knowing about the status of S's believing that P; e.g., concerning its originating or sustaining causes; e.g., concerning how S believes it to be related to evidence. After all, when Watson simply guesses that the culprit is so-and-so, yet without any appreciation of the fact that the evidence confirms it, this does not amount to knowledge.

Klein has discussed the manner in which his account in *Certainty* explains the fact that people's intuitions disagree concerning whether knowledge is possessed in a set of cases including ones that display what philosophers call the 'social aspects' of knowing, and that are not usually regarded as Gettier-type cases. A frequently discussed example was provided by Harman (1968):

The Newspaper S believes a true, bylined report in a generally reliable newspaper that a famous civil-rights leader has been assassinated. The report was written by a reporter who was an eye-witness. Unsuspected by S, those surrounding S do not have any idea of what to think since they have additional information consisting in later news reports to the contrary, which they do not realize were due solely to a conspiracy of other eye-witnesses aimed at avoiding a racial incident.

Klein points out that intuitions are divided as to whether S does fail to know that P10: 'The civil-rights leader was assassinated'. Klein diagnoses this case and all others in which he thinks that intuitions are divided as to whether S has knowledge that P in the same fashion, namely, as being due to a disagreement over the role played by the defeater that describes S as lacking certain crucial information bearing on whether or not P. Some may think that this defeater gives rise to a d-chain containing the false proposition that the civil-rights leader was not assassinated, and does so without involving any false beliefs of S. Since the false proposition in question is a misleading initiating defeater, Klein's analysis upholds the intuition that S knows that P10. In contrast, those who regard the defeater as itself both an initial and effective one will find their intuition upheld that S does not know that P10.

But sometimes a conflict of intuitions is likely without being explicable in Klein's fashion. Consider the following variant on one of the above examples:

The Introspective Inventory of Prejudices Reflecting upon the issue of what S believes, S comes to believe the truth that P11: 'S does not believe that L', where the proposition that L is, unsuspected by S, some racist, sexist or otherwise prejudiced opinion found in many areas of the world.

We can adjust the details of the case, such as the incidence of self-deception in people who judge whether or not propositions like the proposition that P11 are true of themselves, so that when combined with any introspective evidence possessed by S regarding the proposition, the conjunction fails to confirm that P11 but does not constitute strong enough evidence to confirm the falsehood that ~P11.

So Klein will need to explain conflicting intuitions about this case in a new way, perhaps as a disagreement over whether the social facts mentioned do amount to a defeater. Since this means that he has no unified account of cases that provoke conflicting intuitions, it is of interest how other epistemologists characterize the nature of such cases and of the social aspects of knowing.

5.4 Pollock's Defeasibility Account

One way in which a defeasibility analysis proposed by John Pollock (1986) differs from Klein's is by avoiding reference to d-chains. A full exposition of Pollock's analysis is complex but he initially stated part of it by means of the following subjunctive conditional: (C) 'There is a set, X, of truths such that, given any more inclusive set, Y, of truths, necessarily, if beliefs in the truths in Y were added to the set of S's total beliefs, with any beliefs S has in their negations being removed, and S were to believe that P for the same reasons, then S would still be justified in believing that P' (cf. p. 185). Pollock is hoping to deal with levels of complexity concerning the manner in which, roughly put, the impact of a defeater is sometimes cancelled by the impact of a further truth, whose cancelling impact itself may get cancelled by another truth, and where in some cases this alternation continues even further.[5] For instance, the impact of truths that Klein offered as illustrations of misleading defeaters is presumed to be blocked once the antecedent of (C) is satisfied, since that makes S believe the denial of the falsehoods that figure in what Klein calls d-chains generated by those defeaters.

Pollock adds a fifth condition of knowing covering its social aspects, namely, that S's believing that P would not be defeated by the set of all truths that S is 'socially expected' to believe when true.

But Pollock's analysis may be too strong to allow Millikan's knowledge that H if the latter's reasons crucially involve a false belief. Moreover, whether (C) can deal with a variant of the case of the introspective inventory in which S knows that P12: 'S believes that F' and it is unsuspectedly false that F depends upon the nature of introspection and upon what determines the truth values of conditional statements.

Some philosophers think that in such an example, the mental state of S's believing that F is at least part of S's reasons for taking that mental state to be present, i.e., think that a relevant member of set X is the truth that P12. But then it is impossible to satisfy the antecedent of (C), which requires removing S's believing that F because of the addition to the situation of S's believing that ~F, yet also requires retaining S's believing that F so as to have S believe that P12 for the same reasons. If one treats conditionals with impossible antecedents as false, then Pollock's analysis is too strong to allow S to know that P12.

But some philosophers, including Pollock (1992), treat all conditionals with impossible antecedents as true, thereby permitting S to know that P12. Nonetheless, I have argued (1998) that this move renders Pollock's account too weak to rule out knowledge in the following case: S believes a justified but imperfect theory that supports the following propositions, which S does not suspect to be false: A: 'Everyone has memories of early childhood;' B: 'Anyone who has memories of early childhood believes that A'. Not by introspection but on the basis of believing the latter propositions S arrives at the true belief that P13: 'S believes that A'. It is impossible to combine believing that A and believing that B with believing the truths that ~A and that ~B. But S does not know that P13.

Yet since the same type of inappropriate result would occur regarding any truth that S justifiably believes at least partly on the nonsuperfluous basis of believing some falsehood, we might question whether Pollock did formulate (C) as he desired. He wrote that its content was guided by an ethical analogy, where S has a 'subjective obligation' relative to S's beliefs but where S's 'objective obligation' depends upon correcting any of those beliefs that are false and upon what other propositions are true that S has not considered. But the closest epistemological analogue would involve conditional (C′): 'There is a set, X, of truths such that, given any more inclusive set, Y, of truths, necessarily, if beliefs in the truths in Y were added to the set of S's total beliefs, with any beliefs S has in their negations being removed, and S were to believe that P *for whatever of the same reasons still remained*, then S would still be justified in believing that P'. This condition is indeed not satisfied in the case of failing to know that P13. But, unfortunately for Pollock, it also is not satisfied in the example of knowing that P12, at least not if part of S's reason for holding the belief that P12 is S's believing the falsehood that R.

So it is only fair to note that Pollock did reformulate his analysis in an 'official definition' without the use of any subjunctive conditional (cf. 1986, pp. 188–9). He instead spoke of potential 'arguments', letting that term mean a sequence of mental states ending in believing or in disbelieving, with each state being 'based upon' the presence of the former member of the sequence, but where we allow some states to be nondoxastic perceptual or memory states. The full analysis is complex, but a crucial disjunct in one of its conjuncts is the possibility (N): 'There does not exist an argument from the combination of belief in some members of the set of truths not actually believed by S and actual basic mental states of S that defeats S's reason, R, for believing that P [i.e., that supports a proposition that D such that the combination of R with S's believing that D fails to count as a reason for S to believe that P (cf. 1986, p. 38)]'. Indeed, (N) is satisfied with respect to P12, inasmuch as no 'argument' can begin from the impossible combination of the mental state of S's believing that F (assuming that this is part of S's introspective reason for believing the state to be present) with S's believing that ~F.[6]

Nonetheless, since it is controversial what the reasons are for introspective knowledge, it is worth noting that Pollock's account still seems to rule out Millikan's knowledge that H.[7]

6 CAUSAL THEORIES AND RELIABILITY THEORIES

Causal theories analyze S's knowing that P by proposing that some causal relationship holds between S's believing/accepting that P and the occurrence of the state of affairs corresponding to the proposition that P. Symbolizing the latter state of affairs by 'P*', we may say that a causal theory may require, for instance, that the occurrence of P* causes S's believing that P; or that there is a common cause of this belief and the occurrence of P*; or that there is an evolutionary explanation of something's causing the belief in the presence of P*.

Causal theories are distinct from reliability theories to the extent that examples of the latter are formulated in terms of relationships more broadly characterized than as involving a causal relationship. Such a reliability theory holds that S's cognitive or epistemic states (not characterized by mention of knowing anything) are such that, given further characteristics of S,[8] it would be the case that P; or it is nomologically necessary that P; or it is in a specified way probable or likely that P. Defenders of either type of theory differ as to whether they are offering a fourth condition of knowing, or, more radically, are offering the causal condition as a replacement for the justification condition in a JTB analysis.

It has been difficult to see how a causal theory could be applied to knowledge of pure mathematical or logical truths, where the obtaining of P* does not seem susceptible (*pace* Plato) to causal relationships (but cf. Carrier 1976). Lehrer has questioned (1990) whether even an ordinary perceptual belief that a certain object is present is caused by or explained in any clear probabilistic way by the presence of the object. He has also (1979) challenged causal theorists to develop a view that is not too weak to rule out knowledge in the following example:

Tricky Mr. Nogot This is like the original Nogot case except that Nogot has a compulsion to trick people into believing truths by concocting evidence that is misleading in the manner that E was misleading in that case, and Mr. Havit's owning a Ford causes Nogot to realize that P3: 'Someone in the office owns a Ford'.

If we imagine in addition that the compulsion is highly specific to information about automotive facts regarding people in the office, then a probabilistic explanation connects P3* with an interlocutor of Mr. Nogot coming to believe that P3.

6.1 Conclusive Reasons Analyses

One variety of a reliability theory is a 'conclusive reasons analysis'. Some examples of this type of analysis require that the reasons for S's believing that P are such that in S's circumstances, if it were not the case that P then S would not believe that P.

Others require that there is some subset of existing circumstances that are logically independent of the truth of the proposition that P, such that in them if it were not the case that P, then S would not believe that P, or at least such that if it were not the case that P then S would not have the reasons that S does have for believing that P (for discussion see Shope 1983).

6.1.1 Nozick's Conclusive Reasons Account

An instance of a conclusive reasons account is Robert Nozick's account (1981), which includes a requirement that S 'tracks the truth' in the sense that the following subjunctive conditionals are true: (N1) If it were the case that P and S were to use only the belief-forming method that S did use–if any–concerning whether or not P then S would believe that P; (N2) If it were not the case that P and S were to use only the belief-forming method that S did use–if any–concerning whether or not P then S would not believe that P.

Condition (N2) is violated, for example, in Nogot cases where S reaches the conclusion that P by reasoning through a false belief or lemma that Nogot owns a Ford.

Nozick admits that he must drop the second condition in examples where the proposition that P is a truth of pure mathematics or logic or some other necessary truth. Moreover, a number of difficulties affect the analysis. For instance, like a number of other reliability accounts (for discussion, see Shope 1983, p. 137), Nozick's analysis does not explain why intuitions are divided concerning the newspaper example, since Nozick apparently treats S's belief-forming method in that case as forming beliefs on a basis that includes what S read. Nozick writes, "if he had heard the denials, he too would have believed them, just like everyone else" (1981, p. 177). Such a diagnosis also improperly rules out knowledge in some cases where defeaters are misleading.

Again, consider a hypothetical case where the following holds, P14: 'It is true of some of those beliefs that S has concerning beliefs that S might not have them'. It is not clearly impossible in our world that S somehow knows that P14. But conditional (N2) is not satisfied for such a person.

Ernest Sosa (1996, p. 276) faults Nozick's analysis for not permitting us to admit that typically when S knows that P it will also be true that S knows that P15: 'S does not believe falsely that P'. Even if S tracks the truth concerning the former proposition, condition (N2) is not satisfied concerning the latter. S would still believe that P15 even if it were the case that S was believing it falsely. (For further discussion of Nozick see Luper-Foy 1987 and Shope 1984.)

6.2 Further Reliability Accounts

Reliability theorists began moving in the direction of accounts to be considered in the next section when they considered citing facts about the recent causal workings of S's mind in encounters with information as a way of explaining why a reliable process yielding the belief that P should have led to that belief's being true on the present occasion (cf. Morton 1977) or when they took justification to concern processes involving operations of cognitive faculties (cf. Goldman 1979). Yet it is not clear how sticking only to a causal characterization of such matters can provide a way to show that an unreliable process is involved in the following example (cf. Olen 1976) of a lack of knowledge:

The Sports Fan's Surmise On a quiz show, S cannot remember who won a certain award but makes a correct guess on the basis of fragments of recalled information and an attempt to think of what might best explain them.

Some reliability theories might try to deal with knowledge of necessary truths by avoiding the spareness of Nozick's conditionals and instead considering the cognitive processes through which beliefs are formed, perhaps counting as one example of such a process the exercise of reason vis à vis necessary truths. But reliability theories have been bedeviled from the start by what is called 'the generality problem', namely, how to specify the boundaries in the analysis for the type of situation within which a belief-forming process reliably leads to true beliefs (and, perhaps, does not lead to false beliefs) (cf. Goldman 1976). If the boundaries are set too broadly, e.g., if we consider formation of perceptual beliefs on the surface of the earth, then perception counts as reliable and the theory does not explain why myopics fail to know certain things about what they see far away when by chance those things are true. If the boundaries are set too narrowly, the condition may be too easy to satisfy.

6.3 Goldman's Theories

Alvin Goldman has been one of the most prominent reliability theorists, but his developing views are rather complex (for fuller discussion see Shope 1983 and 1989). He has been able to avoid certain counterexamples to his earliest form of such a theory by coming to require both 'local' reliability of a belief-forming procedure, i.e., reliability in the present context of S's believing that P, and 'global' reliability, i.e., reliability for all or at least many uses of the procedure, not just its use in forming the belief in question (cf. 1986). Goldman requires that the procedure sufficiently often produces true beliefs or inhibits false beliefs in actual situations and in relevant counterfactual situations, technically called 'relevant alternatives'. After facing counterexamples to his initial efforts to characterize the relevant alternatives with respect to global reliability, Goldman proposed (1986) that relevant alternatives are the alternatives that are consistent with our general beliefs about the actual world. I have argued (cf. 1989) that this makes his account too strong to permit knowledge in the following example:

Fortunate Beauty S justifiably believes the true statement that P16: 'Beauty is present', on the basis of how Beauty looks, and has acquired a perceptual schema of her through an ordinary learning process. Yet Beauty is fortunate that no mentally disturbed individual has just recently, unsuspected by S, disfigured her in a way that would prevent S's recognizing her on the basis of her visual appearance.

It is consistent with our general views of the world that such disfiguring might occur in many various ways leading to many varying details regarding the present appearance of Beauty, so the relevant truth ratio may not be sufficiently high to permit Goldman to show that S knows that P16.

More importantly, Goldman's handling of Gettier cases, for instance Brown in Barcelona, is problematic. He says that the counterfactual situation in which Jones does not own a Ford and Brown is not in Barcelona cannot be discriminated by S's belief-forming procedure from the actual state of affairs (cf. 1986, pp. 46–47, 54–55; and cf. Audi 1993, p. 218). But such a discriminative ability is also missing in some genuine cases of knowing, such as the case of Mr. Havit, who is not shamming when in S's office he offers evidence of his Ford ownership akin to E in the original Nogot case, and this leads S to believe that someone in the office owns a Ford. While S is obtaining that evidence from Mr. Havit, S cannot discriminate the actual situation concerning someone in the office owning a Ford from a (hypothetical) situation where a terrorist or meteorite has just blown up Havit's Ford far away.

7 SOSA'S VIRTUE EPISTEMOLOGY

The shortcoming of reliability theories has been diagnosed by some as due in part to their failure to nest reliability considerations within a broader requirement concerning a belief's relating to cognitive factors that are manifesting one's intellectual virtues in certain ways or are functioning properly in certain contexts.

Ernest Sosa has sought to characterize intellectual virtues in relation to the goals of believing truths and not believing falsehoods, and has incorporated a reliability aspect when speaking of the likelihood of attaining these goals through the exercise of such virtues:

> At a minimum, for S to believe P at t out of intellectual virtue, there must be a field of propositions F such that P is in F, and there must be conditions C such that S is in C at t (with respect to P) and such that S is nomologically (but not tautologically) likely to be right if S believes a proposition in field F when in conditions C (1991, p. 278).

Properly interpreted and refined, this requirement supposedly leads to an account of the belief that P being prima facie justified out of intellectual virtue or through the operation of a cognitive faculty, so as to provide part of an analysis of S's knowing that P.

The analysis mentions likelihood a second time when it requires that the justification is not overridden. "Possible overriders of such justification would have to be wider intrinsic states of the subject diminishing significantly the probability that the belief in question be true" (1991, p. 241). Sosa offers an example in which one believes that a pink surface is present because it looks to one exactly as if there is a pink surface before one, but a wider state of oneself also includes one's having accepted testimony whose whole content is that there is no pink surface present. Then the likelihood that one's belief is true relative to this wider state is significantly less than the likelihood relative to one's visual experience. So the belief is not epistemically justified and one doesn't know–not even if the testimony is a lie. (Presumably Sosa wants us to consider the impact of the totality of intrinsic states of the believer, since widening even further the state just mentioned might bring in another factor that counteracts the impact of the acceptance of the testimony, e.g., one's also believing that the testifier has repeatedly lied about such matters.)

Sosa does not discuss at length how such estimates of likelihood are to be understood, nor how great must be the lowering of likelihood before it becomes significant in the above way. But he seems to be concerned with an objective propensity approach to probability. During a debate with Alvin Plantinga, Sosa suggests that all that is required for a cognitive faculty to work properly relative to goal G in environment E "is that it be ø'ing where ø'ing in E has a sufficient propensity to lead to G" (1996b, p. 261). Taking G to be the goal of believing truths and not believing falsehoods, Sosa goes on to propose that, at least concerning object-level knowledge, "in the circumstances one would (most likely) believe P iff P were the case–i.e., one (at least probabilistically) tracks the truth . . ". (1996b, p. 267).

Yet the latter comment seems to contain reference to likelihood at an inappropriate place. It appears to present probabilistic tracking as involving the following odd conditionals:

(I) If it were the case that P then there would be a sufficient objective propensity of one's believing that P.
(II) If there were a sufficient objective propensity of one's believing that P then it would be the case that P.

But Sosa's earlier comment about sufficient propensity leads one to expect that he wishes to say that in the circumstances the following is most likely: one would believe that P iff it were the case that P. The circumstances here include, as the instantiation of 'ø'ing', the manifesting of the faculty in question in the formation of one's belief that P. Perhaps a sufficient objective propensity for such manifesting to yield true belief implies that the following conditional is (most likely) true: 'If one were to believe that P then it would be the case that P'.

This understanding of Sosa's intent is supported by two further considerations: (A) He says that the probabilistic tracking is required to occur "because P is in a field of propositions F and one is in conditions C with respect to P, such that believing a proposition in field F, while one is in conditions C with respect to it, would make one very likely to be right" (1996b, p. 267); (B) Shortly after quoting his own probabilistic wording of the biconditional, Sosa drops mention of likelihood when saying that his definition of S's tracking the truth is the following: "S would believe P iff P were the case" (1996, p. 274). Sosa calls this a Cartesian tracking requirement, in order to stress its difference from Nozick's proposal, and symbolizes its subjunctive conditionals as follows:

(C1) $B_S(P)\text{-}{>} T(P)$.
(C2) $T(P)\text{-}{>} B_S(P)$ (cf. 1996a, p. 276).

Yet the preceding comments by Sosa do not explain why he is committed to (C2) or even to its likelihood. Nor does he explain this in the course of trying to avoid using Plantinga's conception of a design plan when characterizing intellectual faculties. Sosa says that a faculty is a special "ability, power, or capacity to accomplish" a normally desirable sort of thing (described at a certain level of generality) and that an intellectual faculty is possessed by a whole intelligent being (1996a, p. 273). Further, functioning, or performing a function, is "a special, distinctive activity that is desirable or at least desired" (1996a, p. 273). At this level of abstraction, relative to some goal, G, what Sosa calls a distinctive activity could also be called a manifestation, M, by the intelligent being of the ability/power/capacity to attain G, which is a means to G on certain occasions or in certain circumstances–for simplicity let us suppose only on an occasion of type K. We would expect Sosa to treat one's functioning properly in producing manifestation M as amounting to the following conditional's being true or at least likely to be true:

(F) If K were to obtain then one would attain G.

Learning about an ability/power/capacity to do A involves coming to understand occasions for its full manifestation, i.e., for its being manifested by the possessor's doing A, and this pertains to understanding conditionals of form (F), for instance,

(F1) If one were in C with respect to propositions in field F then one would believe some true proposition in F.

But such understanding need not always involve expecting the (likely) truth of a conditional of the following form: 'If the ability/power/capacity to do A were to be manifested then it would be fully manifested, i.e., A would be done', e.g., the truth of the conditional

(V) If one were in C with respect to propositions in field F and one were to believe a proposition in F then that belief would be true.

Yet the (likely) truth of (V) is for Sosa an aspect of the presence of a faculty, and this view seems to lie behind his acceptance of (C1).

Consideration (A) above might seem to treat the desirability of having true beliefs as resembling the desirability of having mild colds. It is desirable for the colds that one suffers to be mild. This is not to say that it is desirable to acquire mild colds, which implies that it is worthwhile to seek them out. But Sosa does wish to think of the desirability of epistemic goals as akin to the desirability of food and warmth, as something desirable to acquire and to sustain. Yet it is difficult to see how any of Sosa's remarks canvassed above quite captures this point. Condition (F1) does not, given the role of 'some' in the statement.

7.1 Plantinga's Objections to Sosa

We may eventually make further progress by noting that Sosa has required that conditionals (C1) and (C2) hold *because* conditional (V) holds. Plantinga overlooks this important detail when charging that (C1) and (C2) are insufficient to rule out malfunctioning. He considers an example concerning a necessary truth, e.g., Gödel's first theorem, and supposes that S only believes the theorem because S suffers from a malady that causes S to believe any mathematical statement of a certain level of complexity, even though S displays Cartesian tracking of the truth of Gödel's theorem (cf. 1996, p. 370).

But Plantinga has overlooked the fact that Sosa has in mind an ethical analogue between S's knowing that P and something's being the morally right thing for a given person to do in a given situation. What is initially relevant to determining the latter is that one's circumstances fall under some description, e.g., 'one has promised to do something', forming part of a moral principle, e.g., 'If one has promised to do something then one has a moral obligation to do it'. This is analogous to S's situation's falling under the description in the antecedent of (V). But the preceding moral considerations can be prevented immediately from fixing what is the right thing for the agent to do, all things considered, in case some relevant exception-making circumstance obtains, e.g., the promise's having been extracted under compulsion. In that case, the moral principle is inapplicable to the agent's present moral situation and reveals no moral obligation for the agent to keep the promise. If this type of blocking does not occur, the initial considerations may still fail to fix what is the right thing for that agent to do, all things considered, in cases where the agent has a more important, conflicting moral obligation. The latter type of blocking is viewed by Sosa as analogous to a type of overriding where S's wider/total internal state provides justification for believing that not-P, and the fact that there can be additional cases of overriding is viewed as analogous to the existence in the ethical realm of an additional type of blocking (cf. 1991, p. 239).[9]

In the ethical analogue, whether the following is true:

> (E) If S promised to do that action then performing it is the right thing to do, all things considered,

will depend upon there not being some relevant exception-making factor present and upon S's not having an even more stringent, conflicting moral obligation. This is comparable to Sosa's point that the existence of no overriders with respect to S's belief is required for the truth of condition (C1). Just as the truth of (E) ensures that no exception-making considerations are present and that S has no more stringent or equally stringent conflicting obligations, so (C1) is meant to ensure that no epistemic overriders are present. Thus Sosa no longer needs to make an explicit mention of the absence of overriders in his analysis of knowing once he has included condition (C1).

Since the malady mentioned by Plantinga is an overrider, Plantinga's case does not provide a counterexample and reveals a misunderstanding of the connection between (C1) and (V). Indeed, the reason that Plantinga offers for regarding (C1) as true is merely that in all the nearby possible worlds in which Gödel's theorem is true, the malady does lead to the production of the belief in the theorem. (The same shortcoming affects Plantinga's second proposed counterexample, which concerns believing a contingent truth, e.g., as statement of the value of the force of gravity, that is true in all nearby worlds.)

7.2 Further Considerations about Sosa

Nonetheless, an objection to Sosa's account might be provoked by noticing that nothing yet explains why conditional (V) itself holds. Is it possible that (V) is satisfied when S meets the obsessive, tricky Mr. Nogot, who likes to trick people into believing truths by providing evidence for falsehoods?

One way in which Sosa can deflect this concern by stressing his mention of field F in (V), and his view that such fields are specified widely enough to permit useful epistemic generalizations about the reliability of S as an informant to an epistemic community relative to which knowledge is being ascribed to S (cf. 1991, pp. 281–4). Such generalizations will not narrowly characterize the type of proposition believed by S in the tricky Nogot cases, e.g., as propositions concerning possible Ford owners in offices. And Nogot's obsession does not give him the ability to trick very many people about very many types of propositions within a more usefully characterized field to which the narrower type belongs.

Sosa has highlighted this detail in regard to a case where (C1) does hold but not because (V) holds. The trickery is practiced in Sosa's example by someone who has a car that is a lemon and needs very much for S, who trusts horoscopes, to buy it. So the trickster gets S to believe that P17: 'S will soon be offered a business deal', by planting in the newspaper horoscope under S's sign a forecast that implies that P17 (cf. 1991, p. 239). While Sosa admits that it is not just by accident or luck that S is right in thinking that P17, Sosa points out that S lacks justification because of having no adequate reason for trusting the horoscope.

Yet an external rigging of a match between a belief about certain types of future events and those future events could be imagined to hold more generally as a reason why a statement of form (V) is true. This will show that even when we combine (V) with the other comments by Sosa surveyed above, we do not obtain an analysis of the presence of a faculty or intellectual virtue.

Suppose that there were persons somewhat like D. H. Lawrence's rocking-horse winner, who acquire by some silly process, C, beliefs as to which horses will win some type of race not held very often, and who are thereby led to bet on those horses, and suppose that a group of philanthropists spying on these people were to rig the races so that the horses in question win. Statement (V) might seem to be instantiated for these betters, but they do not know ahead of time which horses will win, most certainly not on the occasions of their original bets when their own track record is not yet established. Nor do these people display a special intellectual faculty in arriving at those beliefs.

Thus it is significant that Sosa adds further details to his account of faculties. A move in the relevant direction is indicated by his shifting away from the view of faculties as belonging to a whole intellectual being, who performs a function, to speaking of a faculty itself as "functioning properly, i.e., performing its distinctive function adequately well", and Sosa goes on to speak of a faculty's "apt functioning" (1991, p. 273). Of course, it will make no sense to speak of the functioning of the occurrence of a state of affairs corresponding to a conditional such as (V). Thus Sosa adds a significant detail when requiring that if the tendency implied by (V) has been displayed then in conditions C S has been persistently right "through" a faculty or intellectual virtue, which is an "intrinsic state" that "adjusts"

S's belief to the facts in field F, so that the belief "turns out right by reason of the virtue" and the virtue "enhances ... [the] differential of truth over error" for relevant beliefs (1991, pp. 239, 277, 282).

Accordingly, Sosa presents an analysis of an intellectual virtue or faculty relative to an environment D which is at least close to being a special case of the following analysis presented by Rom Harré and Edward H. Madden (1975) of something's possessing (in environment E) the power to do A:

> x has the power to do A if and only if
> there are conditions, K, such that if x were in K then x would do A in virtue of x's nature,

where for some powers, e.g., a voluntary agent's power to mow the lawn, 'would' needs to be replaced by 'could'. Sosa writes:

> One has an intellectual virtue or faculty relative to an environment E if and only if one has an inner nature I in virtue of which one would mostly attain the truth and avoid error in a certain field of propositions, F, when in certain conditions, C (1991, p. 284),

provided, of course, that suitable restrictions are imposed on what can be instantiated for 'F' and 'C' so as to avoid triviality.

Thus in the example of the philanthropic belief satisfiers it is not in virtue of the inner nature of the betters that their confidence pays off. So they do not believe out of an intellectual virtue and do not know that the horses in question will win.

For many powers, especially of substances and materials studied by science, Harré and Madden view the relevant nature as including the presence of a mechanism whose operation causally mediates between the obtaining of conditions K and x's doing A. Yet that is compatible with a low probability on such occasions of a full manifestation of the power or ability. Thus Sosa imposes a further demand in specifying that an intellectual virtue satisfies requirement

> (O) There is a significant objective propensity (in E) for a belief formed under conditions C to be right.

7.2.1 Fundamental Intellectual Virtues

Sosa characterizes inner nature I as being that in virtue of which this propensity is present. This invites a comparison to Harré's and Madden's thesis that it is in virtue of x's nature that x will/can do A when in conditions K. They propose that it is not presently reasonable to rule out the possibility that in our world there may be what they call fundamental powers, that is, powers possessed by particulars part of whose nature just is to be such that they will/can do certain types of things under certain conditions, where no additional structure of further particulars constitutes a mechanism for those manifestations under those conditions. For example, they speculate that physics might allow that there is a great field not constituted by any further components, whose nature just is to have the powers to exert various sorts of forces at various points under the conditions of certain types of items being at those points.

Yet could Sosa allow for any fundamental intellectual virtues in this sense, where inner nature I just is the objective propensity mentioned above? This seems to be ruled out by the fact that objective propensities belong to whole set-ups and not to just some of the particulars within a set-up. But that point also renders obscure what it means to say that the objective propensity is there *in virtue of* inner nature I.

Perhaps a way around this difficulty would be for Sosa to replace conditional (V) and requirement (O) with the following conditional:

> (L) If there were a large number of beliefs in propositions within field F formed under conditions C, then a high percentage of them would be true.

Here the mention in the consequent of (L) of a high percentage does not have the oddity of Sosa's having mentioned likelihood in the conditions I and II. Such a change will also permit us to consider whether there can be any fundamental intellectual virtues in the sense that nature I just is the occurrence of the state of affairs corresponding to the instantiation of (L).

First, suppose that there can be such a virtue. This brings Sosa face to face with what I have elsewhere called the problem of newly acquired abilities, roughly, the problem of how to analyze the difference between the presence of a capacity to do A and the presence of an ability to do A (cf. Shope 1999). We might illustrate the problem by an example of a quick learner, who does not have the ability to run a certain type of machine but does have the capacity to do so, and who would quickly in the course of first trying out its controls acquire that ability. It is true even now that if she were to manipulate the machine a number of times then she would run it on a high percentage of those occasions.

Sosa is aware of this type of concern regarding the analysis of virtues in general. He points out that if "conditions C are allowed wide compass while the environment E is narrowly circumscribed, then, as a newborn, Chris Evert already had the virtues of a tennis player" (1991, p. 285). For instance, we might take E to be merely being biologically healthy and C to include a course of development and training, leading eventually to being in a tennis match. In order to distinguish Evert's virtues from capacities, Sosa proposes that we

> restrict the scope of conditions C in such a way that only after a period of maturation and learning does she come to be in an environment E with an inner state I (a *total* relevant epistemic state, including certain stable states of her brain and body) by virtue of which she *would* then perform stellarly when in the conditions C of a tennis match (on the surface of the earth, etc.) (1991, p. 285).

But this provides no general way to distinguish in general between a virtue and a capacity to have a virtue, unless it is proposing that in the case of a virtue, E is not encountered until some process of maturation and development has occurred and C are conditions not including a process of maturation and development. Unfortunately, this has the untoward implication that innate virtues are impossible, which is tendentious (and may unduly constrain Sosa's discussion of what he calls fundamental virtues, in his own special sense of 'fundamental' (cf. 1991, p. 277)).

Moreover, this leaves the analyst with the task of distinguishing a process of development from the warm-up period that might be required in the course of exercising some virtues (cf. Shope 1999).

As hinted at by Sosa's mention of states of Evert's brain and body, he does suggest that there are no intellectual virtues that are fundamental in Harré's and Madden's sense, and he requires regarding conditional (V) that there is one or perhaps a number of alternative "grounds or bases" that form aspects of an intelligent being's inner nature "from which the truth of the . . . conditional derives in turn" (1991, p. 141). His argument is that if S retains the virtue relative to E insofar as the conditional continues to be true even when S is outside E and not in C with respect to a given proposition then "it can only be due to some components or aspects of S's intrinsic nature I", for that is what "fully explains and gives rise to" the truth of (V) (1991, p. 141).

But Harré and Madden regard it as tendentious to require that there be an explanatory basis for the possession of each power, e.g., the powers of the great field, which requires, to use their image, that the world is like a set of infinitely nested Chinese boxes.

7.2.2 Interdependence of Virtues

In addition, there should be concern over whether Sosa's account of a given intellectual virtue is circular and needs to mention the exercise of other intellectual virtues as being among conditions C. Jonathan L. Kvanvig suggests that if a given virtue, e.g., intellectual creativity, is used as a means for supporting the party line rather than in an effort to find truth, it may dispose such a warped individual away from the truth, and Kvanvig protests atomistic analyses of the intellectual virtues, i.e., analyses, like Sosa's, that treat such virtues one at a time (cf. 1992, pp. 118–19).

7.2.3 Explanation of (C2)

Another reason to fault Sosa's account of knowing is that we have not yet found a way to understand his requirement that conditional (C2) holds because (V) holds. And we cannot even use the supposition that (L) holds to support (C2). Requirement (C2) seems instead to pertain to our intuitive understanding of an ability to believe truths in field F as more akin to the ability to obtain food than to the ability to have mild colds.

Perhaps Sosa thinks that (C2) is underwritten by a change in (V) that he eventually introduces in order to deal with myopics who, recognizing their limits, would refrain from making many judgments in field F, where the latter concerns the visual features of things. Sosa proposes that we drop mention of believing from the antecedent of the conditional and instead consider whether, relative to E, in the presence of I the following is true:

> (V′) If P is in field F and S is in C with regard to P then S is very likely to believe correctly with respect to P (1991, p. 286).

Our earlier considerations would then lead us to replace this conditional with the following:

> (L´) If numerous propositions: that P1, that P2, . . . , that Pn, were all taken from those within field F then there would be a high percentage of cases in that sample where S would believe correctly with respect to the proposition.

Sosa may wish to argue that (C2) holds or is likely to hold because of the truth or likelihood of (L´)

7.2.4 Likelihood of Truth

I have deferred until now the further worry of a lack of specificity as to how likely the truth of beliefs formed out of an intellectual virtue must be. If we substitute (L) and (L´) for (V) and (V´), this becomes the issue of how high a percentage is envisioned in the consequents of (L) and (L´). Kvanvig has proposed that intellectual virtues need to be conceived as relative to kinds of beings, since extraterrestrials very much smarter than us, who have powers similar to ours but far outstripping them, would in comparison find us deficient, as they would one of their own kind who only lived up to the level of performance that we display (cf. 1992, p. 126). Kvanvig thinks that this relativizing will require considering not just actual members of a given kind but also what other beings of that kind would have been like, and will introduce a further level of complexity in a search for an analysis of intellectual virtues (cf. 1992, pp. 127–29).

Sufficient questions remain unanswered[10] to agree with Plantinga that Sosa has not yet provided a set of individually necessary and jointly sufficient conditions to analyze the presence of a faculty or intellectual virtue.

8 ZAGZEBSKI'S VIRTUE EPISTEMOLOGY

Linda Trinkaus Zagzebski (1996) avoids Sosa's analysis of intellectual virtues, preferring to regard them as virtues in the same sense as moral virtues. And she does not even require that a virtue must be possessed, let alone be manifested, when the knower comes to believe that P. Nonetheless, she argues, virtues must at least be mentioned in the analysis of knowledge.

Zagzebski words her analysis in two different fashions, sometimes employing the undefined expression, 'achieving cognitive contact with reality', and at other times the expressions 'achieving truth' or 'achieving true belief'. Using the latter terms, her analysis is the following:

> S knows that P if and only if
> (1) S has a belief that P which has arisen out of some act(s) motivated by the disposition to desire the truth of beliefs,
> (2) each act referred to in (1) is of a type that would be/is apt to be/might be performed in S's circumstances by a person with intellectual virtues,

(3) S's general attitude is such that if there were evidence against the belief that P then that evidence would lead S to reflectively consider S's evidence,
(4) S has achieved the truth of the belief through/because of having the motivation referred to in (1) and having performed the type of act(s) referred to in (2); and
(5) if the act(s) referred to in (2) at all involve relying upon some testimony of others in the epistemic community, then S has (also) achieved the truth of the belief that P through/because of that testimony's having been motivated by the disposition to desire the truth of beliefs and having been a type of act that would be/is apt to be/might be performed in the circumstances by a person with intellectual virtues (cf. 1996, pp. 280-1, 295, 297).[11]

Zagzebski suggests that animals and young children can be knowers insofar as condition (2) does not demand that S possess an intellectual virtue, but she follows Sosa in distinguishing two types of knowledge since animals and young children do not satisfy (3). Yet it is questionable whether they even satisfy (1), whose generality seems to impute too much intellectual sophistication and to concern a disposition to desire that for any state of affairs S believes that it obtains only if (if and only if?) it does obtain.[12]

Perhaps the alternative formulation of condition (1) (accompanied by a corresponding reformulation of condition (4)) is meant to avoid this concern by speaking instead of a desire to make cognitive contact with reality, "where this includes more than what is usually expressed by saying that people desire truth" (1996, p. 167). But the problem is really whether animals and very young children only desire something *other* than truth. Zagzebski does speculate that "at the deepest level" all virtues may "arise from the same motivation, perhaps a love of being in general" (*ibid.*). Furthermore, she allows that one's being in cognitive contact with reality may concern one's entire system of beliefs, whose value is holistic and depends upon "such qualities as coherence, clarity, understanding, proper strength of conviction, etc." (1996, p. 316). The applicability to animals and young children is again unclear.

There are several difficulties concerning Zagzebski's treatment of Gettier-type cases. She maintains that for any analysis supplementing the JTB conditions with other, independent conditions, each Gettier case can be constructed in a similar fashion. We start with a situation where the conditions in the analysans other than the truth condition are satisfied but the truth condition is not. This will happen because of a feature of the situation that is not systematically describable in terms of anything ruled out by the other conditions, and so the falsity of the belief is "due to some element of luck. Now amend the case by adding another element of luck, only this time an element that makes the belief true after all" (1996, pp. 288-9). This makes the example fail to satisfy condition (4) of Zagzebski's own analysis, and permits her to count the case as one of ignorance. For instance, in one of Gettier's own examples, "truth is acquired because by accident Brown is in Barcelona" (1996, p. 297).

One difficulty is that there is no way to apply Zagzebski's formula so as to construct the Gettier-type cases concerning tricky Mr. Nogot, who would not provide S with the evidence involved if Nogot did not himself know that P, and who intends that through the cognitive processes that S employs S will be led to form a true belief that P. Since these cases can be modified so that they do not involve direct testimony from Nogot that is relied upon by S, Zagzebski would at least need somehow to widen condition (5), as well as to modify her characterization of Gettier-type cases.

9 PLANTINGA'S PROPER FUNCTIONALISM

In contrast to Sosa, Plantinga wishes to treat proper functioning as involving an intelligent being's functioning according to its 'design plan' in an environment sufficiently similar to the one for which it was designed, either by God or by evolution. Plantinga might, for example, attempt to distinguish the transition from having a capacity to having the corresponding virtue as being a process wherein a detail gets added to the design plan of the being.

In an early version (1993b) Plantinga's view was that one's knowing that P requires not only that one have a true, justified, sufficiently strong belief that P, but also that one meets the following requirements:

> (1) the cognitive faculties involved in the production of one's belief are functioning properly in an environment sufficiently similar to the one for which they were designed,
> (2) the portion of one's design plan covering formation of beliefs when in the latter circumstances specifies that such formation directly serves the function of forming true beliefs,
> (3) if those circumstances include additional beliefs or testimony, then the latter are or express beliefs also satisfying (2)–and so on, backwards through any chain of input beliefs or testimony from one person to another, and
> (4) there is a high statistical or objective probability that a belief produced in accordance with that portion of one's design plan in one's type of circumstances is true.

The analysis has an advantage from the start over Sosa's, namely, its ability to deal with Feldmanlike Gettier-type cases and with Lehrer's case of the cautious reasoner. Sosa, contrary to the stance of many epistemologists, must treat these examples as illegitimate, as cases where S is not justified in believing that someone in the office owns a Ford. Sosa needs to require that such justification, both here and in Gettier's original cases could arise only through S's reasoning through false beliefs, e.g., the belief that Nogot owns the Ford indicated by the evidence. Then such a false belief serves as an overrider.

9.1 Plantinga's Initial Solution of the Gettier Problem

Plantinga, in contrast, suggests that Gettier-type situations are cases in which there is a rather slight deviation of S's situation from the environment in which S's faculties were designed to function, but S's design plan tolerates belief formation in the deviant environment as a trade-off or compromise so that a being with a cognitive system of S's type does not suffer other eventual cognitive losses, e.g., because of carrying too large a brain for ready mobility, and the design plan thus treats this sort of belief formation as only indirectly aimed at truth. In the Nogot cases belief is not "produced by a segment of the design plan directly aimed at truth" (1993b, p. 40).

I have criticized (1998) this solution to the Gettier problem because it does not deal with the case of sincere Mr. Nogot, who continues to believe that he owns a Ford, and may be manifesting evidence to others that he does own one, on the basis of the usual adequate evidence of one's ownership, but, unsuspected by him, his distant vehicle happens to be destroyed by a meteorite simultaneously with his winning a Ford in a raffle. There is no difference in the way that he forms the belief and the way in which the ordinary Mr. Havit, who knows that he owns a Ford, forms a belief that he owns one. So it is difficult to see why there is a violation of any of requirements (1) through (4). Klein proposes a similar example involving destruction by a garbage truck (cf. 1996).

9.2 Plantinga's Modified Solution

Plantinga has responded by taking our objections to be that the comparison to the Havit case shows that these sincere Nogot cases do not violate requirement (2). But Plantinga hopes to revert to his point that there is a bit of "mild environmental pollution" here and characterizes Gettier-type cases as violating requirement (1) (1996, p. 310). Such cases purportedly involve a subtle change in what Plantinga now calls our (cognitive) maxi-environment, an environment similar to

> the one we enjoy right here on earth, [note omitted] the one for which we were designed by God or evolution. This environment would include such features as the presence and properties of light and air, would include such features as the presence of visible objects, of other objects detectable by our kind of cognitive system, of some objects not so detectable, of the regularities of nature, the existence of other people, and so on (1996, p. 313).

Subtle changes in this will be various (cognitive) mini-environments, only some of which are ones for which one's faculties have been designed, since in those circumstances the proper functioning of one's faculties "are more likely to lead to false beliefs than true . . . [and] display a certain deplorable lack of resolution" (1996, p. 316). For instance, the faculties don't enable me to distinguish my owning a Ford because I continue to own the one I did from my owning a Ford only because I have won one just now in a distant raffle.

So Plantinga rephrases condition (1) as follows: 'The faculties involved in the production of one's belief are functioning properly in a macro-environment and in a mini-environment sufficiently similar to the ones for which they were designed'. Although Plantinga could improve (4) merely by requiring that the mini-environment "is favorable" for the exercise of the faculties in question, he considers two ways of rephrasing (4) more sharply (1996, p. 327). First, in a comparison to Sosa's requirement that there be no overriders, Plantinga considers requiring that in the mini-environment "there is a high objective probability that a belief meeting the . . . [other] conditions will be true" (1996, p. 368). Alternatively, Plantinga contemplates rephrasing (4) by appealing to something akin to Sosa's Cartesian tracking and requiring

> (B) If S were to form a belief by way of the faculties mentioned in the previous conditions then S would form a true belief.

This is quite close to conditional (C1) concerning Cartesian tracking, and by incorporating mention of the employment of faculties, it resembles Sosa's requirement that (C1) be true because Sosa's conditional (V) is true.

9.3 Remaining Difficulties

Nonetheless, a significant difference is that the amended (4) in Plantinga's account does not restrict attention, as did Sosa, to aspects of the internal nature of S. So by permitting inclusion of many external factors within the mini-environment, Plantinga's analysis remains too weak to rule out knowledge in a number cases involving genuine defeaters, e.g, in a Grabit case where Tom stole the book but there was unsuspectedly an identical twin in the library at the time. If we add to S's evidence the truth that Tom's twin was at a different place in the library from the spot where the theft took place, this will defuse the defeating impact of the fact that Tom's twin was in the library when the theft occurred. So if the mini-environment is taken to include this difference in location, condition (4) in the amended version of Plantinga's account is satisfied when it is articulated in terms of objective propensities. If we add the further detail to the case that the twin finds stealing abhorrent, then even Plantinga's conditional (B) is satisfied.

Plantinga's amended account is also too strong to handle the following case. S is investigating a portion of a substance in order to find out if it is radioactive and S has background information that if the substance were to produce scintillation in the liquid within which it is immersed, that would show it to be radioactive. But the emissions that produce the scintillation are released very rarely by the substance during a single day of watching it. At almost any moment of a single day, if S's faculties working properly did lead S to form a belief, as required in the antecedent of (B), it would be the false belief that the substance is not radioactive. To be sure, when an actual case of scintillation occurs, S does form the correct belief and comes to know that P18: 'The substance is radioactive'. But Plantinga has written that the mere fact that two such facts obtain does not suffice to render true a subjunctive conditional connecting them, such as (B) (cf. 1996, p. 328). So Plantinga would need to add even more to the amended (4) or to its interpretation in order to allow for knowledge that P18.[13]

It may appear that the difference at issue between Sosa and Plantinga also renders Plantinga's account subject to the earlier counterexamples concerning tricky Mr. Nogot because in those cases the mini-environment includes Nogot's intention, indeed obsession, to trick S into believing truths. But such examples are handled by a further detail that Plantinga includes concerning the content of a mini-environment, namely, that it is a state of affairs as much as possible like the actual maximally specific situation but does not imply the proposition that S forms a true belief nor the denial of that proposition (cf. 1996, p. 315).

Yet the account still clashes with a new variant, the sanitized tricky Nogot case, where we do not speak of Nogot as concerned with truth but merely describe his obsession as the project of tricking people into believing some of what Nogot does about matters concerning officemates' car ownership, about which he is a highly reliable judge.

We saw that Sosa can avoid tricky Nogot cases by relativizing ascriptions of knowing to epistemic communities and to categories of situations that it is useful to generalize over in respect to putative knowers' being reliable sources of information.[14] Analogously, Plantinga might try to place less stress on propensities or conditionals concerning a mini-environment and more on the idea that S's mini-environment is not sufficiently similar to the one for which S's faculties were designed.

9.3.1 Sosa's Critique of Plantinga

Before discussing such a move, it is worthwhile to note Sosa's objection (1996b) that Plantinga improperly implies that no knowledge is ever possessed by Donald Davidson's hypothetical Swampman, a molecule-by-molecule duplicate of Davidson accidentally generated in a swamp, who thereafter behaves exactly as Davidson would (cf. Davidson 1986).

Plantinga has replied (1991) to this type of objection that he is only concerned with a case raised as a putative counterexample when there is reason to think that the case is possible, and Plantinga denies that the Swampman case is of that type. But Sosa has fairly responded that the case may raise trouble if it is at least an "opposing counterexample" in the sense of a case that is not clearly impossible yet which runs contrary to the analysis (1996, p. 257). Sosa suggests that the less clearly impossible such a case is, the less clearly correct is the analysis to which it runs contrary.

9.3.2 An Extension of Sosa's Concerns

Accordingly, some might take the following case to be even less clearly impossible than the Swampman and merely to involve what Plantinga might be willing to characterize as a less than "reasonably successful" designer (1991, p. 206):

The Ultra-X-Files Extraterrestrials These space visitors study us secretly well enough to learn how to design and to produce from raw materials molecular analogues of a human newborn, designed for use in a zoo on their home planet, whose macro- and mini-environments are quite different from earth's. They succeed while here in creating the first prototype, Oscar. But their project then gets cancelled and in the process of embarking for their home planet, Oscar gets misplaced, grows up here, having been mistakenly taken as a human foundling, and subsequently behaves much like us.

Why doesn't Plantinga's account improperly imply that Oscar never knows anything and spreads ignorance through his testimony, upon which many of us rely?

Suppose that Plantinga responds that the case smacks too much of science fiction for him to allow that it is not clearly impossible. In response, it is not totally *ad* *hominem* to construct other cases that Plantinga, a sincere Christian, will find it more uncomfortable to treat in the same dismissive fashion. Plantinga takes seriously the doctrine that God created us with a *sensus divinitatis*, the faculty of being aware of God without the mediation of reasoning, but that this part of our design plan was damaged by the fall, although it now can be repaired through the grace of God acting through the Holy Spirit, so that subsequent to this repair, which involves "a smallish revision in the design plan", the faithful know that God exists as a direct response to His existence (cf. 1996, p. 337–8). Moreover, Plantinga has at times indicated that he does not regard the existence of a devil as inconsistent with the existence of God. So consider how Plantinga's analysis of knowing relates to the following case:

The Faustian Re-Formation Mephistopheles is preparing to drag a sinner, call him Faust, off to hell, and of his own free will effects a change in Faust's present design plan so as to repair Faust's *sensus divinitatis*, with the aim of having it function in the environment of hell so as to make Faust suffer even more by the recognition of what he has lost. God does not interfere with the repair, since He, unsuspected by the devil, will be taking Faust up to heaven, where Faust, using his repaired faculty, will come to believe that God exists.

Won't Plantinga's analysis have the intuitively questionable implication that Faust, in contrast to the faithful in heaven, will fail to know that God exists, even though he will believe it, simply because the relevant detail of Faust's eventual design plan was provided by a tinker who thought of the circumstances for its exercise as that of perdition. If Plantinga decides that Faust does know because the macro- and mini-environments of heaven are sufficiently similar to those of hell, some clarification of standards of similarity begs revelation.

9.3.3 Plantinga on the Cognitive Environment

Consider a further puzzle about the environment for which we were designed. Suppose some of us travel quite far into space. Whether we come to know anything there will depend, given Plantinga's account, on whether we were designed by God to use our faculties on the surface of our specific planet, helping one another and raising our families, rather than pursuing exotic phenomena far out in space.[15] But even if our intended environment was less restrictive, suppose that for us to use a space ship to reach certain distances, it is required that the crew remain in constant darkness and operate by feeling and hearing within the confines of the ship. Since Plantinga has described our earthly macro-environment as not involving constant darkness, how are we to decide whether the space travelers' environment is sufficiently similar?

9.3.4 Being Made to Have Beliefs

A final puzzle to be noted here concerns Plantinga's response to Lehrer's case of Mr. Truetemp, who, unsuspected by Truetemp but because of a medically benevolent operation, has been implanted with a small device that functions to cause him every hour to produce correct beliefs as to his temperature. After the operation he forms firm beliefs of this sort but admits that he has no idea why he has them (cf. Lehrer 1996, p. 32). Plantinga says that Mr. Truetemp lacks knowledge of his temperature because he has a defeater consisting in his belief that he is constructed like the rest of us and none of us has the ability to form direct and correct beliefs as to our temperature; moreover "everyone he meets scoffs or smiles at his claim to have such an ability" (1996, p. 333). Plantinga adds that if Mr. Truetemp does not have a defeater here, he also lacks knowledge, "since proper function, in his situation, requires that he *have* a defeating belief . . . ". No doubt Plantinga could make a similar point about the case of the philanthropic belief satisfiers, and even about the rocking-horse winner in D. H. Lawrence's story.

But Plantinga will apparently be committed to saying that knowledge of one's temperature is indeed possessed in the case of little Lord Truetemp, a young child, not much learned in the ways of the world, and subject to the functioning of a similar, unsuspected device, who has started forming the beliefs as to his temperature but has not told anybody yet.

Many will find this result implausible[16] but perhaps not Plantinga, who maintains that there would be knowledge by someone that God exists in another of Lehrer's cases, where God just directly implants in S the belief that God exists (cf. Plantinga 1996, p. 338). Plantinga says that this implantation is at least a cognitive process, and can be seen as a limiting case of Plantinga's analysis, even though it involves no operation of S's faculties. But is it plausible to combine that verdict with its implication, namely, that if we have actually been designed by evolution and yet God does exist, then not even God could implant direct knowledge in us by implanting such a belief–simply because He was not the origin of the species? Moreover, if little Lord Truetemp's device had been mistakenly imbedded during the operation because of a technical mix-up, does this leave him without knowledge of his temperature simply because, à la Swampman, nothing designed his cognitive faculties to operate in the relevant way?

10 THE ANALYSIS OF KNOWING AS A BROAD CATEGORY

As in one of the preceding horse-betting examples, a 'seer' who only muses, but with unsuspected accuracy, about the future or who only has images of it, without any inclination to believe that the corresponding events will occur, is someone who might be regarded as revealing that both the belief condition and the justification condition of a JTB analysis are too strong. Other examples challenging those conditions have been proposed (for discussion, see Shope 1983). This reminds us of the risk incurred by focusing on adult knowers, since dogs and young children violate the justification condition in many of its formulations at the times that they are said to know that such-and-such. Some regard the latter knowledge ascriptions as metaphorical, but certain reliability theories analyze knowing as a broad enough category to cover such cases, as does Sosa's virtue epistemology. Such analysts can then seek to treat the favorite philosophical examples of knowing as belonging to a special species that can be characterized by some adjustments to the JTB analysis retaining something close to its original conditions.

I have sketched an approach along these lines (1983; 1989), which, in contrast to suggestions of philosophers such as Sosa, does not need to analyze the genus by means of any belief condition (whose inclusion raises concerns about infants' knowing various simple things, besides facing challenges alluded to above). One requirement of the analysis can be put succinctly by speaking of a certain sort of representing (where 'P*' continues to symbolize the state of affairs corresponding to the declarative sentence instantiated for 'P'): (R1) 'S has the power to proceed in a way such that S's proceeding in that way represents the condition of its being the case that P, i.e., represents the situation's involving P*'.

10.1 Representing

The preceding requirement concerns a special type of representing in which x can represent y even if x is not about y and not an item ordinarily called a representation, e.g., the tree rings' being of a certain number in a cross section of a tree can represent the age of the tree in years. Moreover, just as the rings themselves do not represent the age of the tree, neither does S represent something in knowing that P or in proceeding in a way referred to in (R1).

I have defended (Shope, 1999) a constitutive analysis of 'x represents y' that relativizes this analysandum both to a contextually salient what-question concerning y, such as Q1: 'What is the age of the tree in years?' and to various contextually salient propositions being justified, e.g. the proposition that the growth conditions of the tree have been normal. Relative to such details, X, the tree ring's numbering n, representing Y, the age of the tree in years, is analyzable, roughly, as the occurrence of a state of affairs involving Y having an affect upon the occurrence of a state of affairs involving X, where this relationship makes justified to at least some degree an answer to Q1 (relative to various other contextually salient propositions' being justified). In particular, the occurrence, O, for n years of a certain state of affairs concerning the tree's growth has had (in a 'nondeviant' way; cf. Shope 1999) some affect upon the occurrence of a state of affairs concerning the determinable: the-rings'-being-of-a-certain-number, since O was the nondeviant cause of that determinable's taking the determinate form that it did; furthermore, this causal relationship makes justified to at least some degree the following answer to Q1: 'The age of the tree is n years' (relative to various other propositions' being justified, such as that the growth conditions of the tree have been normal).

Similarly, suppose that when S is the baby or family dog, the occurrences of certain past relationships to Mommy's–or Master's–coming through the door at a certain time of day has been the (nondeviant) cause of the creature's proceeding in a certain way, say, the infant's looking toward the door–or the dog's stationing itself by the door–shortly before that time of day. Relative to other salient propositions being justified, this causal relationship makes justified to at least some degree as an answer to Q2: 'What is some of the domestic situation?' the proposition that P19: 'Mommy/Master will soon appear'. So S satisfies requirement (R1) for knowing that P19.

The relevance of a salient what-question to representing something to be the case, and derivatively to knowing something to be the case, might mislead one into accepting Alan R. White's position that to know that P is to be able to give an answer, namely, that P, which is the correct answer to a possible question (cf. 1982, pp. 119–20; and cf. Craig 1990). The dog and infant do manifest knowledge but not by producing it in the sense of displaying the answer to a question. Since they proceed in a way referred to in (R1) as a consequence of the earlier events that I have mentioned, they might be said to have shown what some of the domestic situation is and perhaps be said to have yielded an answer to Q2. But they still have not given an answer to a question, not even nonverbally.

10.2 Justified Propositions

Of course, when S is instead an older human, one way of proceeding referred to in (R1) is to assert the proposition answering the relevant what-question. It is only due to this possibility, involved in the cooperative inquiry within epistemic communities, that the knowledge of brutes and infants is of interest, and it is relative to such communities that a justified status is attributable to propositions, such as the answer to a relevant what-question or the propositions to which an ascription of representing is contextually relative.

I have proposed (1983) that a proposition is a justified proposition if and only if the rationality of members of a contextually relevant epistemic community would be more fully manifested in relation to epistemic goals by members accepting that proposition instead of competing ones and instead of withholding acceptance of any of these propositions. I have suggested that by taking scientific methodology as our best present guide to what it is like for rationality to be manifested, we may deal with examples of the social aspects of knowing.

One reason that we are willing to treat the more complex kind of state that interests philosophers who struggle to hone a JTB analysis as constituting a species of knowing as a broad category is that infants typically grow up to become members of epistemic communities to which ascriptions of the more complex type of knowing is relative. A second reason is that in providing an analysis of the latter type of knowing, which Sosa (1991) calls reflective knowledge and part of which I have called (1983) justified factual knowledge, we need to mention–in a manner that I shall shortly sketch–some states belonging to the genus, without implying that they belong to the narrower species.

10.3 A Capacity Concerning Thoughts

There is a second requirement for an analysis of the genus: (R2) 'S has the capacity to have the thought of the occurrence of the state of affairs P* be causally involved in S's proceeding in the way referred to in (R1)'. This capacity to have reality in mind when proceeding is manifested as an infant matures by the development of a corresponding power or ability. The manifesting of the latter power may then be partly involved in S's asserting that P to other inquirers.

This explains why it is only metaphorical to speak of some machines as knowing that P, for instance, to say that a door-opening device hooked up to an electric eye knows that something is coming. These machines satisfy (R1) but not (R2). Brutes such as dogs may fail to form epistemic communities, but come along as free riders to knowing, provided that they can have thoughts and the capacity mentioned in (R2), which seems to accord with how many people speak of such animals.

There are reasons (cf. Shope manuscript) for concluding that knowing as a broad category will require an analysis taking a recursive form, rather than merely including the conjunction of requirements (R1) and (R2). Because of this, we cannot say that knowing is the representational power and cognitive capacity mentioned in those two requirements, but we can regard it as a state whose embodiment at least always partly involves their presence with respect to some states of affairs.

10.4 Justified Factual Knowledge

I have sketched (1983) a way to solve the Gettier problem by analyzing the species of the above genus to which a belief and justification condition are relevant provided that we view epistemic matters as falling under the same general interests that we have when seeking adequate explanations in other domains. In order for S's knowing that P to belong to this species, S must be justified in believing that P through grasping an initial portion of what I technically call a 'justification-explaining chain' (JEC) connected with the proposition that P, a certain sequence of justified propositions beginning with a proposition of the form, 'M and that makes the proposition that P justified', and related in certain ways that involve adequate explanations of why members of the sequence are justified propositions and, roughly, of how they make the preceding member in the chain justified. Regarding any domain, an explanation is inadequate when it contains falsehoods at certain places, so by specifying the general structure of a JEC we can specify the places where falsehoods are prohibited for this reason from appearing in it. Gettier-type cases then turn out to be situations where an attempt to construct a JEC modelled on genuine cases of knowing requires inclusion of a falsehood at a proscribed place. But the definition of a JEC does not require that S have believed or accepted that falsehood (for details see Shope 1983 and for defense see Shope manuscript).[17]

The analysis of knowing as a broad category permits fleshing out my earlier sketch, which had analyzed S as having justified factual knowledge that P if and only if P, the proposition that P is justified, and S's belief that P (or acceptance of the proposition that P) is justified in relation to epistemic goals either through S's grasping portions of a JEC connected to the proposition that P or independently of anything making it justified. It can be argued (cf. Shope manuscript) that S grasps enough portions of such a chain when, roughly, it is in virtue of grasping the portions that S does that S possesses the representational power previously mentioned in (R1). It can be further argued that what it is for S to grasp a member of the chain, e.g., the proposition that K, is for S to know that K as an instance of knowing taken as a broad category. (This permits us to attribute justified factual knowledge to some children who possess but have not yet manifested the capacity mentioned in (R2).) An analysis of how S's belief that P (or acceptance of the proposition that P) becomes justified through grasping in this fashion some portions of a JEC can then complete the clarification of the analysans.[18]

Since the importance of a proposition's being justified through input from members of epistemic communities helps to *explain* the *need* for requirement (R2) in the analysis of knowing as a broad category, and since the latter type of knowing is *mentioned in* the analysis of justified factual knowledge, we may say that the analysis of the genus and the analysis of the species stand in a kind of symbiotic, but not viciously circular, relationship.[19]

11 THE RELATION OF KNOWING HOW TO KNOWING THAT

11.1 White's Proposal and Critique of Ryle

Alan R. White (1982) accepts Gilbert Ryle's distinction (1946, 1949) between knowing how something is (e.g., how the patient is; how the tune goes) and knowing how to do something (e.g., how to swim), but disputes a number of Rylean theses concerning the latter, including the following:

(1) Learning how to do something is improving an ability.
(2) Learning how to do something is never acquiring information.
(3) In leaning how to do something, the knowledge is imparted by practice, not by being told.

Counterexamples to each of those theses include learning how to enter a concert without paying and learning how to control the speed of an engine via fitting it with a governor. In addition, knowing how to spell 'cat' is a counterexample to claim

(4) Knowing how but not knowing that admits of degrees

(And consider one's only partly knowing the history of the Napoleonic wars). Such spelling know how is also contrary to claim

(5) The distinction between knowing or not knowing how to do something is the distinction between being clever or stupid and between being knowledgeable/ intelligent or ignorant.

White's further counterexamples to (5) include not knowing how to play chess, and being considered clever because of possessing much erudite information. White also objects to claim

(6) Knowing how to do something is being able to do it.

He points out that ships are able to float and machines are able to calculate. Furthermore, a normal person is able to hear traffic, and may be able to see farther than someone else as well as to distinguish light from dark. According to Ryle the following is true:

(7) One's being able to say how to do something is never sufficient to prove that one knows how to do it.

But White objects that being able to tell us how to do something may be "the only resource left to the crippled driving instructor, the paralyzed swimming coach and the arthritic seamstress who wish to pass on their knowledge" (White, 1982, 26).

White proposes that claim (7) was the basis that Ryle used to defend the following thesis:

(8) Knowing how to do something is not the same kind of knowledge as knowing how something is.

Having undercut this argument for (8), White attempts to disprove (8) by advocating the following premises about both knowing how to do something and knowing how something is:

(A) Knowing how is the ability either to show or to tell how.
(B) The latter ability is the ability to produce for a (potential) audience the answer to a question (such as, 'What is the way x As?' or 'What is the way Aing is to be / should be done?' or 'What is the way to A?' – where the latter is either tantamount to the preceding question, or else asks about the way that Aing can be done).

Since producing such an answer involves intentions, White concludes that it involves knowing that the answer to the relevant question is such-and-such, and so knowing how and knowing that are not two kinds of knowledge but instead knowledge of two kinds of things.[20]

Yet premise (A) can be challenged by a version of White's example concerning the paralyzed swimming coach when combined with the example of knowing how a tune goes. A totally paralyzed coach will manifest the latter knowledge in recognizing the tune when it is played to the coach through earphones, even though the coach cannot show us or tell us how it goes. But the recognition does show the way the tune goes at least in the sense of being part of what makes an answer justified to the question, 'What is the way the tune goes?' Namely, the answer, 'That way.' What needs attention here is the difference between a person's showing something in White's sense and an occurrence's showing something in the latter sense, where the occurrence involves the person.

11.2 Types of Powers and Capacities

Since it was the latter sense that was pertinent to the way that representing figured in the analysis of knowing provided above by conditions (R1) and (R2), these requirements prompt us to consider whether we might analyze knowing how if we modify their description of what gets represented. In the following analysis, we may instantiate for 'ø' phrases of the forms, 'to do A' (e.g., 'to swim'; 'to spell "deceive"'), or 'x As' (e.g., 'one swims'; 'the tune goes'), or 'one should A' (e.g., 'one should address a magistrate'; 'one should dance the step'):

(KH) S knows how ø/the way ø if and only if

(I) S has the power to proceed in some way or fashion, f, such that S's proceeding in that way represents one/the way ø being what it is[21]
(II) S has the capacity to have the thought of f be causally involved in S's proceeding in way f.

Ryle's thesis (8) is now called into question insofar as we have analyzed the difference between knowing how and knowing that as often[22] involving merely different types of representational powers and accompanying cognitive capacities.

Perhaps a further difference is that in some cases of knowing how we waive the requirement that condition (II) of (KH) be satisfied. Although White is correct that an instinctual ability to do something, e.g., to secrete adrenalin, typically may not involve knowing how, this may be because what is done is something the subject is made to undergo rather than a way of proceeding. In contrast, we do seem willing to say that some infants from birth know how to follow a bright light with their eyes. Such an infant also possesses the related capacity mentioned in condition (II). But simpler animals will not always satisfy that clause, as in a case where a biologist says, 'The male stickleback knows how to engage in a crucial type of finning display near the end of the courting ritual.' Perhaps condition (II) needs to be waived regarding lower animals, or the biologist's remark needs to be regarded as anthropomorphizing or loose speech. Alternatively, perhaps the combined conditions provide an analysans for 'S possesses knowledge of how ø' whereas condition (I) alone provides an analysans for 'S knows how ø'.[23]

Robert K. Shope
University of Massachusetts–Boston

NOTES

[1] When a statement is asserted by uttering the sentence or the word is employed in a relevant way in a sentence used to make an assertion, the proposition involved in the contextual implication is not entailed by the truth of the sentence but is nonetheless affirmed in asserting the sentence, unless it is 'cancelled' in special ways by additional features of the context of utterance, which for some contextual implications is not linguistically permitted.

[2] When interest (1) is viewed as involving a concern with reliable exercise of the ability to discriminate the occurrence from the nonoccurrence of a state of affairs, some versions of what is called a reliability theory of knowing (see below) have considered this to be a manner in which a kind of justification may belong to the exercise of that ability.

[3] Paul Moser's defeasibility analysis (1989) bears some resemblance to Klein's, but does not attempt to explain the nature of introspective knowledge.

[4] It also carries over the difficulty that none of the characterizations of the relationship in question have made sense of Klein's claim that the relationship is reflexive, which he appeals to in order to apply his analysis to noninferential knowledge (cf. 1981, p. 150).

[5] But Pollock does not prove that this alternation always ceases after a finite number of stages.

[6] Plantinga (1993a, pp. 219n–20n) reports an objection to Pollock raised by Richard Foley, namely, "that if I miss an obvious and nearby defeater *q*, then I don't really know *p*, even if *q* is in turn defeated by some defeater I would encounter, if at all, only after enormously

protracted reflection and investigation." Perhaps Pollock can reply that the obviousness in question makes the information that *q* something that S is socially expected to believe.

[7]The very complex defeasibility theory most recently defended by Lehrer (1990) has been criticized by Plantinga (1996) for treating knowing as overly intellectual in the sense of requiring too many thoughts to have occurred to S, e.g., the thought that S's system of beliefs is responsive to sensory experience and to external reality. Lehrer's analysis also gives the wrong verdict on a variant of the case of the clever reasoner (cf. Shope manuscript).

[8]According to some reliability theories, these characteristics are external to any of which S can easily become aware.

[9]Sosa speaks of the possibility of overriding justification as the "defeasibility" of justification (cf. 1991, p. 239n). But this may be misleading, since defeasibility theories concern the way that various statements that S does not believe/accept and which are nonetheless true relate to S's evidence/believing/having grounds for believing. Typically, a defeater's being true is not an aspect of the inner nature of S.

[10]Including some further concerns about the typical limitations of conditional analyses of powers and abilities (cf. Shope 1999).

[11]Since the testifier might be relying on another testifier, this last condition may more appropriately be worded by making the analysis recursive.

[12]Compare J. Greco's postulating (1990) in his treatment of intellectual virtues that children countenance epistemic norms without their having cognitive access to that commitment or to the fact of such conformity.

[13]This difficulty is close to what I call (1999) the recovery problem for conditional analyses of powers and abilities.

[14]His additional way of dealing with such cases is to treat them as involving S's reasoning through false beliefs.

[15]Consider Stanislaw Lem's novel, *Solaris*, and the film based upon it.

[16]For instance, Zagzebski may say that Truetemp fails to know because his belief-forming process does not imitate one that would or might be used in the circumstances by a person with intellectual virtues as a way of manifesting one or more of those virtues. Zagzebski does build a reliability requirement into the explanation of what a virtue is. Although she suggests that the accompanying generality problem can be solved by an empirical study of how people generalize their habits of belief formation, she admits that this might not be applicable to belief-forming processes "that are too automatic and close to the instinctive to count as habits" (1996, p. 311n).

[17]Roderick Chisholm (1989) also proposes to block Gettier-type examples by focussing on the role of falsehoods. For objections, see Plantinga 1993a, p. 63 and Shope 1998; Jason Kawall has pointed out to me that Chisholm will also not be able to admit the presence of knowledge in cases resembling Hilpinen's example about Millikan.

[18]This approach can deal with Millikan's knowledge that H (where for simplicity we ignore the relevant ranges of error). Consider the proposition that Millikan did accept that V: 'The charge of the electron is n'. The fact that E: 'Millikan obtained the experimental data that he did in the fashion that he did' is part of what makes justified the true proposition that C: 'The proposition that V counts as a justified proposition relative to the scientific community of Millikan's day'. This connection is part of considerations that make justified the true proposition that V′: 'The charge of the electron is reasonably/quite/significantly close to n'. One JEC connected with the proposition that H includes the proposition that V′, as well as the propositions that C, that E, and that I: 'The proposition that H is rationally inferable in the fashion followed by Millikan from the proposition that V'. The JEC will mention the existence of an argument paralleling Millikan's inference to H, in which the proposition that V′ figures in place of the proposition that V. Since Millikan did justifiedly accept at least the propositions that C, that E, and that I, it can be shown (Shope manuscript) that such a grasp of

a portion of the JEC was enough for him to have the representational power mentioned in my analysis of his knowing that H, where his having accepted the proposition that H represents the existence of state of affairs H*.

[19] Also see Sosa 1991 on the relationship between what he calls animal knowledge and what he calls reflective knowledge.

[20] White provides additional helpful discussion of a number of other philosophers' views concerning knowing how.

[21] I.e., represents one/the way ø being *that* (where the latter reference is to what is one/the way ø).

[22] We may agree with White that 'know' followed by an interrogative, e.g., 'how', and a verb phrase in the indicative mood. e.g., 'the engine broke down', deals with a variety of knowing that. In such cases S knows how x As if and only if S knows that such-and-such is the way x As.

[23] I am grateful to Peter Klein and Alvin Plantinga for helpful discussion.

REFERENCES

Almeder, R.: 1992, *Blind Realism: An Essay on Human Knowledge and Natural Science*, Rowman & Littlefield, Lanham, MD.

Audi, R.: 1993, *The Structure of Justification*, Cambridge University Press, Cambridge, England, New York and Melbourne.

Barker, J.: 1976, 'What You Don't Know Won't Hurt You?', *American Philosophical Quarterly* **13**, 303–8.

Carrier, L. S.: 1976, 'The Causal Theory of Knowledge', *Philosophia* **6**, 237–57.

Chisholm R.: 1989, *Theory of Knowledge*, 3rd edn., Prentice Hall, Englewood Cliffs, N.J.

Clarke, D. S. Jr.: 1989, *Rational Acceptance and Purpose: An Outline of a Pragmatist Epistemology*, Rowman & Littlefield, Totowa, N.J.

Craig, E.: 1990, *Knowledge and the State of Nature: An Essay in Conceptual Synthesis*, Clarendon Press, Oxford.

Davidson, D.: 1986, 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association* **60**, 441–58.

Dretske, F.: 1972, 'Contrastive Statements', *Philosophical Review* **81**, 411–37.

Feldman, R.: 1974, 'An Alleged Defect in Gettier Counter-Examples', *Australasian Journal of Philosophy* **52**, 68–9.

Feldman, R.: 1996, 'Plantinga, Gettier, and Warrant', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 199–220.

Gettier, E.: 1963, 'Is Justified True Belief Knowledge?', *Analysis* **23**, 121–3; reprinted in M. Roth and L. Galis (eds.), 1970, *Knowing: Essays in the Analysis of Knowledge*, Random House, New York, pp. 35–8.

Goldman, A.: 1976, 'Discrimination and Perceptual Knowledge', *Journal of Philosophy* **73**, 771–91; reprinted in G. Pappas and M. Swain (eds.): 1978, *Essays on Knowledge and Justification*, Cornell University Press, Ithaca and London.

Goldman, A.: 1979, 'What is Justified Belief?', in G. Pappas (ed.), *Justification and Knowledge: New Studies in Epistemology*, D. Reidel, Boston, Dordrecht and London, pp. 1–23.

Goldman, A.: 1986, *Epistemology and Cognition*, Harvard University Press, Cambridge, MA and London, England.

Greco, J.: 1990, 'Internalism and Epistemically Responsible Belief', *Synthèse* **85**, 245–77.

Grice, H. P.: 1961, 'The Causal Theory of Perception', *Proceedings of the Aristotelian Society*, Supplementary Volume **35**, 121–52; reprinted in R. J. Swartz (ed.), *Perceiving, Sensing and Knowing*, Doubleday, New York, 1965.

Harman, G.: 1968, 'Knowledge, Inference, and Explanation', *American Philosophical Quarterly* **5**, 164–73.

Harman, G.: 1973, *Thought*, Princeton University Press, Princeton.

Harré, R. and E. H. Madden: 1975, *Causal Powers*, Basil Blackwell, Oxford.

Hilpinen, R.: 1988, 'Knowledge and Conditionals', in J. E. Tomberlin (ed.), *Philosophical Perspectives, 2: Epistemology*, Ridgeview, 1988, pp. 157–82.

Klein, P.: 1971, 'A Proposed Definition of Propositional Knowledge', *Journal of Philosophy* **68**, 471–82.

Klein, P.: 1981, *Certainty: A Refutation of Scepticism*, University of Minnesota Press, Minneapolis, MN.

Klein, P.: 1996, 'Warrant, Proper Function, Reliabilism, and Defeasibility', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 97–130.

Kvanvig, J. L.: 1992, *The Intellectual Virtues and the Life of the Mind: On the Place of the Virtues in Epistemology*, Rowman & Littlefield, Savage, MD.

Lehrer, K.: 1965, 'Knowledge, Truth, and Evidence', *Analysis* **25**, 168–75; reprinted in M. Roth and L. Galis (eds.), *Knowing: Essays in the Analysis of Knowledge*, Random House, New York, 1970, pp. 55–66.

Lehrer, K.: 1970, 'The Fourth Condition for Knowledge: a Defense', *Review of Metaphysics* **24**, 122–8.

Lehrer, K.: 1974, *Knowledge*, London, Oxford University Press.

Lehrer, K.: 1979, 'The Gettier Problem and the Analysis of Knowledge', in G. Pappas (ed.), *Justification and Knowledge: New Studies in Epistemology*, D. Reidel, Boston, Dordrecht, and London, pp. 65–78.

Lehrer, K.: 1990, *Theory of Knowledge*, Westview Press, Boulder and San Francisco.

Lehrer, K.: 1996, 'Proper Function versus Systematic Coherence', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 25–45.

Lehrer, K. and T. Paxson. Jr.: 1969, 'Knowledge: Undefeated Justified True Belief', *Journal of Philosophy* **66**, 225–37.

Luper-Foy, S.: 1987, *The Possibility of Knowledge: Nozick and His Critics*, Rowman & Littlefield, Totowa, N.J.

Morton, A.: 1997, *A Guide Through the Theory of Knowledge*, Dickenson, Enrico and Belmont.

Moser, P.: 1989, *Knowledge and Evidence*, Cambridge University Press, Cambridge.

Nozick, R.: 1981, *Philosophical Explanation*, Harvard University Press, Cambridge, MA.

Olen, J.: 1976, 'Is Undefeated Justified True Belief Knowledge?', *Analysis* **36**, 150–2.

Plantinga, A.: 1991, 'A Reply to James Taylor', *Philosophical Studies* **64**, 203–17.

Plantinga, A.: 1993a, *Warrant: The Current Debate*, Oxford University Press, New York and Oxford.

Plantinga, A.: 1993b, *Warrant and Proper Function*, Oxford University Press, New York and Oxford.

Plantinga, A.: 1996 '*Respondeo*', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 307–78.

Pollock, J. L.: 1986, *Contemporary Theories of Knowledge*, Rowman & Littlefield, Totowa, N.J.

Pollock, J. L.: 1992, 'Reply to Shope', *Philosophy and Phenomenological Research* **52**, 411–3.

Russell, B.: 1948, *Human Knowledge: Its Scope and Limits*, Allen and Unwin, New York.

Ryle, G.: 1946, 'Knowing How and Knowing That', *Proceedings Aristotelian Society* **46**, 1–16.

Ryle, G.: 1949, *The Concept of Mind*, Hutchinson, London.

Shope, R. K.: 1983, *The Analysis of Knowing: a Decade of Research*, Princeton University Press, Princeton.

Shope, R. K.: 1984, 'Cognitive Abilities, Conditionals, and Knowledge: A Response to Nozick', *Journal of Philosophy* **81**, 29–47.

Shope, R. K.: 1989, 'Justification, Reliability, and Knowledge', *Philosophia* **19**, 133–54.

Shope, R. K.: 1992, 'Propositional Knowledge', in J. Dancy and E. Sosa (eds.), *A Companion to Epistemology*, Basil Blackwell, Oxford, pp. 396–401.

Shope, R. K.: 1998, 'Gettier Problems', in E. Craig (ed.), *Routledge Encyclopedia of Philosophy*, Routledge, London and New York, p. 4, pp. 54–9.

Shope, R. K.: 1999, *The Nature of Meaningfulness: Representing, Powers and Meaning*, Rowman & Littlefield, Lanham, MD.

Shope, R. K.: *Knowledge as Power*, manuscript.

Slaught, R. L.: 1977, 'Is Justified True Belief Knowledge?: A Selective Critical Survey of Recent Work', *Philosophy Research Archives* **3**, 1–135.

Sosa, E.: 1991, *Knowledge in Perspective: Selected Essays in Epistemology*, Cambridge University Press, Cambridge, England, New York, Port Chester, Melbourne, and Sydney.

Sosa, E.: 1996a, 'Postscript to "Proper Functionalism and Virtue Epistemology"', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 271–80.

Sosa, E.: 1996b, 'Proper Functionalism and Virtue Epistemology', in J. Kvanvig (ed.), *Warrant in Contemporary Epistemology: Essays in Honor of Plantinga's Theory of Knowledge*, Rowman & Littlefield, Lanham, Boulder, New York and London, pp. 253–70.

Swain, M.: 1974, 'Epistemic Defeasibility', *American Philosophical Quarterly* **11**, 15–25; reprinted in G. Pappas and M. Swain (eds.), *Essays on Knowledge and Justification*, Cornell University Press, Ithaca and London, 1978.

Swain, M.: 1981, *Reasons and Knowledge*, Cornell University Press, Ithaca and London.

White, A. R.: 1982, *The Nature of Knowledge*, Rowman & Littlefield, Totowa.

Zagzebski, L. T.: 1996, *Virtues of the Mind*, Cambridge University Press, Cambridge, England, Melbourne, New York.

MARIAN DAVID

THEORIES OF TRUTH

1. TRUTH BEARERS

A theory of truth is supposed to tell us what it is for something to be true or false. But what kind of "something" is at issue here? What are the bearers of truth and falsehood? Philosophical discussions of truth tend to revolve around four broad categories of *truth bearers*: sentences (utterances), statements (assertions), beliefs (judgments, thoughts), and propositions. The significant differences among these four categories are responsible for a considerable amount of complexity in the theory of truth. It seems unlikely that a single feature could account for '*x* is true' with '*x*' ranging over truth bearers of arbitrary type. Instead, a general theory of truth will have to take the form of a disjunction: *x* is true just in case *x* is either a true sentence or a true statement or a true belief or a true proposition. Here the general notion of truth is partitioned into four subnotions, *sentence-truth*, *statement-truth*, *belief-truth*, and *proposition-truth*, each of which will have to be given its own account. Still, the over-all theory will aim to show that the general notion of truth is not just a bare disjunction; it will aim to show that the subnotions are all tied (each in its own manner) to a single explanatory ground. One way to achieve this is by selecting a category of basic truth bearers and taking the associated subnotion of truth as primary, while treating the others as derived. The primary notion has to be explained in terms of some feature that does not itself contain any notion of truth. The derived notions have to be explained in terms of the primary notion together with certain relations that hold between the basic truth bearers, on the one hand, and the non-basic truth bearers on the other. A theory with this structure will give at least some unity to the general notion of truth.

Since knowledge requires true belief, epistemologists have a special interest in the prospects for beliefs as truth bearers: Are beliefs primary bearers of truth and falsehood, or are their truth values derived from the truth values of other, more basic truth bearers?

1.1 The Standard Analysis of Belief

According to the standard analysis of belief, the mental state of believing involves a relation between a believer and a *proposition*. The latter is an object of a somewhat special sort. If S believes that bats are mammals, then *what she believes* is said to be the proposition *that bats are mammals* – the proposition is the object of the belief relation and the content of the belief state: propositions are "content-objects." They are most perspicuously referred to by 'that'-clauses – schematically, 'that p', where the schematic letter 'p' abbreviates a declarative sentence, as in 'She believes that

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 331-414.
© 2004 *Kluwer Academic Publishers.*

p'. According to the standard analysis, propositions are the primary bearers of truth. The truth of a belief is explained in terms of the truth of the proposition that is its content. In fact, most proponents of this analysis would argue that, strictly speaking, belief states are not truth evaluable at all. The noun 'belief' is ambiguous. Sometimes it is used to refer to the state of believing a proposition; sometimes it is used to refer to the proposition believed. Strictly speaking, truth and falsehood go with the second use only: we say "What she believes is true"; we do not say "Her believing it is true." So the form 'Her belief that p is true' should be construed as 'She believes the proposition that p and this proposition is true'; and the impersonal form '*The* belief that p is false' should be construed as 'Someone believes (or might believe) the proposition that p but this proposition is false'.[1]

On the face of it, the standard analysis appears to be by far the most popular analysis of belief among contemporary analytic epistemologists. Talk of propositions as truth bearers and as content-objects of belief and knowledge is almost ubiquitous in epistemological literature. However, this talk does not always imply a deep commitment to the analysis; and there are reasons why epistemologists might want to avoid such commitment. The nature of propositions and, more fundamentally, the very need for introducing such entities in the first place, is subject to ongoing debate – a debate that has recurred regularly throughout the history of philosophy. It began when the Stoics were attacked by their contemporaries for introducing propositions (*axioma*) as incorporeal and non-mental truth bearers; and it continued through the Middle Ages when Abelard and Gregory of Rimini defended a recognizably Stoic position against the majority view which opted for verbal and mental sentences.[2] The debate died down in the modern period but was permanently reopened when the Stoic tradition enjoyed a full revival at the hands of Bolzano, Frege, Husserl, Meinong, and early Moore and Russell. The debate has turned out to be closely analogous to, and easily as protracted as, the debate about the nature of universals; with realist positions, conceptualist positions, and nominalist positions. Most contemporary epistemologists try to steer clear from getting too deeply involved in this debate. The hope is that significant work can be done in epistemology while staying as neutral as possible about the issue. In practice this means that many epistemologists adopt realist language for convenience, talking about propositions as contents of beliefs and as bearers of truth and falsehood, without necessarily committing themselves to genuine realism about propositions.

1.2 The Sentential Analysis of Belief

Concern for truth leads naturally to the study of logic – and as the practitioner soon finds out, whatever theory one might happen to hold about the subject matter of logic, in practice logic is done with sentences and formulas. The great success of 20th-century formal logic has made the sententialist approach very popular. Indeed, some authors talk as if there were no other serious candidates for the role of truth bearers, which suggests that they take sentence-truth to be basic and would explain belief-truth in terms of it, e.g.: A belief is true just in case every sentence expressing it is true. But the idea behind this approach seems misguided. The truth values of our sentences must depend in part on their conventionally assigned meanings. But why should conventional linguistic meaning enter into the explanation of truth and falsehood for beliefs? It seems quite wrong to think that true beliefs are true *because* of the meanings of certain sentences; if anything, it would be more plausible to think that the order of explanation should go the other way round.

In *Sophist* 263^e Plato famously identified thought with inner speech. Recent interest in philosophical psychology and cognitive science has brought back Plato's picture in form of the *language-of-thought analysis* of belief, advanced by Harman (1973), Fodor (1975, 1978), Field (1978), and others. This analysis, like the standard analysis, is a relational analysis. But it construes believing in terms of an underlying belief-generating relation, B, usually conceived of as a computational relation, to be spelled out by cognitive science. This relation relates the believer to a mental representation, more specifically, to a mental sentence of the language of thought: S believes that p iff there is some mental sentence *s* such that S stands in relation B to *s* and *s* means that p; and S's belief is true iff *s* is true. Thus, the notion of belief-truth is derivative; it is derived from mental-sentence-truth. The approach is supposed to preserve the advantages of taking sentences as truth bearers while it aims to avoid the objection raised above against deriving belief-truth from sentence-truth. The objection does not apply in this case, because the mental sentences that constitute the language of thought are said to have their meanings not by convention, but by nature.

1.3 The Non-Relational Analysis of Belief

The non-relational analysis of belief denies that believing is to be analyzed as a relation to some object or other. On this analysis, truth and falsehood belong to the state of believing that p rather than to the object referred to, or specified by, 'that p' – there is no such object. Accordingly, truth and falsehood are regarded not like properties of objects but more like different "modes" of states of believing. A sentence of the form 'Her belief that p is true' is to be understood along the lines of 'She runs quickly'; that is, its underlying logical form should be construed as 'She believes truly that p', with 'true' functioning as an adverb rather than a predicate. Since on this analysis there is no object from whose truth value the truth value of a belief state could be derived, the analysis treats belief states like primary truth bearers and belief-truth as the primary notion of truth. This feature recommends the non-relational analysis to those who want to steer clear from propositions, sentences, and mental sentences in their account of belief and belief-truth. But the analysis faces a serious obstacle.

1.4 The Problem of Logical Complexity

There is a grave difficulty with taking beliefs to be primary bearers of truth and falsehood. The problem arises due to their *logical complexity*. Assume S holds a logically complex belief, for example, the belief that p or q. Obviously, the truth value of S's disjunctive belief must depend on the truth values of the disjuncts. Yet, someone who believes that p or q may well not believe that p nor believe that q. Similarly, someone who believes that, if p, then q may well not believe the antecedent nor the consequent; and someone who believes that not-p will typically not believe that p. In all these cases, the truth value of S's complex belief depends on the truth values of its constituents, although the constituents may well not be believed by S or by anyone. But this means that a view according to which belief states are primary truth bearers seems unable to account for how the truth values of logically complex beliefs are connected to the truth values of their logically simpler constituents – to do that one needs to be able to apply truth and falsehood to belief constituents *even when they are not believed.* This difficulty arises in much the same form for views that want to take judgments, statements, or assertions as the ultimate truth bearers. The difficulty is not easily evaded. Talk of unbelieved beliefs and unstated statements is either absurd or simply amounts to talk of unbelieved and unstated propositions or sentences. It is hard to overstate the significance of this problem; unless it can be resolved, it rules out a whole host of popular candidates for the role of primary truth bearers in one fell swoop. The problem is largely responsible for the fact that philosophers in the late 20th century have displayed an inclination to prefer propositions or sentences (including mental sentences) as the primary truth bearers.[3]

1.5 Truth Without Bearers?

Relational analyses of belief treat 'true' as a predicate applicable to certain objects, the truth bearers. The non-relational analysis treats 'true' as an adverb and holds that the primary notion of truth does not require truth bearers in the full-blown sense. But even this analysis requires truth bearers in a wider sense of the term. For, if truth is a mode of believing, there has to be a state of believing for truth to modify. This means that all three analyses are committed to *recasting* locutions of the form 'it is true that p', for such locutions exhibit a prima facie "bearerless" use of 'true' as part of the sentential operator 'it is true that'. The propositional and sentential analyses will recast such locutions into subject-predicate form; viz., 'the proposition that p is true', and 'the sentence 'p' is true'. The non-relational analysis will opt for an adverbial form, like 'someone believes truly that p'. Now, there is a basic schematic principle about truth that spells trouble for these analyses, because it makes use of the bearerless operator-sense of 'true':

(T) It is true that p if and only if p.

The instances of this schema are self-evident truths about truth. But once 'it is true that p' is recast in one of the three ways just mentioned, the right-to-left direction of any given instance of (T) will be in doubt, because its left-hand side will carry an existential commitment to propositions, or sentences, or believers – a commitment not carried by its right-hand side. In short, theories that construe the primary notion of truth as requiring truth bearers of some sort will not handle (T) very smoothly. Advocates of such theories have learned to live with (T) as a somewhat awkward anomaly. They argue, or assume, that 'it is true that p' must be recast, for otherwise its connection to the uses of 'true' that do require truth bearers of some sort would be lost; and these uses are in the majority and appear to be theoretically more important.

2. THE AIM OF A THEORY OF TRUTH

The first thing to remember about a theory of truth is that it ought to account for falsehood as well as truth. It is hard to say much more about the proper aim of a theory of truth without begging the question against one or another of the projects that have been pursued under this title. There are various, and often competing, views about what the proper aim for a theory of truth should be: (i) A theory of truth should explain the nature of truth, specify the property in virtue of which true things are true, explain what it is that makes true things true, specify the conditions constituting something's being true, or explain what it is for something to be true; (ii) it should analyze the concept of truth or analyze the meaning of the word 'true'; (iii) it should explain what it is to grasp the concept of truth or what it is to understand the meaning of the word 'true'; (iv) it should describe the use, or the correct use, or the proper use, of the word 'true'; (v) it should specify the linguistic function of the word 'true' or specify what we do when we say that something is true; (vi) it should explain what the purpose or point is of saying that something is true; (vii) it should specify the criterion or test of truth, i.e., the conditions under which we can recognize something as true or the conditions under which we are justified in believing that something is true.

The differences between these projects reflect philosophical differences about how to address "What is?"-questions in general and the question "What is truth?" in particular. The entries collected under (i) range from "thicker", more metaphysically loaded, to "thinner" formulations. They describe variants of the classical approach according to which a theory of truth pursues a more or less metaphysical project. The approach is associated with so-called "realist" theories of truth, especially with correspondence theories. Its proponents generally regard projects (ii)-(vii) as, at best, ancillary to (i); and they will oppose the idea that project (i) could be pursued by way of pursuing one of the other projects – especially if the other project is taken from (iii)-(vii). They hold that such a strategy seriously confuses the question "What is truth?" with a number of entirely different questions, a confusion that typically leads into some form of relativism, idealism, or anti-realism about truth. Project (vii) is often singled out as paradigmatic for confusions of this sort. According to the classical approach, the conditions under which something is true or false have to be sharply distinguished from the epistemic criteria that allow us to tell whether something is true or false: the idea that we could pursue (i) in terms of (vii) badly confuses the aim of a theory of truth with the aim of epistemology.

Coherence theories of truth, verificationist theories, and some pragmatist theories are often labeled "anti-realist" precisely because they generally incline towards pursuing project (i) by way of (vii) – a strategy that tends to have anti-realist consequences. Their proponents are frequently accused of muddling the distinction between the theory of truth and the theory of knowledge. Advocates of so-called deflationary views of truth maintain that truth has no nature, that there is no genuine property in *virtue* of which true things are true. Consequently, they object to the "thicker" entries under (i) on the grounds that they beg the question against deflationism in the very description of the aim for a theory of truth. Although deflationists should be able to accept some of the "thinner" descriptions under (i), they tend to prefer (ii)-(v), especially (v). Nevertheless, many deflationists agree with proponents of (i) that (vii) describes the project of epistemology and should by no means be mixed up with the theory of truth.[4]

One would naturally expect a theory of truth to aim for a *definition* of truth. There is some disagreement about what precisely a definition is supposed to amount to – traditionally, it would have to be a biconditional with the strength of a necessary equivalence or the strength of a meaning equivalence (but there is some disagreement about what the latter is supposed to amount to). Moreover, a lonely definition will hardly deserve to be called a theory. A *theory* of truth worthy of the title should offer further explanations and illustrations, and maybe further definitions, pertaining to the notions in terms of which truth is being defined. Still, one would expect a theory of truth to offer *at least* a definition of truth in some reasonable sense of 'definition'.

There is, however, a serious difficulty concerning even the seemingly modest requirement that a theory of truth should at least offer a formulation that has the *logical form* of a traditional definition, i.e., the form of a universally quantifiable biconditional – '*x* is true iff *x* is D' – enabling us to replace the term 'is true' in all its occurrences by its definiens 'is D'. Tarski (1933) argues that we cannot define 'true' for the totality of declarative sentences of our ordinary language: the attempt to construct such a definition will inevitably run up against the paradox of the liar – the contradiction that comes to the fore when one tries to evaluate sentence L: 'L is not true'. This indicates that a theory of truth should not even attempt to offer a definition of this sort. It seems truth can be defined only for restricted artificial languages or for carefully circumscribed fragments of ordinary language. On the other hand, one might point out that Tarski's argument takes 'definition' in the traditional sense according to which only an explicit (eliminative) definition is regarded as formally correct. Tarski does not preclude a "definition" in a more lenient sense in which a set of characteristic axioms or principles might be said to provide an implicit (non-eliminative) definition of truth.[5] It is noteworthy that philosophical debates about truth regularly revolve around proposals stated in terms of explicit definitions, or similar formulations, that would give rise to paradox if taken in full generality. Indeed, it is customary to set aside Tarski's result for the purpose of philosophical (as opposed to logical) discussions of truth. The custom is sustained by the idea (the hope?) that basic philosophical issues about truth can be treated in terms of "provisional" explicit definitions presumed to be tacitly restricted to instances that are not "liar like" so that Tarski's argument is kept at arm's length. The task of turning such provisional definitions into something more final (and more consistent) is then, as it were, left to the logicians.

3. CORRESPONDENCE THEORIES

"Truth is a relation to reality; therefore, it has to be explained in terms of a relation to reality." This is the fundamental intuition characteristic of correspondence theories of truth. But the idea that truth is a relation to reality is just a pattern that has been spelled out in a number of different ways, resulting in a rather large family of theories. The family takes its name from accounts that call the truth-making relation *correspondence* and maintain that *facts* are the portions of reality that make truth bearers true. Other versions offer alternative conceptions of truth makers or alternative conceptions of the truth-making relation; moreover, different versions concentrate on different types of truth bearers.

3.1 Precursors

The correspondence theory might be traced back to Plato's *Sophist* 263^b, but the most frequently cited *locus classicus* comes from Aristotle:

> To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true. (Aristotle, *Metaphysics* 1011^{b}25)

Here Aristotle defines truth for sayings (*logoi*) which is probably intended to apply to verbal sayings (sentences, utterances) as well as mental sayings (thoughts). The definition is a bit indeterminate as to what kinds of sayings he had in mind: saying of what exists that it exists? of what is the case that it is the case? or of what is F that it is F? More importantly, although Aristotle's definition is often cited as the *locus classicus* for the correspondence approach to truth, it does not actually much emphasize the basic correspondence intuition. To be sure, it does invoke a relation to reality (*saying* something *of* something), but the relation is not made very explicit; nor is there an explicit specification of what, on the part of reality, should count as a truth maker. Aristotle sounds much more like a genuine correspondence theorist in the *Categories*, where he talks of "underlying things" that "make" statements true and implies that these truth makers are logically structured situations or facts.[6]

The Stoics employed *semantic* terminology in their account of truth, defining the simple proposition (*axioma*) that this man is sitting as true "when the predicate belongs to the thing which falls under the demonstrative." Related semantic accounts can be found in medieval writers, who tended to list separate clauses for sentences of different grammatical categories, e.g., John Buridan: "Every true particular affirmative is true because the subject and predicate stand for (*supponere pro*) the same thing or things. And every universal affirmative is true if whatever thing or things the subject stands for, the predicate stands for that thing or those things..." Thomas Aquinas cites a correspondence definition which he claims to find in Isaac Israeli: *Veritas est adaequatio rei et intellectus* – Truth is the adequation of thing and intellect. He approves of this, saying himself: "The judgment is said to be true when it conforms to the external reality." Authors of the modern period convey the general impression that some form of correspondence theory is regarded as far too obvious to deserve much discussion. Descartes: "I have never had any doubts about truth, because it seems a notion so transcendentally clear that nobody can be ignorant of it...the word 'truth', in the strict sense, denotes the conformity of thought with its object..."; and Locke: "Truth then seems to me...to signify nothing but the joining or separating of signs, as the things signified by them, do agree or disagree one with another"; and Leibniz: "Let us be content with looking for truth in the correspondence between propositions which are in the mind and the things which they are about." Correspondence formulations can also be found where they might come as a surprise given the author's overall philosophical position, e.g., in Spinoza: "A true idea must correspond with its ideate or object"; and in Kant: "What is truth? The nominal definition of truth, that it is the agreement of [a cognition] with its object, is assumed as granted."[7]

3.2 Truth Makers: Facts, Objects, and Properties

The classical version of the correspondence theory is motivated by intuitively plausible judgments of the following sort: the belief that bats are mammals is true because it corresponds to the fact that bats are mammals; if the belief that the cat is on the mat is false, then it is false because it does not correspond to any fact. The central definition of the theory is a straightforward generalization from such data. The now classical formulations were given by Russell: "Thus a belief is true when there is a corresponding fact, and is false when there is no corresponding fact" (1912, 129); and by Moore: "To say that this belief is true is to say that there is in the universe *a* fact to which it corresponds; and to say that it is false is to say that there is *not* in the universe any fact to which it corresponds" (1953, 277). It is noteworthy that the authors cited in the previous section – with the important exception of Aristotle (see note 6) – all seem to be talking about a relation to *things* rather than facts. Correspondence to *facts*, it seems, was a novelty in the days of Russell and Moore.[8]

Correspondence truth, as defined by Russell and Moore, applies primarily to beliefs. But they tended to discuss it with reference to the (declarative) sentences that are standardly used to express beliefs; and they treated synonymous sentences as *one*, presumably on the grounds that synonymous sentences would normally express the same belief. For starters, it will be convenient to be equally ambiguous about truth bearers. Thus (CF) will be treated as if it were a definition, even though it does not specify to what type of truth bearers it is intended to apply:

(CF) x is true $=_{\mathrm{df}}$ there is a fact y such that x corresponds to y;
x is false $=_{\mathrm{df}}$ there is no fact y such that x corresponds to y.

A preliminary issue that arises right away concerns the definition's intended breadth as to *subject matter*. Suppose, for example, (CF) is intended to apply to declarative sentences. Ordinarily, we have no qualms about calling sentences 'true' that concern ethical, aesthetical, logical, or mathematical subject matter. But some philosophers are strongly opposed to the idea that there are facts corresponding to such sentences. To avoid conflict with (CF), different strategies have been tried. Emotivists for example have claimed that ethical declaratives are not truth evaluable at all, in spite of their grammatical appearance. Positivists have claimed that the truth of logical sentences is not a matter of correspondence to facts, that the notion of truth relevant to logic is a different notion than the one covered by (CF). Both strategies have difficulties explaining what seems to be an important phenomenon:

there are instances of clearly valid reasoning in which ethical (logical) premises are used in conjunction with "factual" premises. It is hard to see how such reasoning could be valid, if there were no (univocal) notion of truth applicable to all the premises employed in such reasoning. An alternative strategy is to suggest that the allegedly problematic items *do* correspond to facts but to facts of a "different" subject matter, e.g., true ethical sentences correspond to facts about the customs extant in a given society, true mathematical sentences correspond to psychological facts, etc.

Definition (CF) will constitute the central definition of a classical correspondence theory of truth; but the definition is not the whole theory. A correspondence *theory* – even in a very loose sense of "theory" – should go beyond the mere definition and discharge a *triple task*: it should tell us about the workings of the correspondence relation; it should tell us about the ontological nature of facts; and it should tell us which truth bearers correspond to which facts. It is quite natural to tackle this triple task by construing correspondence in a strong sense, namely as a *structural isomorphism* between truth bearers and truth makers – let us call this the correspondence-as-*congruence* approach. The basic idea is that truth bearers and facts are both complex structured entities: facts are composed of things, properties, and relations; truth bearers are composed of words or concepts. The account then plans to proceed by showing how the correspondence relation that holds between a truth bearer and a fact can be generated from underlying relations that hold between the constituents of the truth bearer and the constituents of the fact. One part of the project will be concerned with these correspondence-generating relations; eventually, it must lead into a theory that addresses the question how simple words or concepts can be *about* things, properties, and relations. If the truth bearers are sentences, this part of the project will merge with a theory of linguistic reference, i.e., with *semantics*. If the truth bearers are beliefs or thoughts, it will merge with a theory of mental reference, i.e., with a theory of *intentionality*. The other part of the project, the specifically ontological part, will have to provide identity criteria for facts, and it will have to explain how the simple constituents of facts combine into complex wholes. Putting all this together should yield an account that tells us which truth bearers correspond to which facts.

The standard objection against correspondence theories is quite simple: correspondence is a mysterious relation. Remembering that the correspondence relation for sentences/beliefs belongs to the family of semantic/intentional relations helps put this objection in some perspective. It reminds us that, as a relation, correspondence will not be significantly more (but also not less) mysterious than semantic and intentional relations in general. Sure enough, such relations raise a host of difficult questions – most notoriously: Can semantic/intentional relations be explained in terms of natural (causal) relations? or do they have to be regarded as irreducibly non-natural aspects of reality? To mention a more specific puzzle: How do such relations manage to "reach out" over space and time to allow us to refer to, say, Nefertiti, who lived in a distant land and has been dead for more than 3000 years? The fact that Nefertiti *is* queen of Egypt does not seem to be around anymore. How does it manage to make true the sentence 'Nefertiti was queen of Egypt'? or should we say that the sentence is made true by the fact that Nefertiti *was* queen of Egypt, a fact that exists *now*? And what about reference to the future? Puzzles like these illustrate that semantic and intentional relations have some curious features. And there are philosophers who have argued that these relations are too mysterious to be taken seriously. But one should not lose sight of the fact that this is a very general and very radical complaint. The common practice to aim it specifically at the correspondence theory is misleading. It seems that, as far as the intelligibility of the correspondence *relation* is concerned, the correspondence theory of truth should stand, or fall, with the general theory of reference and intentionality. But maybe the standard objection is at bottom not really aimed against the nature of the correspondence relation, but rather against one of its relata, i.e., against *facts*. In this case, the standard objection would be of one piece with another very popular objection to the correspondence theory, to wit: facts are mysterious entities. Let us see, then, what correspondence theorists have had to say about truth makers.

On the most straightforward implementation of the correspondence-as-congruence theory, correspondence will be a one-one relation between truth bearers and congruent facts. The resulting account will be committed to all sorts of logically complex facts and logical objects that are often viewed with much suspicion. Consider *molecular* truth bearers, i.e., logically complex truth bearers that have other truth bearers as constituents, e.g., the sentence ‘not-p’, the belief that p or q, the statement that, if p, then q. Such molecular truth bearers will require negative facts, disjunctive facts, and conditional facts of arbitrary complexity – a fact for each true truth bearer no matter how complex. Moreover, these complex facts will contain logical objects corresponding to the logical constants ‘not’, ‘or’, ‘if-then’; and these objects will have to be regarded as constituents of the world. Many philosophers have found it hard to believe in the existence of all these “funny” facts and objects. Aiming to avoid ontological commitment to such objectionable entities, correspondence theorists have proposed various accounts on which correspondence is not a one-one relation between truth bearer and truth maker.

Wittgenstein (1921) and Russell (1918) offer a more “sophisticated” version of the correspondence theory as part of their program of *Logical Atomism.* The truth values of molecular truth bearers are to be accounted for *recursively* in terms of their logical structure and the truth values of their simpler constituents: a sentence of the form ‘not-p’ is true iff ‘p’ is false; a sentence of the form ‘p and q’ is true iff ‘p’ is true and ‘q’ is true; a sentence of the form ‘p or q’ is true iff ‘p’ is true or ‘q’ is true. These recursive clauses (called “truth conditions”) can be reapplied until the truth of a molecular sentence of arbitrary complexity is reduced to the truth or falsehood of its elementary constituents. (CF) is now restricted to *elementary* truth bearers whose truth makers are *atomic* facts; it serves as the base-clause for the truth-conditional recursions. The aim is to abolish the need for complex facts *by not* assigning any entities to the logical constants. Logical complexity belongs to the structure of language; it is not a feature of the world: “My fundamental idea is that the ‘logical constants’ are not representatives; that there can be no representatives of the *logic* of facts” (Wittgenstein 1921, 4.0312). According to atomism, there is no need for any facts but *atomic* facts; they are the sole truth makers – conjunctive facts are usually permitted because they are mere aggregates of atomic facts. Every truth has a truth maker, but not every truth is congruent with its truth maker(s). Loose talk of correspondence to facts is now explained recursively in terms of logical structure and the relation of correspondence in the strict sense (congruence) that holds exclusively between elementary truth bearers and atomic facts.[9] While the logical atomists seem to have held that the constituents of atomic facts are to be determined on the basis of *a priori* considerations, David Armstrong (1997) advocates *a posteriori* scientific atomism. On his view, atomic facts are composed of particulars and simple universals (properties and relations). The latter are objective features of the world that ground the objective resemblances between particulars and explain their causal powers; accordingly, what particulars and universals there are will have to be decided on the basis of total science. Still, Wittgenstein, Russell, and Armstrong agree on the basic thesis that facts do not supervene on, hence, are not reducible to, their constituents. Facts are entities *over and above* the particulars and universals of which they are composed: *a's having R to b* and *b's having R to a* are not the same fact even though they have the same constituents.[10]

Facts, even atomic facts, are sometimes spurned, because they appear to be *relatively abstract* candidates for truth makers. They are typically referred to by 'that'-clauses – e.g., the fact that Caesar died – but such a fact is not easily located in space-time. Even a "present" fact, like the fact that Clinton is president of the United States, seems hard to locate in space-time.[11] Moreover, the 'that'-clause 'that Caesar died' appears to refer to the truth that he died as well as to the fact that he died, which has suggested to some that facts are too closely tied-up with truths to serve as appropriate truth makers. In short, the charge is that facts are spurious sentence-like slices of reality that a good theory of truth should do without – a charge eloquently expressed by Quine (1987, 213): "Here again we have fabricated substance for an empty doctrine. The world is full of things variously related, but what, in addition to all that, are facts? They are projected from true sentences for the sake of correspondence"; cf. also Strawson (1950), and Davidson (1969).

Mulligan, Simons, and Smith (1984) have proposed a version of atomism that attempts to address such concerns. They propose to make do without facts (and without universals) and to return to the older tradition of correspondence to things of some sort. They argue that things and their *moments*, rather than facts, are the ultimate truth makers. Moments are concrete spatio-temporal particulars that depend for their existence on the substances on which they are founded and cannot be shared with any other substances. The individual accidents of the Aristotelian tradition, like Socrates' *individual whiteness*, are examples of moments. In general, all events, processes, actions, states, conditions, boundaries, configurations, and disturbances that require an underlying thing or medium are said to be moments.[12]

Both versions of atomism are based on the observation that a mere object *a* is not sufficiently *articulated* to serve as an adequate truth maker. If *a* were the sole truth maker of 'a is F', then 'a is not-F' would have to be true too; so the truth maker for 'a is F' needs at least to involve *a* and *Fness*. The advocate of facts will argue that, since *Fness* is a universal, it could be instantiated in another object, *b*, hence the mere existence of *a* and *Fness* is not sufficient for making true the claim 'a is F': *a* and *Fness* need to be tied together in the fact of *a's being F*, which will be a relatively abstract entity containing a universal. The advocates of moments will respond that *a's Fness* can be construed as a moment of *a* (e.g., Caesar's death). As such it cannot exist without *a*, hence, it cannot possibly be instantiated in anything but *a*. The moment of *a's* individual *Fness* would offer a "thingy," spatio-temporal, yet sufficiently articulated truth maker for 'a is F'. Some will regard this step in the direction of nominalism as a distinct advantage of the factless view. However, the view is likely to run into difficulties over relations – it is hard to see the *fasterness* of *a's* motion relative to *b's* as a spatio-temporally located moment – which would count as a serious disadvantage compared to atomism with facts.[13]

All forms of atomism propose to get by without logically complex truth makers by accounting for the truth values of complex truth bearers recursively in terms of their logical structure and atomic truth makers (atomic facts, moments). This strategy can be pushed even further by going, as it were, *subatomic*. Sentences can be broken up into their subsentential components. The relation of correspondence can be broken up into semantic subrelations appropriate to these subsentential components: names *refer* to objects; predicates (open sentences) are *satisfied* by objects. Satisfaction of complex predicates can be handled recursively in terms of logical structure and satisfaction of simpler constituent predicates: an object *o* satisfies '*x* is not F' iff *o* does not satisfy '*x* is F'; *o* satisfies '*x* is F or *x* is G' iff *o* satisfies '*x* is F' or *o* satisfies '*x* is G'; and so on. These recursions are anchored in a base-clause addressing the satisfaction of *primitive* predicates: *o* satisfies '*x* is F' iff *o* instantiates the property expressed by 'F'. Truth for a singular sentence, consisting of a name and an arbitrarily complex predicate, is defined thus: A singular sentence is true iff the object referred to by the name satisfies the predicate. Logical machinery provided by Tarski (1933) can be used to turn this vastly simplified sketch into a more general definition of truth – a definition that can handle quantified sentences as well as sentences with relational predicates. How general a definition of this sort can be made available is not known. It depends on the difficult question of how much of natural language is amenable to the kind of structural analysis whose availability is presupposed by the subatomic recursive approach. How much of the language of thought is amenable to this kind of analysis is obviously an even more difficult question. (The more an account of truth wants to exploit the internal structure of truth bearers, the more its range of applicability will be limited by the limited availability of appropriate structural analyses of the relevant truth bearers.)[14]

The subatomic version of the correspondence theory partitions the correspondence relation into two relations: reference and satisfaction. The task of accounting for these relations is part of the task of a semantic theory (probably the most important part). So, as far as the relation of correspondence is concerned, the correspondence theory merges with semantics and/or the theory of intentionality. With respect to facts, an advocate of this form of recursive account can maintain that talk of facts is metaphorical at best: facts are not really needed; they dissolve into objects and properties. It is contentious whether this dissolution of facts is more than cosmetic. Consider the sentence 'Snow is white'. What is the truth maker for this sentence? – not snow, nor the property of being white, nor both of them taken together. It seems, rather, that what makes the sentence true is the fact that snow instantiates the property of being white (in conjunction with the semantic facts that 'snow' refers to snow and 'white' expresses the property of being white). To this the response is likely to be that talk of truth makers should be dropped together with talk of facts – at bottom, all such talk should be regarded as metaphorical. The subatomic recursive approach promises a correspondence account of truth that has no use for facts or truth makers. Whether it can make good on this promise is subject to ongoing debate.[15]

3.3 The Problem of Falsehood

A theory of truth has to account for falsehood as well as truth. But falsehood raises a peculiar problem. The problem was first pointed out by Plato whose reasoning can be paraphrased like this: If S believes that the cat is on the mat, then *what S believes* is that the cat is on the mat; but, if the belief is false, then *that the cat is on the mat* is not the case and there is no such thing as *that the cat is on the mat* – and this seems to imply, absurdly, that one does not believe anything if one believes falsely.[16]

To see how this problem arises it is helpful to reconsider atomism and (CF). So far, (CF) has been taken to apply to beliefs and/or sentences – to the latter either for their own sake, or because they were regarded as representatives of the beliefs they would normally be used to express, or because they were construed as the mental sentences needed for a language-of-thought analysis of belief. Now, in the course of working out the details of her correspondence theory, the atomist will eventually have to confront truth bearers (beliefs, sentences, believed mental sentences) of the form

(1) Susan believes that p.

Truth bearers of this form are complex; they contain a truth evaluable constituent. Yet, they are *not truth-functional*: the truth value of the complex is not determined by the truth value of its truth evaluable constituent – replacing ‘p’ by a sentence with the same truth value can change the complex from a truth into a falsehood, and vice versa. But this means that (1) is beyond the reach of the atomists' truth-conditional recursions. Hence (1) as a whole, if true, must correspond to a fact. But what are the constituents of this fact? What, in particular, is the constituent that should be assigned to ‘p’ or to ‘that p’? Since (1) can be true even though Susan's belief may be false, no truth maker (fact) can be assigned to ‘p’. What would be needed here is something that can function as a *false maker* of Susan's belief. But the (CF) framework for beliefs and sentences does not provide for such things. It explains falsehood as absence of truth makers; thus, the truth of (1) as a whole is so far unaccounted for.[17]

One might want to reply that, despite appearances, (1) is not complex, hence, it does not contain ‘p’ as truth evaluable component. The theory would then simply declare that a truth bearer of the form (1) corresponds to some fact or other while remaining completely silent about the inner makeup of the relevant fact. But the reply is shortsighted. Although the truth value of (1) does not depend on the truth value of the embedded ‘p’, the truth values of truth bearers (beliefs, sentences) of the form

(2) Susan's belief that p is true
(3) Susan's belief that p is false

do depend on the truth value of the embedded 'p'. Moreover, (1) in conjunction with 'it is true that p' implies (2) and in conjunction with 'it is false that p' implies (3), while both (2) and (3) imply (1); and (1)-(3) all imply that there is something such that Susan believes it. On the face of it, none of this can be accounted for, if (1) is treated as logically unanalyzable. At this point, one might remember the language-of-thought analysis of belief (section 1.2): Why not simply assign a mental sentence as false maker to 'that p' in (1)-(3)? This should be feasible, provided the language-of-thought analysis is feasible at all. Still, the analysis tells us that S believes that p iff for some mental sentence *s*, S stands in relation B to *s* and *s* *means* that p. And this raises the question of how to analyze complex truth bearers of the form

(4) *s* means that p,

which pose difficulties concerning falsehood precisely analogous to the ones posed by (1)-(3).[18]

3.4 False Makers: Propositions and States of Affairs

One traditional response to the problem of falsehood argues that an adequate theory of truth must look beyond truth makers (facts), if it is to account for truth *and* falsehood – it must admit *propositions* as primary *truth bearers* because they can function as *truth makers* and as *false makers* for sentences and beliefs. The basic argument is simply this: the problems raised by (1)-(3) can all be accounted for on the assumption that a belief is made true, or false, by the truth, or falsehood, of the proposition that is the content of the belief. The same goes for (4) and its kin with propositions functioning as the contents of sentences. Now, the crucial point is that the issue of how to analyze (1)-(4) arose *within* the (CF)-account of truth and falsehood for beliefs and sentences. Evidently, one cannot coherently cling to (CF) *and* give a propositional analysis of truth bearers of the form (1)-(4): once these are analyzed in the propositionalist manner, (CF) has become unhinged from within – *a new account has just been given.* The argument from falsehood to propositions is, in effect, an argument for a two-tiered correspondence definition:

(CP_B) x is a true (false) belief $=_{df}$ x is a belief and there is a proposition y such that y is the content of x and y is true (false);

(CP_S) x is a true (false) sentence $=_{df}$ x is a sentence and there is a proposition y such that x expresses y and y is true (false);

(CP) x is a true proposition $=_{df}$ x is a proposition and there is a fact y such that x corresponds to y;
x is a false proposition $=_{df}$ x is a proposition and there is no fact y such that x corresponds to y.

Although (CP) is simply (CF) applied to propositions, sentence-truth and belief-truth are now derived from proposition-truth in combination with the *content-relation* and the relation of *expressing* (or *meaning*). The new definition of *truth* for beliefs and sentences is, strictly speaking, not incompatible with the one given by

(CF); it can be regarded as an analysis of the latter. But the new definition of *falsehood* brings a structural change: false beliefs and sentences fail to correspond to facts in virtue of "corresponding" to something after all.

The propositionalist correspondence theory outlined by the (CP)-framework goes naturally with the standard analysis of belief (which is no accident – the argument from falsehood to propositions is one of the major arguments *for* the standard analysis of belief). The theory could be regarded as the *standard view* – not because it is universally accepted (it is not), but because it constitutes, as it were, a "default" version of the correspondence theory, a position that the correspondence approach will naturally tend to revert to when left unattended. Also, it tends to be endorsed implicitly in areas like epistemology and the philosophy of mind where truth comes up regularly but its precise nature is not in the focus of interest. It is, however, often unclear to what extent such implicit endorsement can be taken seriously. There is a widespread practice (introduced in Russell 1918) of employing proposition-talk as a mere *façon de parler* – as a convenient shorthand standing in for more cumbersome talk about significant sentences or about contentful mental states like thoughts and beliefs (or somehow for all of those mixed together). One can even find theories of truth, including correspondence theories, that employ such *convenience-propositions* as truth bearers. The relative "popularity" of the (CP)-framework is due in some measure to this practice of not taking propositions too seriously. But convenience-propositions can hardly be primary bearers of truth and falsehood. They must be mere abstractions from language or thought and ontologically entirely dependent on them; they must be dispensable in favor of sentence-talk or belief-talk. Yet the problem of falsehood indicates that propositions are needed as truth makers and as false makers for sentences and beliefs, which means that one cannot endorse the (CP)-framework while hoping to get away without genuine commitment to propositions. (CP_B) and (CP_S) invoke propositions as *grounds* for truth or falsehood of beliefs and sentences: if propositions were somehow just "constructions" from beliefs or sentences, they could not be truth *makers* or false *makers* for beliefs and sentences. The argument from falsehood to propositions is, then, an argument for propositions as they are conceived of in the Stoic tradition (cf. section 1.1): abstract entities that are ontologically independent of mind and language but nevertheless fit to function as the meanings of our sentences and as the contents of our thoughts and beliefs. There is considerable opposition to propositions so construed – opposition that is based on a variety of considerations – and some philosophers, most famously Quine (1960, 1970), would argue that a good theory should do without them. It has however proven to be exceedingly difficult to come up with an account of belief and meaning that addresses the problem of falsehood without invoking abstract propositions or their kin.[19]

A related approach to truth invokes entities called *states of affairs*. They are typically denoted by 'that'-clauses or by gerundival clauses, e.g., the state of affairs of *Socrates' being snubnosed.* One important motivation for states of affairs derives from consideration of facts. Assume S believes that p: if S's belief is true, then it is a fact that p; if S's belief is false, then it is not a fact that p. What does the clause 'that p' in 'it is a fact that p' and in 'it is not a fact that p' refer to? Surely *not* to a fact, for there is no such fact if S's belief is false. The clause must refer to something that can both be a fact and fail to be a fact: the state of affairs that p. If the state of affairs obtains, then it is a fact; if it does not obtain, then it is not a fact. The crucial feature of states of affairs is, then, that they can be said *to obtain* or *fail to obtain*, to be the case or not to be the case, that is, they exist even when they are not concretely manifested or realized. Thus, states of affairs can function as truth makers and as false makers for sentences and beliefs:

(CS) *x* is true $=_{\text{df}}$ there is some state of affairs *y* such that *x* corresponds to *y* and *y* obtains;
x is false $=_{\text{df}}$ there is some state of affairs *y* such that *x* corresponds to *y* and *y* does not obtain.

Since a fact is a state of affairs that obtains, the definition allows one to say that truth is correspondence to a fact. But this now means that a truth corresponds to a state of affairs and that the state of affairs it corresponds to obtains. The relation of correspondence involved here is of a different sort than the one involved in (CF), for it gives truths as well as falsehoods something to correspond to. Actually, in (CS) 'correspondence' is used as a convenient "cover" term to be spelled out differently depending on what type of truth bearers (CS) is applied to. When it is applied to sentences, one could say that true as well as false sentences *represent* states of affairs; when it is applied to beliefs, one could say that true as well as false beliefs *have as their content-objects* states of affairs.[20]

States of affairs and propositions are both said to be denoted by 'that'-clauses. This might suggest that the (CS)-framework and the (CP)-framework are little more than notational variants. But on the face of it, there appear to be ample reasons for thinking that propositions and states of affairs are distinct. Propositions, though objective entities, are closely linked to mind and meaning. A proposition is composed of word-senses or *concepts* – where a concept should be taken as an objective way of conceiving of a thing or a property (a "mode of presentation"). Frege, who called propositions *thoughts* (1892, 1918), used this view of propositions and their constituents to explain how it is possible for S to believe that Muhammad Ali is a boxer even though S does not believe, or even disbelieves, that Cassius Clay is a boxer: S conceives of the same person in two different ways, i.e., two different concept-propositions are involved. Propositions are, then, rather fine-grained concept-entities. States of affairs, on the other hand, are much more coarse-grained. Intuitively, they are just like facts except that – very much unlike facts – they can fail to obtain without going out of existence. One naturally thinks of them as being constituted by "worldly" objects and properties, rather than senses or concepts. The state of affairs *that M.A. is a boxer* is constituted by M.A. himself and the property of being a boxer; thus, it is the same state of affairs as the state of affairs *that C.C. is a boxer*. If propositions and states of affairs are distinguished in this manner, the two accounts of truth might be regarded as competing with each other or they might be regarded as supplementary to each other. In the latter case, it is natural to think of the distinction between propositions and states of affairs along the lines of Frege's sense-reference distinction: two sentences can express different propositions but represent the same state of affairs.

Recent philosophy of mind and meaning has been much concerned with the notion of *content*. In particular, it has been much concerned with the question whether contents are "conceptual" entities (propositions) or "worldly" entities (states of affairs). The discussion relies on a widely accepted principle that ties the notion of content to the notion of truth: *Necessarily, if x has the same content as y, then x has the same truth value as y.* The principle codifies the idea that, whatever content is, the content of a belief/sentence must be something that *determines* the truth value of the belief/sentence. Well-known and much discussed arguments by Kripke (1972), Putnam (1975), and Kaplan (1977) have been taken to show that, at least in the case of proper names, indexicals, and natural-kind terms, the contents of the relevant sentences and beliefs should not be identified with concept-propositions: concept-propositions, it seems, do not determine truth value. Instead, contents should be identified with worldly states of affairs that are constituted by objects and properties. The latter thesis is presently under debate.[21]

The respective fates of the (CP)-framework and the (CS)-framework would seem to depend entirely on this debate in the theory of content. If the contents of sentences and beliefs are concept-propositions, the appropriate correspondence theory will be based on the (CP)-definitions. It will require semantic/intentional relations relating words and psychological episodes to concepts, and it will require relations relating concepts to objects, properties, and facts. If, on the other hand, the contents of sentences and beliefs are states of affairs, the appropriate theory will be based on (CS)-definitions. It will get by with semantic/intentional relations relating words and psychological episodes directly to objects, properties, and facts, without the mediation of Fregean senses or concepts. However this works out in detail, the larger point is that the shape of a correspondence theory – what sort of relations and entities it invokes – will be determined by requirements coming from semantics and the theory of intentionality. The point applies not just to the issue whether the (CP)-framework or the (CS)-framework offers a better account of truth; it applies quite generally. We have seen earlier that the correspondence approach is frequently criticized on the grounds that facts are dubious sentence-like slices of reality. In response to this, atomistic theories attempt to get by without complex facts; and "subatomistic" theories attempt to dissolve facts entirely. But the issue that is ultimately at stake here – whether an account of truth will invoke sentence-like slices of reality – is one on which the theory of truth must simply await the verdict, if any, from semantics and the theory of intentionality. *If* the analysis of belief and meaning requires content-entities, like propositions or states of affairs, then the theory of truth will be irrevocably committed to sentence-like slices of reality. Moreover, since we can think logically complex thoughts, the theory will be committed to logically complex sentence-like slices of reality. If so, Wittgenstein's appealing idea that logical complexity is a trait of language, or thought, rather than the world, will not be upheld: the idea overlooks that the content of thought is itself part of the world, so its complexity must be part of the world too. It seems, then, that the most basic issues that arise within a classical correspondence theory – the question of the nature of the correspondence relation, the question of the need for sentence-like slices of reality, and the question of the logical complexity of the world – have their proper home in the theory of meaning and/or intentionality. A classical correspondence theory of truth is little more than a spin-off from semantics and the theory of intentionality.

3.5 Minimal Correspondence

Full-fledged correspondence definitions raise a host of tangled issues in ontology, semantics, and the theory of intentionality. The intractability of these issues makes it natural to look for a more austere analysis of truth, one that preserves the basic correspondence intuition but somehow manages to steer clear from deep involvement in metaphysical affairs. Aristotle's definition in *Metaphysics* 1011^b25 (cf. section 3.1) is surely the paradigm for such a minimal correspondence approach. It also displays the chief disadvantage of the approach: excessive vagueness; it is simply not clear what the definition actually says. Some medieval authors liked to render Aristotle's definition in a peculiarly truncated style: A (mental) sentence is true just in case *sicut significat, ita est* – as it signifies, so it is. The formula seems skillfully designed so as to be maximally elusive with respect to what it is that is responsible for the truth of a sentence. Evidently, it all depends on the crucial semantic issue of what kind of thing(s) a sentence is taken to signify.

Contemporary advocates of the minimal correspondence approach tend to produce descendants of the elusive medieval formula. According to Strawson (1964, 79), 'S's statement is true' can be analyzed by way of 'It is as S states' or 'It is as S says it is'. Mackie makes a similar proposal: "*To say that a statement is true is to say that things are as, in it, they are stated to be*" (1973, 50). The semantic relation of signification that occurred in the medieval formula has dropped out here; but this is due solely to the switch from sentences to statements – some semantic relation will be needed to handle truth for sentences (the switch to statements is not entirely fortuitous, for they at best problematic candidates for primary truth bearers; cf. section 1.4). In Mackie's analysis, the relation of *stating* does duty for correspondence. This is a bit worrisome. Can one seriously maintain that statements can *state* how things are? – or is there a suppressed reference to persons who do the actual stating, as is suggested by Strawson's proposals? In the latter case, *stating* would be an intentional relation between persons and the *things* they state, which raises the non-trivial question what those things are. What about truth makers? According to Strawson's formulations, a statement is made true by something called "it". Mackie's reference to "things" might indicate that statements are made true by objects together with the properties they instantiate; alternatively, "things" might be a reference to states of affairs – it is hard to tell.

Mackie eventually arrives at the following analysis, which he regards as tantamount to his original proposal (cf. 1973, 59): x is a true statement iff, for some p, x is the statement that p, and p. Here the worrisome relation of *stating* has dropped out entirely; moreover, there is no explicit reference to any thing or things that could be identified as truth makers. It seems the idea that truth consists in a relation to reality has been diluted to the point of vanishing – nowadays, this would be regarded as a *deflationary* account of truth (see section 8). Alston, who also aims for a relatively minimal correspondence analysis of truth, takes the opposite route: "A statement is true if and only if what a maker of the statement is saying to be the case in making that statement, actually is the case" (1996, 23). Although Alston's wording does not make this very explicit, his analysis is committed to states of affairs. For on his account, the truth (falsehood) of a statement requires that there be something *x* such that *x* is said to be the case and *x* is (is not) the case. Only states of affairs, or similarly abstract entities that can be said to be the case or not to be the case, to obtain or fail to obtain, will fill this bill (the role of the correspondence relation is played by the intentional relation of *saying* which relates persons to states of affairs). Alston's analysis, unlike Mackie's, is recognizably a correspondence analysis – a version of (CS) from section 3.4 – but it is not obvious in which sense it is more "minimal" than standard correspondence definitions.

4. THE IDENTITY THEORY

Propositions are often introduced, quite generally, as the referents of 'that'-clauses. This very liberal way of introducing propositions leads rather directly to a surprisingly simple theory of truth; it has been called the *identity theory*:

(PI) *x* is a true proposition iff *x* is a fact;

All advocates of propositions maintain that the linguistic form 'it is true that p' is grammatically misleading. Its underlying logical form is said to be revealed by 'that p is true', in which the expression 'that p' functions as the logical subject referring to a proposition. On the liberal way of introducing propositions, the form 'it is a fact that p' must then be similarly recast as 'that p is a fact', where 'that p' refers again to a proposition. (PI) results from applying these recastings to an elementary observation: it is true that p if and only if it is a fact that p.

(PI) is put as a basic *principle* about truth, rather than a definition, because some of its advocates tend to *not* regard it as a definition. Indeed, at the time at which G.E. Moore espoused (PI) he also maintained that truth is at bottom indefinable – the same goes for Russell and Frege.[22] How does the identity theory relate to the correspondence theory? Chisholm construed identity as a limit case of correspondence: "There is no question, then, about the sense in which true propositions may be said to 'correspond with' facts. They correspond with facts in the fullest sense possible, for they *are* facts" (1977, 88). Moore (1901-02) and Frege (1918), on the other hand, regarded the two theories as competitors. It seems best to follow Moore and Frege on this point. Correspondence theorists normally hold that truths are not to be identified with facts. For they want to say that facts are truth makers – that truths are true *because* of the facts. Such claims would be pointless if identity were counted as a correspondence relation in the intended sense. One should remember, though, that (PI) is concerned with *proposition-truth* only. (CP_S) and (CP_B) are still in place. Sentence-truth still requires the semantic relation of *expressing* to hold between a true sentence and the fact that makes the sentence true. Belief-truth still requires the intentional relation of *having-as-content* to hold between a true belief and the fact that makes the belief true. Consequently, an identity theorist should still come out as a correspondence theorist with respect to sentence-truth and belief-truth.

But this point requires a qualification. How substantive the correspondence relation is taken to be might depend on a further issue. Note first that sentences have their contents contingently: words signify *ad placitum*. A (token) physicalists will treat (token) belief states similarly. He will hold that S's state of believing that p is a brain state of S's, and brain states have their contents contingently (although not by convention). In both cases the relation between truth bearer and truth maker is contingent; hence, it is naturally seen as a full-fledged correspondence relation. But it is often held – against (token) physicalism – that belief states have their contents essentially; that is, it is said that S's state of believing that p would not be the state it is, if it did not have the proposition that p as its content. On this view, it is *constitutive* of S's belief that p that it stand in the content relation to the proposition that p. And it might be thought that such a constitutive relation is "too intimate" to count as a full-fledged correspondence relation. If so, the identity theorist who is *not* a (token) physicalist will not be regarded as a full-fledged correspondence theorist about belief-truth.[23]

On the standard analysis of belief, the contents of our beliefs and thoughts are propositions. When (PI) is combined with this analysis, the result is rather startling: true propositions are facts, hence the content of the true thought that p *is* the fact that p – the fact itself, not some stand-in or representative of that fact. But isn't this more than a little bizarre? We think of facts as belonging to, or rather, constituting *the world*. The identity theory evokes the tantalizing picture of the world itself entering the mind. Or is the picture rather one of the mind stepping out into the world? Or are we being told that the world is constituted by the mind? It is sometimes said that a correspondence theory of truth opens up a "gap" between our thoughts and reality – a gap that, once opened, turns out to be unbridgeable, thus making it impossible for our thoughts to come into contact with reality. Similarly, it is said that a correspondence theory makes the attainment of knowledge impossible because the confirmation of a belief would require an impossible comparison between a thought in the mind and a fact of the world. Setting aside the question of how much force such worries have against standard correspondence theories, the identity theorist will claim that they have no force against his theory. If the content of a true thought is a fact, the issue of matching or comparing thought-content with fact can never arise in the first place: the identity theory has some nice consequences for the metaphysics of mind and knowledge.[24]

But are these nice consequences (assuming they are really there) not bought at too high a price? Well, what the identity theory actually amounts to depends very much on the underlying view of the nature of propositions. If propositions (the referents of 'that'-clauses) are identified with coarse-grained entities (states of affairs) that are constituted by objects and properties, then (PI) taken by itself is not at all bizarre. What may be considered peculiar is the thesis that propositions (in this sense) are the *contents* of our beliefs; that "worldly" objects (trees, pigs, stars) and properties make up the contents of our thoughts. However surprising this may seem, it is one of the major contemporary views about the nature of content. Defenders of the Kripke-Putnam-Kaplan inspired coarse-grained theory of content are committed to the thesis that the content of a thought is a state of affairs; hence, they are committed to the thesis that the content of a true thought is a fact (see section 3.4). If, on the other hand, propositions are construed as fine-grained concept-entities, as in Frege (1892, 1918) and Chisholm (1976, 1977), then (PI) is committed to a peculiar view about *facts*. Although facts will be objective mind- and language-independent entities, they will be constituted by concepts, rather than objects and properties. Facts will be as fine-grained as concept-propositions; hence, there will be as many different facts about, say, Aristotle and water as there are ways of conceiving of Aristotle and water. Most philosophers will object that it is absurd to maintain that there are as many facts as there are ways of thinking about things. If, finally, propositions are construed as ultimately mental or linguistic entities, (i.e., if they are construed as convenience-propositions), then the identity theory turns out to be a version of idealism. No wonder, then, that realists, like Moore, Russell, Frege, and Chisholm, as well as idealists, like Hegel and Bradley, can all be seen advocating "the" identity theory of truth.[25]

5. THE PRAGMATIC THEORY

"Truth is utility" – this, in a nutshell, is the view commonly referred to as the *pragmatic* theory of truth. It originates with the pragmatists F.C.S. Schiller and William James who liked to say that truth is what works, is useful, is expedient, pays. Some quotes from James will help convey the flavor: "An idea is 'true' so long as to believe it is profitable to our lives" (1907, 42); "The possession of true thoughts means everywhere the possession of invaluable instruments of action" (1907, 97); "*'The true', to put it very briefly, is only the expedient in the way of our thinking*...Expedient in almost any fashion; and expedient in the long run and on the whole of course" (1907, 106); "On pragmatic principles, if the hypothesis of God works satisfactorily in the widest sense of the word, it is true" (1907, 143). I shall go with common use and refer to James's view as the "pragmatic" theory of truth, even though the label might suggest that most pragmatists would endorse James's view, which is by no means the case. James's fellow pragmatists, C.S. Peirce and John Dewey, advocated an epistemic theory of truth. It should be noted that there is considerable overlap between epistemic and pragmatic theories; e.g., the epistemic virtues of verifiability, coherence, and explanatory power play a prominent role in James's theory because they are important ingredients of Jamesian utility.[26]

The kind of utility James has in mind is, to a first approximation, the utility that accrues from *believing* a truth. So, a rough formulation of the pragmatic definition, applied to statements, will look like this: A statement is true $=_{df}$ believing the statement is useful. The same form of definition will fit propositions – although James would have been much opposed to admitting the latter, unless they are construed as mere convenience-propositions. Applying the pragmatic definition to beliefs, taken as primary truth bearers, requires a minor adjustment: A belief is true $=_{df}$ holding the belief is useful.[27]

These rough formulations lack much detail and have to be fleshed out. For example, there are at least two broad senses of 'useful' that can be relevant to beliefs. On the one hand, a belief might be useful in that it leads to the fulfillment of an expectation or prediction; on the other hand, a belief might be useful in that acting on the basis of it leads to the satisfaction of a desire. What if these two come apart? What if, say, Smith's expectation that it is about to rain will be fulfilled, but it will also cause him to stay at home, thereby leading to the frustration of his desire to have a picnic? Another point raises concerns of a more logical nature. To say that a true belief is useful is evidently elliptical: we have to specify *for whom* the belief is supposed to be useful. The natural response is that it has to be useful for the believer in question. But this has the consequence of allowing both the statement that p and the statement that not-p to be true. After all, believing that p might be useful for Smith while believing that not-p might be equally useful for Jones. It is, however, an adequacy condition on a satisfactory account of truth that it not allow for violations of the law that a statement and its negation cannot both be true – a version of the principle of non-contradiction. Since the problem arises because utility can vary from person to person, one might try to address the difficulty by allowing truth to vary accordingly from believer to believer: A statement is true for S $=_{\text{df}}$ believing it is useful for S. The definition would commit pragmatism to relativism about truth, for it defines truth as the relativized notion *truth for S*. On this interpretation, the pragmatic theory will encounter much resistance: commitment to truth-relativism is widely regarded as a *reductio* of any serious attempt to come to terms with the nature of truth (see section 7). An alternative, non-relativist, proposal would be that a true belief has to be useful for *everyone* concerned, that is: A statement is true $=_{\text{df}}$ believing it is useful for everyone who believes it. This formulation raises questions about falsehood. If a false statement is defined as one that is useless for everyone to believe, then many more statements than is intuitively plausible will come out as neither true nor false. If, on the other hand, falsehood is defined as the absence of truth, then the definition will allow violations of the law that a statement and its negation cannot both be false: there being someone for whom the belief that p is useless does not guarantee that there is not also someone for whom the belief that not-p is equally useless.

The standard objection to the pragmatic theory of truth was raised by Moore (1908) and Russell (1908). They point out that the utility of a belief is, quite obviously, neither necessary nor sufficient for its truth. While it can be granted that, by and large, true beliefs are useful and, by and large, useful beliefs are true, there are many exceptions. False beliefs about one's own moral or intellectual qualities, or about the superiority of one's own culture, can be useful indeed; and they can remain useful throughout one's life. Moreover, all sorts of false beliefs can be useful if the facts conspire in our favor – missing a plane-crash because of false beliefs about departure times would be one (extreme) example. Conversely, true beliefs about trivial matters are often too inconsequential to be useful for anything; and worse, acquiring true beliefs about, say, the number of flowers on the wallpaper can be counterproductive, distracting one from thinking about more important matters. It is tempting to circumvent the latter problem by defining a true belief as one that *could* be useful under some circumstances or other. This, however, exacerbates the first problem: almost any falsehood could be useful under some circumstances. Utility, it seems, is at best a fairly reliable companion of truth.

Russell and Moore also argued that, even if it so happens that all and only useful beliefs are actually true, the pragmatic definition is still unacceptable because it has the wrong modal consequences. It *might* be useful to believe that A exists, even if A did not exist; and it *might* be useless to believe that A exists, even if A did exist. If truth were the same as utility, it would follow, absurdly, that it might be true that A exists, even if A did not exist, and *vice versa*. The pragmatic definition divorces truth from the facts: "The pragmatic account of truth assumes, so it seems to me, that no one takes any interest in facts, and that the truth of the proposition that your friend exists is an adequate substitute for the fact of his existence" (Russell 1908, 123). This type of objection plays a fundamental role in the theory of truth and is used regularly against (most) competitors of the correspondence theory. It relies implicitly on the schematic principle

(T) It is true that p iff p,

which can lay claim to being (partly) constitutive of truth in the sense that any purported theory of truth that comes into conflict with it must be regarded as having missed its mark. The objection points out that, on the one hand, (T) is a necessary principle involving 'true', while on the other hand, there are actual or possible counterinstances to 'it is useful to believe that p iff p'. Consequently, (T) would fail if 'useful to believe' were substituted for 'true'; hence, truth is not utility.

The pragmatist will have to insist that the identification of truth with utility cannot lead to failure of (T); that is, he will have to insist that it's being useful to believe that p *entails* that p, and it's being useless to believe that p *entails* that not-p. At this juncture, the debate about truth leads naturally into the debate over *anti-realism*. For the only way one can sustain the relevant entailments is by embracing the thesis that *what is the case* is determined by what it is useful to believe. In other words, the pragmatist seems committed to respond to Russell and Moore that the identification of truth with utility does not imply that it might be true that A exists, even if A did not exists, because whether A exists *depends on* whether it is useful to believe that A exists. Since this makes reality crucially dependent on *us*, this move will earn the pragmatist the accusation of succumbing to anti-realism. Note, however, that the anti-realism involved here is not a simple form of subjectivism. For whether believing that p is useful would seem to be an objective issue that does not depend on whether one *believes* that believing that p is useful. Still, the utility of believing that p does seem to depend on a person's goals and on her background beliefs. Large-scale revisions in her background beliefs may leave her belief that p so isolated that it becomes useless for promoting her goals; and if she changes her goals, the belief may not serve any of her new goals. So, on the pragmatic definition, the truth value of her belief that p, and, because of (T), whether it is the case that p, might vary in accordance with variations in the believer's goals and background beliefs.[28]

Richard Rorty – maybe the most outspoken contemporary pragmatist – once proclaimed that, in the "homely" and "shopworn" sense, "true" means roughly "what you can defend against all comers" (1979, 308). It is fairly obvious though that this is not the homely and shopworn sense of 'true'. On the contrary, it is precisely the homely sense of 'true' that makes us realize – with some uneasiness – that we may be unable to defend the truth if the circumstances conspire against us. To take an example from Goldman (1986, 18): A crime has been committed. All the evidence points against you. Your case is indefensible. You are lost. Nevertheless, the truth is that you are innocent. According to Rorty's anti-realist proclamation, you are guilty. It is similarly obvious that truth is not utility: we quickly realize that there may be, and probably are, useless truths and useful falsehoods. Philosophers of a realist bent will tend to agree with Moore's impolite evaluation that pragmatic accounts of truth, when taken literally, are "intensely silly" (1908, 115). It turns out, however, that advocates of the pragmatic approach, including Rorty, generally exhibit a realist bent; they tend to shrink from the more radical anti-realist consequences their own views seem to have *when taken literally*. There is, then, some question as to whether these views should be taken literally. Advocates of the classical approach to the theory of truth – project (i) of section 2 – will find ample evidence in support of this surmise in James's writings. He exhibits a marked tendency to talk in terms of what it is for an opinion to *count* as true or to be *adopted as* the true one – as opposed to what is for it to *be* true (cf. e.g., 1907, 35-36). This suggests that James might be concerned with projects (iv) and (vii) from section 2. That is, he might be concerned with the proper use of 'true', where the relevant sense of 'proper' is most likely broadly epistemological, pertaining to the conditions under which it is rational to regard a statement or belief as true. In short, James's theory of "truth" could be construed as a theory of rationality – a theory according to which verfiability, coherence, explanatory power *and* a potential for leading to successful action and to satisfactory emotions are all relevant to rationality. An advocate of the classical approach to the theory of truth may allow, if only for the sake of argument, that James's theory of "truth" is a viable theory of rationality; but she will insist that it is not a viable theory of truth.

Is James, then, simply confused about the difference between the theory of knowledge or rationality and the theory of truth? It is difficult to tell. There are indications that he is aware of the distinction but wants to maintain that a theory of truth is nothing but a theory of rationality: "The reasons why we call things true is the reason why they *are* true" (1907, 37). As a pragmatist, James is a staunch advocate of Peirce's (1878) "pragmatic maxim," according to which the content of an idea (concept, word) is to be defined in terms of the experiental and practical consequences of its application. In James's words: "There can *be* no difference anywhere that doesn't *make* a difference elsewhere" (1907, 30). James argues that his "proper-use" approach to truth follows from the pragmatic maxim (cf. 1907, 97). If this is the last word on the subject, then the pragmatic theory is anti-realist after all. For, as a theory of *truth* it has to honor (T), which can only be done by turning anti-realist; and maybe this is precisely what James means to do.[29] There is, however, an alternative interpretation of pragmatism (suggested by Rorty; cf. the introduction to 1982, and 1986). On reflection, it is far from obvious that the account of truth as utility follows from the pragmatic maxim. Instead of defining truth as utility, a pragmatist could maintain that (T) is his "theory of truth." This claim would seem to go well with the pragmatic maxim. For, according to the maxim, the concept of truth is given by the pragmatic (experiential *cum* practical) difference between truth and falsehood. And (T) tells us that the pragmatic difference between 'it is true that p' and 'it is false that p' simply reduces to the pragmatic difference between 'p' and 'not-p', whatever that difference might be in each case. In other words, the pragmatic maxim applied to (T) tells us that 'it is true that p' means the same as 'p'. The pragmatist could claim that this is already all that needs to be said about truth by way of a "theory". Anything of any real interest in the vicinity of the topic *truth* will concern the study of rational belief and its practical effects. With this deflationary attitude towards truth (cf. section 8), the pragmatist may hope to avoid anti-realist consequences, if they are unwelcome. At the same time, he could coherently pursue a theme that is surely one of the primary motivations for the pragmatic theory of truth: the rejection of the classical correspondence theory with its metaphysical "extravagances".

There is yet another way of interpreting the message James is trying to convey. Kirkham (1992, chap. 3.4) suggests that James and other pragmatists may be understood as trying to make a negative point about the *value* of truth as the traditional goal of rational inquiry. The point would be that truth, in the ordinary sense of the term, is not a value worth striving for: correspondence to reality, or some such thing, is not an intrinsically valuable property of a belief. We should stop seeing it as the goal we ought to be aiming at in rational inquiry. We should replace it with other values – values that really count, like verifiability, coherence, explanatory power, the ability to promote successful actions and satisfaction of desires.[30] But if the point is to reject truth as a value, then it is not fortuitously expressed by *redefining* 'truth' in terms of the values that one wants to put in place of truth. James seems to assume that the word 'truth' must stand for the ultimate and intrinsically valuable goal of rational inquiry *whatever* that turns out to be – if it turned out to be happiness, then truth would be happiness. Having effectively rejected truth for the part of the ultimate goal, he still wants to keep the word. This generates paradoxical results and is needlessly confusing. Moreover, it suggests that the pragmatist still hankers for the aura of respectability and intellectual purity surrounding the old idea of *the pursuit of truth.* Nietzsche, it seems, was significantly bolder in this respect: "*What* in us really wants "truth"?...*why not rather* untruth?" (1886, 9). He left the definition of truth alone and proceeded to roundly praise the value of life-promoting falsehoods.[31]

6. EPISTEMIC THEORIES

A theory of truth is an epistemic theory if it aims to account for truth in epistemic terms, like *justification*, *evidence*, *rationality*, *verifiability*, *warranted assertibility.* Compared to the correspondence theory, epistemic theories are relative newcomers. Their chief motivation derives from a particular diagnosis of the failure of the Cartesian project to secure knowledge by refuting skepticism. The diagnosis, to put it very briefly, is this: Descartes took for granted the correspondence theory of truth. But if truth is correspondence to reality, rational belief and true belief (about the external world) can always come apart, which means that there is *no guarantee* that even the most rational procedures will lead us to the truth. Hence, the skeptic cannot be answered and knowledge cannot be secured. The diagnosis already suggests the intended remedy. Descartes must have been wrong about truth: true belief *is* rational belief, or something close enough so that the two cannot come too far apart.

6.1 Infallibilism and Evidence-Transcendence

To understand what is involved in an epistemic approach to truth, it is helpful to look at the relation between truth and justification (evidence, warrant) and to reflect on the role these notions play in epistemology. Consider the "classical" analysis of knowledge:

(K) S knows that p iff (i) S believes that p, (ii) it is true that p, (iii) S has adequate justification (evidence, warrant) for believing that p.

The primary focus of concern for a theory of knowledge, especially for one intent on addressing skepticism, is the connection between conditions (iii) and (ii) – sometimes referred to as *the truth-connection.* The Cartesian project requires that this connection be absolutely tight. The very possibility of error has to be excluded. Thus, the truth-connection must be governed by a principle that makes condition (ii) logically redundant – justification must be *infallible*:

(INF) It is not possible that S has adequate justification for believing that p even though it is false that p.

Cartesian infallibilism takes for granted that truth is correspondence and understands INF as a constraint on the notion of justification. Nothing can count as the kind of justification required for knowledge, unless its possession *guarantees* truth and excludes the possibility of error. It is widely held that the Cartesian project was doomed to failure because this constraint is far too demanding. Descartes proposed indubitability as the source of infallible justification; unfortunately, it is not indubitable that indubitable beliefs cannot be wrong – and other candidates seem equally problematic. More importantly, it seems that the standards set by INF cannot be met by any methods for justifying beliefs about the external world that are actually available to us. Our empirical methods – observation, induction, inference to the best explanation – all fall far short from providing infallible justification; and purely *a priori* methods for justifying beliefs about the external world do not appear to be available to us.

One response to the failure of Descartes's project is skepticism. A second and more popular response is *fallibilism.* Like the Cartesian and the skeptic, the fallibilist takes for granted that truth is correspondence. Unlike the Cartesian and the skeptic, the fallibilist rejects INF. This allows that we can obtain knowledge about the external world through the fallible empirical procedures that are actually available to us. But fallibilism comes at a cost. First, there are Gettier's (1962) counterexamples to (K); they cleverly exploit the fact that truth and fallible justification can come apart. Some fourth condition has to be added to (K), or some other revision has to be made. But finding the right repair has turned out to be surprisingly difficult. Second, there is the problem of the truth-connection. Having rejected INF, the fallibilist is still expected to secure *some* connection between justification and truth. At least, she should be able to explain why, and how, justification based on our fallible empirical methods makes it *likely* that a belief is true. This, too, has turned out to be difficult. The worry here is that when it comes to the issue of skepticism, the fallibilist may not have anything more to offer than a weak defense of the mere possibility of knowledge.

Epistemic approaches to truth offer alternatives to skepticism and fallibilism. Two types have to be distinguished. The first equates truth directly with the *possession* of adequate justification. The proposal is, in effect, to retain INF but to give it an interpretation contrary to the one taken for granted by the Cartesian infallibilist. Instead of interpreting it as a constraint on justification, it is now interpreted as constitutive of the nature of *truth*. Whereas Descartes let (correspondence) truth set the standard for what counts as adequate justification, this position lets our criteria for what counts as adequate justification set the standard for what is true: beliefs that are justified by proper methods are true *because* they are justified by proper methods. Since the position aims to rescue the Cartesian project by holding on to INF, it is best seen as a version of infallibilism – post-Cartesian infallibilism, as it were. As such it can hope to preserve (K), since infallible justification makes (K) immune to Gettier's examples. The crucial difference to Cartesian infallibilism is that the new version allows for our empirical beliefs to be infallibly justified by the empirical justification procedures that are actually available to us – provided the procedures meet independently specified criteria of adequacy.[32]

The second type of epistemic approach is best introduced by recalling the hypotheses of hyperbolical skepticism: "I am systematically deceived by an evil demon," or "I am a brain in a vat." Our actual methods for justifying beliefs about the external world all depend in one way or another on empirical evidence. The skeptical hypotheses, on the other hand, *transcend* all empirical evidence: nothing within our experience can tell us whether we are victims of the evil demon or the evil brain-surgeon. But these hypotheses also entail the falsehood of all our beliefs about the external world. So, if it is so much as possible that the evil-demon hypothesis is true, our beliefs about the external world might all be false, even if we had done all that can be done to certify them: even ideal empirical justification could not guarantee truth. The idea now is that to rescue the Cartesian project the skeptical hypotheses have to be deactivated, i.e., the very possibility of evidence-transcendent truth must be excluded. Truth must be subject to an *epistemic constraint*:

(EC) Necessarily, x is true only if there is some evidence for x that is (in principle) available – only if belief in x is (in principle) justifiable.

The contemporary dispute over epistemic approaches to truth tends to focus more on EC than on INF. This is because one of the most influential advocates of the approach, Michael Dummett, motivates epistemic truth via a *verificationist* theory of meaning or cognitive significance – the theory that the meaning of a statement is determined by the empirically recognizable conditions for verifying and/or falsifying it. Since the truth/falsehood of evidence-transcendent hypotheses could not possibly make any difference to our experience, such hypotheses are regarded as cognitively defective in a manner that keeps them from being either true or false. Verificationism and its companion, EC, are still in the business of pursuing the Cartesian project, for they promise to eliminate hyperbolical skepticism, one of the most serious threats to that project. But the connection to infallibilism is less direct. An EC-based advocate of epistemic truth may want to define truth as some form of in-principle verifiability, which may not directly underwrite INF. For example, if 'in-principle verifiable' is taken to mean 'verifiable by the best scientific methods', it may well be held that it is possible for S to have adequate justification (justification that would otherwise be good enough for knowledge), even though S's belief is false because it would be falsified by the best scientific methods. But there is an important indirect connection to INF. If S's belief *is* justified by the best scientific methods, then it is guaranteed to be true. So an EC-based advocate of epistemic truth can be seen as defending a form of infallibilism – best-justification infallibilism.[33]

Classical fallibilists (i.e., correspondence-truth fallibilists) tend to think of the advocates of epistemic truth as philosophers who refused to learn the lessons of Descartes's failure and are now trying to save the Cartesian project by *cheating* in the theory of truth. Redefining truth to "prove" the skeptic wrong does not solve Descartes's problem – it just sweeps it under the rug. The problem will simply resurface in a verbally different form, e.g.: How can we *guarantee* that our "true" (justified, in principle justifiable) beliefs correspond to the way things really are? According to classical fallibilists, this problem cannot be solved. There is no such guarantee. Knowledge about the external world is indeed possible – but only when fallible justification combines with correspondence truth. In response, advocates of epistemic theories will usually repudiate the accusation that they are *re*defining the notion of truth. They will argue that an epistemic account of truth spells out what we meant all along by the word 'true'. In addition, they can return the charge of "redefinitionism" to the fallibilists; for fallibilists are potentially vulnerable to the charge that they *re*define the notion of *justification* to make knowledge at least possible.[34]

6.2 Truth as Justification

There are many different ways in which one might attempt to define truth epistemically. To get some handle on the range of possibilities, it seems best to begin with an extremely generic formulation:

(ET) x is true $=_{\mathrm{df}}$ there is sufficient justification (evidence, warrant) for x; x is false $=_{\mathrm{df}}$ there is sufficient justification (evidence, warrant) for the negation of x.[35]

The intended truth bearers are left unspecified because advocates of the epistemic approach to truth do not agree on whether beliefs, statements, sentences, or propositions should be regarded as the primary truth bearers. Propositions are generally frowned upon, unless they are mere convenience-propositions. Beliefs and statements are usually preferred, although they are at best problematic candidates for the role of primary truth bearers (see section 1.4).

Any account of truth along the lines of (ET) must address four groups of issues. First, we generally acknowledge that evidence can be misleading. There clearly is a sense of 'evidence' in which having evidence for a belief is not sufficient for the truth of one's belief. So an epistemic account of truth has to specify what kind, or degree, or level-of-quality of justification is supposed to be sufficient for truth. Second, any epistemic account must avoid making it entirely impossible for there to be true but unjustified beliefs. Surely, we do sometimes have unjustified beliefs, and sometimes they happen to be true. Third, justification (evidence) is always justification *for someone* (at a time). Ultimately, abstract talk about justification must be cashed out in terms of someone's being justified in holding a belief, or someone's being justified in believing a statement or a proposition. Advocates of (ET) like to use impersonal formulations – "truth consists in the existence of evidence" – or they refer to *us* – "truth is what is justifiable for us" – leaving it somewhat unclear who is included: some humans? all humans? other cognitive beings? Fourth, it has to be specified how "available" the relevant evidence has to be. Is a belief true only if someone actually possesses adequate evidence for it? or is it enough if the evidence is available in principle? – how available is "available in principle"?

To address the first point of the preceding paragraph, advocates of epistemic truth tend to require very "high-grade" justification, often referred to as *conclusive* justification or verification, as opposed to mere confirmation. Let us consider, then, two candidates for making (ET) a bit more precise. (I will continue to gloss over the subtler differences among the various epistemic notions – terms like 'evidence', 'warrant', or 'verification' might be used in place of 'justification'):

(E_1) x is true/false $=_{df}$ someone has conclusive justification for/against x;

(E_2) x is true/false $=_{df}$ it is possible for someone to have conclusive justification for/against x.

What is meant by "conclusive" justification? This notion is most naturally characterized as *truth-entailing*: a belief or statement is conclusively justified iff the evidence for it entails or establishes its truth.[36] Of course, this characterization would make both proposals circular; nevertheless, it is instructive to stay with it for a moment. On this truth-entailing interpretation of conclusive justification, both (E_1) and (E_2) offer conditions that are sufficient for truth. But is either of them necessary? If 'someone' is taken unrestrictedly so that it may subsume God, the answer to this question must be positive – and (E_2) may even be acceptable to some atheists on this reading. Defining truth by explicit or implicit (or veiled) reference to God is, in fact, a historically important position, one that may have been held by a number of philosophers who are sometimes regarded as defending some form of coherence theory of truth: Spinoza and Joachim might be interpreted in this manner. However, since God is "defined" as believing all and only the truths, invocation of God at this point appears to be less than helpful. Moreover, on the resulting definition, truth will be at least as inaccessible as on any correspondence definition – a feature that rather undermines the typical motivation for advancing epistemic accounts of truth in the first place. Restricting 'someone' to humans, it is often objected that (E_1) and (E_2) are both inadequate because of the vast number of *truths* (and falsehoods) that nobody has ever thought of, nobody will ever think of, and nobody could ever think of – let alone conclusively verify (or falsify). Although this objection raises a very important issue, it is somewhat tangential. It requires ontological commitment to truth bearers that exist independently of any (human) minds, i.e., to propositions in the serious sense of the term. Since defenders of epistemic truth are generally opposed to serious propositions, the objection leads away from the most central issue, namely: Assuming (E_1) and (E_2) to range over human beliefs or statements, is either of them satisfactory? And here it seems that both proposals are far too strong. Our empirical beliefs about the world are not justifiable by us in a manner that logically *guarantees* their truth – not in any sense that is consistent with us being us. The two definitions would not allow for any truths or falsehoods about anything but one's own conscious states and about subject matters that are accessible to *a priori* insight. Indeed, these definitions are just versions of original Cartesian infallibilism about justification which, if insisted upon, lead to a notational variant of radical skepticism. Whereas the standard skeptic argues that we know hardly anything there is to know, this type of skeptic says that we know most of what there is to know – but only because there is so little to know: our empirical hypotheses about the world are not even candidates for knowledge; they are all neither true nor false.

The main difficulty facing an epistemic account of truth is sketching out a notion of conclusive justification that is strong enough to be intuitively sufficient for truth without making reference to truth itself, but not so strong as to make it entirely impossible for there to be any true, or false, empirical beliefs about the world. Verificationists sometimes characterize conclusive justification as justification that is sufficient for *knowledge*. Since knowledge is supposed to be explained in terms of truth, this proposal is not serviceable. But it is suggestive – it suggests making use of a notion epistemologists have developed in the attempt to define knowledge, i.e., the notion of *evidential indefeasibility*: S has indefeasible justification for *x* iff S has justifying evidence for *x*, and there is no additional evidence (in principle) available to S that would defeat S's evidence for *x*; in other words, indefeasible justification is justification that remains stable no matter what additional information might become available.[37] More needs to be said about *how* available the potentially defeating evidence is supposed to be, but this rough sketch should suffice to make it plausible that there is a purely evidential notion of conclusive justification that could be used to spell out (ET):

(E3) x is true $=_{df}$ someone has indefeasible justification for x;

(E4) x is true $=_{df}$ it is possible for someone to have indefeasible justification for x;

(E5) x is true for S $=_{df}$ S has indefeasible justification for x;

(E6) x is true $=_{df}$ anyone in possession of all evidence relevant to x would be justified in believing x.

Opponents will object that these proposals are either too weak, or too strong, or both. Let us briefly look at some objections that might be raised. Justification is relative to persons (and times) and is highly sensitive to differences in background beliefs and to differences in personal experience. Hence, two persons might have equally adequate justification for logically conflicting beliefs. Of course, if conclusive justification were characterized as truth-entailing such cases would be excluded by definition. But is indefeasible justification powerful enough to *guarantee* that such cases cannot arise? – as it has to, since it is impossible for logically conflicting beliefs to both be true. There is the worry that two inquirers might be so separated from each other (for example, in time) that the evidence supporting the belief of one is not available to the other as defeater for her conflicting belief, not in any sense of "available" compatible with their being human inquirers – this worry about relativity applies to (E3), hence to (E4).[38] A related problem arises from the *underdetermination* of theories by evidence. It is often argued that conflicting high-level theories might be equally supported by whatever counts as evidence. Again this raises the specter of each of a pair of conflicting beliefs coming out as true according to (E_3) and (E_4). Alternatively, if it is said that the two theories would defeat each other, the consequence would be that of a pair of conflicting theoretical beliefs neither one would come out as false. One way of accommodating the relativity of justification is to turn to (E_5) and to define the relativized notion *truth for S* instead of truth. Of course, this would commit the advocate of epistemic truth to relativism about truth (see section 7); moreover, it is not clear how (E_5) can handle any of the other difficulties (primarily because it is not at all clear what it even means to say that *x* is true *for* S). Returning to (E_3), one may object that all available evidence might strongly indicate, say, that Caesar crossed the Rubicon, while the truth is that Caesar took to the sea to get quickly to Rome, though all traces of this clever move have been lost to time: we are now indefeasibly justified in believing that Caesar crossed. So it seems (E_3) – hence (E_4) – is too weak, unless the notion of defeating evidence being *available* to us is spelled out in a way that involves physical impossibilities like time travel, etc.[39]

(E_3) is usually discarded right away on the grounds that it is far too strong in any case. Surely, it is possible to hold a true belief for which no one has indefeasible justification. A more flexible formulation – "there is, was, or will be someone with indefeasible justification" – seems still much too strong. But what about (E_4)? Does it offer a necessary condition for truth? Well, couldn't there be true beliefs the evidence for which is inaccessible to us? Couldn't the belief that there is a golden mountain on some planet be true, even though the relevant planet is so far away that it is physically impossible for us to ever acquire any information about it due to limitations imposed by the speed of light? But maybe physical modality is too restrictive; maybe (E_4) should be interpreted as referring to what is *logically* possible for humans. But is it even logically possible *for humans* to break the laws of physics? A different type of problem arises from our vulnerability to additional evidence. It seems that for many justified true beliefs about the external world there will be some *misleading* evidence nearby that would defeat our justification. You just saw the President in New York City. You recognized him immediately. In your hand you hold today's newspaper which reveals to its readers that the President is at home and has sent a body-double to New York. The newspaper is mistaken. Still, had you read it, your justification for believing you saw the President would have been defeated. So your justification was not indefeasible. Yet, your belief was true nonetheless. Might there not be some (many) true beliefs about the world with respect to which it is impossible for humans to gather evidence that is invulnerable to this kind of problem? (E_4) says no – again, much depends on how one thinks of what is to count as humanly possible. In the end, one may have to talk about indefeasibility with respect to "the total body" of available relevant evidence – hence (E_6). Whether the notion of a total body of relevant evidence makes any clear sense is difficult to decide – accounts of this sort will be considered a bit further in section 6.4.

Proposals like (E_4) raise a divisive question: Is it possible for humans to acquire indefeasible justification for every truth they are capable of believing about the external world? Critics of the epistemic approach will give a negative answer; they hold that our capacity for belief may well transcend our capacity for acquiring justification. Advocates of the epistemic approach are committed to a positive answer; for they are committed to the view that true (or false) belief cannot entirely outrun our capacity for acquiring justification. I have given only a small sample of the kinds of disputes occasioned by epistemic accounts of truth. Ultimately, these disputes will turn on the issue whether a given proposal – one that is intuitively strong enough to be at least taken seriously as a candidate for defining truth – employs a conception of what evidence it is humanly possible to acquire that is still consistent with humans being humans. If the proposal requires superhuman capacities, it would seem to have become detached from any conceivable motivation for adopting an epistemic approach to truth in the first place. Some new motivation must then be given, if the position is not to come off as entirely *ad hoc*. Another issue that tends to come up sooner or later is the issue of anti-realism. Counterexamples to the effect that a given proposal is intuitively too weak or too strong for truth can always be "met" by an anti-realist rejoinder, i.e., by the claim that the *truth* about the subject matter at hand – e.g., the truth about what Caesar did, or did not do, to the Rubicon – is determined by what can be justifiedly believed about it because the *fact of the matter* is determined by what can be justifiedly believed about it.[40]

6.3 Truth as Coherence

Some authors who have been interpreted as advocating coherence theories of truth employ a primarily ontological notion of coherence; e.g., Spinoza's and Bradley's coherence theories of truth, if they even held such a theories, seem thoroughly metaphysical. Here we shall be concerned with theories that construe coherence as a primarily epistemological notion. To a first approximation: a coherence theory of truth is what becomes of the generic epistemic account of truth, (ET), when its definiens is spelled out by a coherence theory of epistemic justification (or some natural extension thereof). The term 'coherence' is commonly used in two different senses. In one sense it refers to a relation between an individual belief and a system of beliefs; in the other sense it refers to a holistic property that applies to a whole system of beliefs in virtue of the relations among its members. Accordingly, the coherence theory of truth is a times characterized as the view that a true belief is one that *coheres* with a designated system of beliefs, while at other times it is characterized as the view that a true belief is one that belongs to a *coherent* system of beliefs. As long as one remembers that there are two logically distinct notions of coherence in play, the theory might as well be characterized like this: x is a true belief $=_{\mathrm{df}}$ x coheres with a coherent system of beliefs. This gives the truth-coherentist the option to count a belief as true in virtue of its coherence with a coherent system to which it does not itself belong as a member.[41]

System coherence, which is the notion of primary importance, depends on a number of *coherence-conferring virtues*. Logical consistency is minimally necessary but far from sufficient. In addition, a coherent system is expected to be comprehensive; it must be a rich system containing observational, memory, introspective, and self-evident beliefs, together with higher-level generalizations and very high-level theoretical beliefs. Moreover, the members of the system have to be tied together by a dense web of inferential, probabilistic, and explanatory relations – and certain global virtues, especially simplicity and conservativeness, are also frequently required. The truth-coherentist may hope to leave the business of spelling out the inner workings of coherence to epistemologists; after all, it seems a coherence theory of truth should be little more than an extension of a coherence theory of epistemic justification. There are, however, some difficulties with this idea. A glance at the two most prominent coherence theories of justification – Lehrer (1990) and BonJour (1985) – reveals that they both make ample use of truth-linked notions. Beliefs about one's own reliability, about the reliability of one's own cognitive faculties, and about the truth-likelihood of one's own beliefs play a crucial role in both theories. Since these theories are not developed to serve as accounts of truth, this does not give rise to any internal problems. It does, however, make it difficult to see how such a theory could be adapted to serve as an account of truth.

The most serious difficulty, however, arises right at the first step: epistemic coherence, like all justification, is relative to persons (and times). The truth-coherentist has to specify *whose system* is supposed to be the one that is relevant to the truth of a given belief. The first idea is to say: S's belief is true $=_{\mathrm{df}}$ it belongs to the maximally coherent subsystem of beliefs constructible from S's total belief-set (the maximally coherent subsystem will be the subset of S's total belief-set that possesses the best mix of coherence-conferring virtues to the highest degree). Granting for the moment that there is only one such maximally coherent subsystem per person, the proposal has the unacceptable consequence of allowing each of a pair of logically conflicting beliefs to be true. After all, it is entirely possible that S's belief that p coheres with his maximally coherent subsystem, while S*'s belief that not-p coheres with her maximally coherent subsystem. To avoid contradiction, an advocate of this sort of account has to embrace a form of subjective relativism about truth, defining the relative notion *truth for S* instead of truth: S's belief that p is true *for S* $=_{\mathrm{df}}$ it belongs to the maximally coherent subsystem constructible from S's total belief-set. Attempting to avoid relativism, one might try to cast the net wider and specify the relevant system as one whose members are not actually believed by any single person. Many options are available here. To give just a few examples: S's belief is true $=_{\mathrm{df}}$ it coheres with the maximally coherent subsystem of (*a*) the beliefs held by the scientists, or (*b*) the statements underwritten by science, or (*c*) the beliefs held by one's cultural peers, or (*d*) the beliefs held by mankind. (Further specifications will be required, e.g.: Does (*b*) refer to the science of today, of this month, this year? What exactly counts as science? etc.) Not all these options actually avoid the problem. Since it is possible for two persons from different cultures to have logically conflicting beliefs that cohere with the systems of their respective cultures, option (*c*) will require embracing some form of cultural relativism about truth. Moreover, all these options talk about *the* maximally coherent subsystem; hence, they all face the problem of *uniqueness*. One cannot simply presuppose that there will always be a single subsystem of a given total belief-set that is more coherent than all other subsystems of that set. If there is more than one maximally coherent subsystem constructible from a total belief-set, then the proposals above entail, absurdly, that there are no true beliefs whatsoever. If the uniqueness requirement is weakened (if the 'the' is loosened to a 'some'), then there is the possibility that each of two logically conflicting beliefs might cohere with *some* maximally coherent subsystem constructible from a total belief-set. The problem is difficult to solve because there are so many coherence-conferring virtues, generating many different ways of weighting their relative import. Moreover, the demand to be met is very stringent. It is logically impossible for each of a pair of contradictory beliefs to be true. The coherentist has to show that it is logically impossible for two subsystems of a given belief-set to satisfy the coherence-conferring virtues to a maximal degree – a daunting task.[42]

There are three types of responses to the problem of uniqueness. The first is *optimism*: the scientists will come to agree, sooner or later (cf. Hempel 1935, 57). But optimism is besides the point. What needs to be shown is that it is not possible for them to go on disagreeing. The second response is more sophisticated: If system A and system B satisfy all the coherence-conferring virtues to a maximal degree, then A is *identical* with B (cf. Blanshard 1941, 276ff.). The idea is that, if "two" theories are equally comprehensive, equally explanatory of all the data, equally conservative and simple, and so on, then they are really just two versions of the same theory. It is comparatively easy to see how this reply is supposed to work, provided one thinks of truth bearers as sentences, and provided one is prepared to accept some form of *meaning holism* – the view that the meaning of a sentence is determined by its position in the sentence-network (theory) to which it belongs. If the sentence 'there are witches' coheres with system A and the sentence 'there are no witches' coheres with system B, and if A and B are maximally coherent, i.e., express the same theory, then, on this view, the two sentences simply do not have contradictory meanings despite their grammatical appearance. How the idea is to be applied to beliefs, however, is not so easy to see. Say the belief that there are witches coheres with belief system A and the belief that there are no witches with system B. What could it even mean to say that these two beliefs do *not* contradict each other? The claim that, when maximally coherent, system A is the same as system B seems to boil down to the mere stipulation that it is not possible for the one belief to cohere with A and the other to cohere with B. The third response to the uniqueness problem – which may well be combined with the second – is to turn towards *idealized* systems: the truth of a belief is determined "ultimately by its coherence with that further whole, all-comprehensive and fully articulated, in which thought can come to rest" (Blanshard, 1941, 264). The idea here is that the uniqueness problem will disappear, if truth is defined as coherence with "the" system of beliefs that we (scientists, mankind) would adopt, if we were able to make our theories as perfect as possible. A coherence theory of truth that takes this form belongs to a type of epistemic approach that deserves separate treatment.

6.4 Epistemically Ideal Conditions

An important version of the epistemic approach to truth derives from C. S. Peirce: "The opinion which is fated to be ultimately agreed to by all who investigate is what is meant by truth" (1878, 38); "The truth of the proposition that Caesar crossed the Rubicon consists in the fact that the further we push our archaeological and other studies, the more strongly will that conclusion force itself on our minds for ever – or would do so, if study were to go on for ever" (1901-02, 718). As it stands, this is more an expression of considerable intellectual optimism than a satisfactory account of truth; taken literally, these formulations do not even require that future investigations should proceed in any rational manner at all. Hilary Putnam has made a proposal that is close to the spirit of Peirce's idea but constitutes a significant improvement: "'Truth'...is some sort of (idealized) rational acceptability – some sort of ideal coherence of our beliefs with each other and with our experiences *as those experiences are themselves represented in our belief system* – and not correspondence with mind-independent or discourse-independent 'states of affairs'" (1981, 50f.); "Truth is an *idealization* of rational acceptability. We speak as if there were such things as epistemically ideal conditions, and we call a statement 'true' if it would be justified under such conditions" (1981, 55).[43]

Although Putnam emphasizes that he is not trying to give a definition of truth but merely an "informal elucidation" (1981, 56), one may still investigate the merits of this proposal when taken as a definition:

(IC) x is true/false $=_{\mathrm{df}}$ if anyone were in epistemically ideal conditions with respect to x, they would be justified in believing x/the negation of x.

The definition has the consequence that it is impossible to hold a true (false) belief whose truth (falsehood) would not be disclosed under epistemically ideal circumstances. Opponents are quick to point out that this does not seem impossible at all. Also, the issue of underdetermination of high-level theories by whatever counts as evidence becomes relevant again (see section 6.2): What is the reason for thinking that such underdetermination *must* resolve once we are in epistemically ideal circumstances? Setting aside these more external criticisms, attention focuses on the crucial role played by *epistemically ideal conditions* – a notion that is clearly in need of some further clarification. Can it be characterized without reference to truth or falsehood? In particular, What reason could there be for thinking that it is impossible to have justified but false beliefs under ideal conditions, if these conditions are not tacitly understood as conditions in which all our justified beliefs are true? To respond to this worry, ideal epistemic conditions have to be characterized in purely evidential terms: S is in epistemically ideal conditions with respect to x iff S is in possession of *all* evidence *relevant* to x. Nothing short of this will suffice. For, if S does not possess all the relevant evidence, then there could be some evidence that would defeat her justification, so that she would not be justified if she were in possession of that evidence. Hence, a weaker characterization of ideal epistemic conditions would reopen the door to the problems arising from the relativity of justification – precisely the problems that (IC) was intended to avoid: different persons could be justified in holding logically conflicting beliefs relative to their respective suboptimal epistemic circumstances. An ideal epistemic condition with respect to a particular belief has to be characterized as one in which there is no further evidence to be had that is relevant to the belief.

There are two serious problems with this characterization. First, as William Alston (1996, 205) points out, it is obscure what possessing all evidence relevant to a belief x is supposed to amount to. Does it involve believing *every* proposition that bears evidentially on x? Surely not, for that would involve holding lots of contradictory beliefs. Does it involve believing every *true* proposition that bears evidentially on x? This characterization is not available to an advocate of (IC). So maybe it involves believing every justifiable proposition that bears evidentially on x. But justifiable for whom? We are back to the relativity of justification (logically conflicting beliefs may be justifiable for different persons) and it won't do to say "justifiable for someone in ideal epistemic conditions," for that would amount to spelling out ideal epistemic conditions in terms of ideal epistemic conditions. Second, as Crispin Wright (1992, 45) points out, it is difficult to see that the 'relevant' in 'possessing all relevant evidence' can impose any restriction at all. Evidence is a highly *holistic* property. Whether e is evidence for a person's belief depends on the person's background beliefs: anything can be evidence for anything to someone given the right background beliefs (a mere falling apple can be evidence for the workings of the planetary system).[44] Consequently, the idea of possessing all evidence relevant to a particular belief boils down to the idea of possessing all evidence, period. So it turns out that the conception of ideal epistemic circumstances required for (IC) is far removed from any epistemic circumstances human beings could possibly find themselves in. A being that possesses all evidence cannot be a human being; indeed, it is hard to see how any being but God could reasonably be said to possess all evidence. It seems (IC) comes down to a definition Spinoza might have approved of: x is true $=_{\text{df}}$ if God did exist, He would be justified in believing x.

Alvin Plantinga has uncovered a crippling defect that afflicts all accounts of truth along the lines of (IC), no matter how 'epistemically ideal conditions' is spelled out in detail. The gist of his argument can be stated fairly briskly: If I were in epistemically ideal conditions with respect to the statement *I am in epistemically ideal conditions*, then I would be justified in believing that I am in epistemically ideal conditions – otherwise the conditions would not be ideal with respect to that statement. So, according to (IC), it is true that I am in epistemically ideal conditions, hence, according to (IC), I *am* in epistemically ideal conditions – *reductio*.[45]

6.5 Conceptual Primacy

Many epistemologists hold that the notion of truth is prior to any epistemological notions. We cannot make sense of *epistemic* justification as distinct from moral or pragmatic justification without conceiving of it as *a means to truth*. Epistemic justification is an evaluative notion, and what makes it distinctly epistemic is that it is used to evaluate beliefs relative to the cognitive goal of attaining truth and avoiding falsehood. Defining truth in terms of justification is more than pointless on this view, because such a definition tries to invert a fundamental conceptual dependence relation. If anything, truth might be used to define epistemic justification – as is proposed by reliabilists (cf. Goldman 1986). Moreover, defining truth in terms of justification must completely undermine the theory of justification itself. Such a theory consists basically in a specification of a set of standards that our beliefs have to meet in order to be adequately justified. But the proposed standards must not be arbitrary. That is, the theory has to be able to answer the question why just *these* standards are the *correct* standards of justification; and the only satisfactory way to answer this question is by showing that the proposed standards are adequately *truth-conducive*. Redefining truth in terms of justification short-circuits this vindication and deprives it of all force: *any* set of epistemic standards, however arbitrary, can be "vindicated" in this question-begging manner (cf. BonJour 1985, 5-10, 108-110). So, at bottom, the claim is that we cannot make any sense of notions like justification, evidence, warrant, verification, without thinking of them as indicators of *truth*. The response to be expected from advocates of the epistemic approach is easily stated: On the contrary, justification, evidence, warrant, verification, and especially, rationality and reason, are conceptually prior to truth – it is truth and its cognates (reference, reliability, probability) that cannot be understood without anchoring them in epistemic notions. The view is typically combined with a late-Wittgensteinian view according to which the correctness of our epistemic standards is ultimately grounded in our ordinary social practices of giving and receiving reasons for our actions and assertions. This issue of conceptual priority between the realms of the epistemic and the alethic marks one of the deeper divides in contemporary analytic philosophy. In the background lurks a maybe even deeper divide marked by disagreement over conceptual priority between the *normative* and the *descriptive*. The champions of the normative construe '*x* is true' as intrinsically evaluative, as expressive of the speaker's evaluation that asserting *x* is proper or correct or responsible. The champions of the descriptive construe '*x* is true' as paradigmatically fact-stating. They will argue that it is a fallacy to go from the premise that truth is (often) our cognitive goal to the conclusion that '*x* is true' is normative rather than descriptive; this is like moving from the premise that people tend to desire large amounts of money to the conclusion that 'Bill Gates has large amounts of money' is normative rather than descriptive.[46]

The late-Wittgensteinian theory of meaning is sometimes seen as leading fairly directly to an epistemic account of truth. According to this theory, the meaning of a term is to be explained in terms of the conditions for its *proper use*. Since it is held that it is proper to affirm a statement's truth if and only if one has adequate justification for believing it, it is natural to infer that the meaning of 'true' must be equated with adequate justification or some close relative (compare projects (iii), (iv) and (vii) of section 2). Opponents will insist that this fatally conflates the definition of truth with the test of truth. Setting such external criticisms aside for the moment, it appears that there is an interesting internal difficulty, one that is maybe not sufficiently appreciated. Let it be granted, for the sake of argument, that the proper-use condition for 'true' is indeed given by the equation: it is proper to affirm that it is true that p iff one has adequate justification for believing that p. A moment's reflection on human fallibility indicates that it is quite proper to affirm: "Some of my adequately justified beliefs may well be false," and "I may well have unjustified beliefs that are true." On the face of it, this suggests that there must be some mistake in the inference from the proper-use condition for 'true' to the thesis of the epistemic meaning of 'true' – even by the lights of a proper-use theory of meaning. The inference is too narrowly focused on uses of 'true' where it is applied to individual statements as opposed to uses within generalizations.

A related issue concerns the nature of epistemic justification. If truth is defined epistemically, i.e., by some specification of the generic formulation (ET) from section 6.2, then the theory of truth is but a spin-off of epistemology. But which epistemology? Can every theory of justification be supplemented with (ET) to yield an account of truth? No. In fact, one may even question whether (ET) can be combined with any of the available theories of justification. ***Reliabilism*** would evidently turn into nonsense if truth were defined along the lines of (ET); after all, it is the core thesis of reliabilism that justification has to be explained in terms of reliability which is directly defined with reference to truth (a belief-forming process is reliable iff it tends to produce true beliefs rather than false ones). Some versions of *foundationalism* explicitly require that basic beliefs – the beliefs from which all justification ultimately derives – have to be infallible. Since infalliblity is defined in terms of truth, such versions of foundationalism cannot have any use for (ET) either. That leaves fallibilist versions of foundationalism; and it leaves *coherence* theories of justification. But even for these theories it is not obvious that they can be meaningfully combined with (ET). All theories of justification tend to invoke the logical notions of *entailment* and *consistency*, notions that are normally defined directly in terms of truth. So the question whether any epistemological theory can still make sense when combined with (ET) is closely tied-up with the question whether fundamental logical notions can be understood without essential reference to truth. This connects with the dispute over conceptual priority. The conceptual priority of the epistemic requires an epistemology that has no need for truth – not even to make sense of logic.[47]

6.6 Anti-Realism and Realism

Discussions of epistemic theories of truth lead into the debate over anti-realism along the very same road described earlier with respect to pragmatic theories of truth (see section 5). Take some candidate for an epistemic definition of truth, say, a notion of in-principle justifiability, and assume that this notion is not tacitly (and circularly) defined in terms of truth. Opponents of the epistemic approach will begin their attack by invoking intuitive counterexamples to the (necessity of the) equation "it is true that p iff it is in principle justifiable that p" – the counterexamples will often elicit subjunctive or modal intuitions. To bolster their arguments and make the disparity between the two notions come out more vividly, they will then point out that instances of the necessary principle (T) – it is true that p iff p – fail (to be necessary) when 'in-principle justifiable' is substituted for 'true'. In response, the advocate of the epistemic definition could give up on (T); but that would mean giving up on the claim that her theory is a theory of truth (moreover, it is in practice virtually impossible to write about issues concerning truth and reality without making constant use of (T) or of some closely related principle). So the advocate of the epistemic definition will insist on her identification of truth with in-principle justifiability *and* on (T). This will commit her to some form of anti-realism, because, in the face of the intuitive counterexamples, the double-insistence can be sustained only by the claim that whether p must depend on whether believing that p is in principle justifiable – whether dinosaurs were warm-blooded must depend on whether believing that dinosaurs were warm-blooded is in principle justifiable. The precise form and the extent of the anti-realist commitment will be determined by the details of the proposed epistemic analysis of truth.

It should be noted that the dispute between realists and anti-realists is likely to remain, even if there should be a notion of in-principle justifiability that seems intuitively coextensive, or even necessarily coextensive, with truth. For even then the parties to the dispute will likely disagree over what Crispin Wright has aptly called the *Euthyphro contrast*: Are true beliefs true *because* they are in principle justifiable? or are they in principle justifiable *because* they are true? The anti-realist is committed to the former while the realist is committed to the latter.[48]

Michael Dummett (1978) suggests that a realism/anti-realism dispute about a certain domain is best understood as a dispute over the notions of meaning and truth appropriate to statements about that domain: the realist should be understood as assigning to the relevant statements truth conditions as their meanings, whereas the anti-realist should be understood as assigning verification conditions. Another way to put this, according to Dummett, is to say that realists employ a non-epistemic notion of truth while anti-realists employ an epistemic notion of truth. Dummett's proposal to tie the realism/anti-realism dispute definitionally to issues in the theories of meaning and truth has met strong opposition and has generated spirited discussions about how to think of the difference between realists and anti-realists. Philosophers who are generally regarded as "anti-realists" tend to take a favorable stance towards Dummett's proposal, whereas philosophers who are generally regarded as "realists" reject it and opt for a more traditional characterization along the lines suggested by Peirce (1878, 36): "Thus we may define the real as that whose characters are independent of what anybody may think them to be." This is a good approximation, but it does not clearly mark idealism as a form of anti-realism, for idealism maintains that material objects are *constituted* by mental objects or events (ideas). So, a rough version of the traditional characterization of anti-realism could go like this: an anti-realist about Fs holds that (i) whether something is an F depends on whether someone believes that it is an F, or he holds that (ii) the Fs are constituted by mental states – which part of this criterion is more appropriate will typically depend on the subject matter at hand. The basic idea is that anti-realists about Fs regard Fs as mind-dependent, either doxastically or constitutionally.[49]

The traditional characterization of realism and anti-realism is purely metaphysical. But even traditionalists may have the lingering feeling that Dummett is on to something: Are the debates about the nature of truth and the debates about realism vs. anti-realism not just two sides of the same coin? Four points may help clarify this matter. *First*, pragmatic and epistemic definitions of truth have anti-realist consequences *provided* principle (T) is in play. Without (T) there is no necessary connection between the *truth* about the metabolism of dinosaurs and the metabolism of dinosaurs. Since (T) is so obvious, it is easy to overlook it and to think of pragmatic and epistemic definitions as being anti-realist all by themselves. *Second*, the anti-realist consequences of pragmatic and epistemic definitions – in conjunction with (T) – are usually not immediate. Whether believing x is useful or justifiable does not obviously depend on whether anyone *believes* that believing x is useful or justifiable. But there will be anti-realist consequences right around the corner, for whether it is useful or justifiable to believe x does depend on our goals, experiences, background beliefs, cognitive capacities (etc.); hence the nature of dinosaur metabolism will equally depend on our goals, experiences, background beliefs, and cognitive capacities.[50] *Third*, there is a debate about the existence of mind- and language-independent truth bearers, i.e., about propositions in the serious sense. This debate, which is about the nature and existence of *truths*, should be distinguished from the debate about the nature of *truth*. Advocates of pragmatic and epistemic accounts of truth tend to deny the existence of serious propositions, which seems quite appropriate given their overall views. But a realist about truth could also deny the existence of mind-independent *truths* (i.e., of serious propositions). She may want to hold that only beliefs and statements are bearers of truth and falsehood and may reject any attempt to construe beliefs and statements as relations to serious propositions. What a realist about truth should not deny is that beliefs and statements are true or false independently of whether we think that they are. *Fourth*, a correspondence definition of truth, taken by itself, is not an expression of realism about material objects or any other subject matter. To be sure, talk of correspondence to reality or facts does suggest a realist attitude; but this suggestion can be canceled without inconsistency. As Kirkham (1992, 133) puts it, correspondence definitions, taken by themselves, are only *quasi realist*. An anti-realist may embrace a correspondence definition of truth *and* maintain that the reality to which a truth corresponds is mind-dependent.[51]

6.7 The Epistemological Argument

Historically, the most important motivation for epistemic theories of truth derives from an epistemological argument *against* the correspondence theory. Basically, the line of reasoning is that a correspondence theory of truth must inevitably lead into global skepticism because the required correspondence between our beliefs and mind-independent reality is not ascertainable. Some epistemic account of truth is then offered on the grounds that it avoids global skepticism. Note that the argument interprets the correspondence theory as an expression of metaphysical realism. As we have seen at the end of the previous section, this is a mistake: the correspondence theory is realist only when realism is explicitly added to it, i.e., it isn't realist at all when taken by itself. But since the argument is so popular, let us gloss over this point and assume that the correspondence theory is meant as a statement of realism. The epistemological argument can be interpreted in a number of different ways.

On one interpretation the argument amounts to the objection that a correspondence theory fails to establish justification infallibilism, coupled with the assumption that infallible justification is required for knowledge. The latter assumption is of course highly contested – see section 6.1. Moreover, most epistemic theories of truth do not really establish justification infallibilism either. They only establish "optimal-justification infallibilism," for they define truth as some form of ideal or optimal justification (as in-principle indefeasible justifiability, or as coherence with an ideally coherent system, or as justifiability under ideal circumstances). Since our ordinary methods of justifying beliefs about the external world – observation, induction, inference to the best explanation – do not yield any of these forms of optimal justification when applied by beings like us, most epistemic theories of truth do not establish infallibilism with respect to the kind of justification that is actually available to us. The only epistemic accounts of truth that could be of help here are ones that identify truth with the ordinary, low-grade justification provided by our actual methods as actually employed by us in this world and in our ordinary circumstances. But identifying truth with this low-grade justification is too obviously absurd for such accounts to even be serious contenders as theories of truth.

On a second interpretation the epistemological argument is primarily concerned with removing the threat of hyperbolical skepticism, i.e., the brand of skepticism that is based on the evil-demon hypothesis or on one of its relatives. To what extent the argument speaks in favor of epistemic truth-theories on this score depends on how the threat is interpreted. Sometimes the evil-demon hypothesis is said to make knowledge about the world impossible on the grounds that we cannot know anything about the world as long as it is possible that all our beliefs about the world are false. This also presupposes some version of infallibilism and will be rejected by epistemological fallibilists. Moreover, optimal-justification infallibilism does not really remove the possibility that everything we believe is false. It only removes the possibility that our optimally justified beliefs are false. Since it is dubious whether we possess optimal justification for our beliefs about the world, it is not clear whether epistemic theories of truth are of much help here. However, the threat posed by hyperbolical skepticism is more often interpreted as deriving from an argument that does not obviously presuppose infallibilism: (1) I know I have a hand, only if I know that I am not deceived by an evil demon; (2) I do not know that I am not deceived by an evil demon; therefore, (3) I do not know that I have a hand. Now, on an epistemic account of truth, the evil-demon hypothesis cannot possibly be true because it is radically evidence-transcendent. This looks promising; for, based on an epistemic account of truth, I can answer premise (2) with (2*): I know that it is *not true* that I am deceived by an evil demon. But this success is not unambiguous. Since the evil-demon hypothesis is evidence transcendent, its negation is evidence transcendent too, which means that its negation is not true either. So I cannot infer the negation of (2) from (2*). After all, the negation of (2) would require that I know that I am *not* deceived by an evil-demon. Since knowledge requires truth, such knowledge would require that the negation of the evil-demon hypothesis be true, which is not possible on an epistemic account of truth. The overall result would seem to be that it is quite obscure how much an epistemic account of truth can actually achieve against the evil-demon.

The most influential version of the epistemological argument is the "circle of belief" argument: We cannot step outside our own minds to compare our thoughts with mind-independent reality. We cannot get outside the circle of our ideas. Yet, on a correspondence theory of truth, this is precisely what we would have to do to gain knowledge. We would have to access reality as it is *in itself* – independently of our cognition of it – and determine whether it corresponds to our thought. Since this is impossible, since all our access to the world is mediated by our cognition, the correspondence theory makes knowledge impossible. Kant puts this much better (1800, intro. vii): "According to [the correspondence definition of truth] my cognition, then, to pass as true, shall agree with the object. Now I can, however, compare the object with my cognition only *by cognizing it.* My cognition thus shall confirm itself, which is yet far from sufficient for truth. For since the object is outside me and the cognition in me, I can judge only whether my cognition of the object agrees with my cognition of the object." For ease of reference, let us call this sort of argument the "Kantian" argument – always remembering that it is very popular and comes in many different versions. It brings up a host of issues in epistemology, the philosophy of mind, the theory of truth, and general metaphysics. All that can be done here is to hint at a few pertinent points.[52]

To begin with, the Kantian argument – as it is understood here – is an argument *from* epistemology *to* metaphysics and should be distinguished from purely metaphysical arguments against correspondence. Assuming the unacceptability of skepticism, the Kantian argument derives the metaphysical conclusion that a realist correspondence theory must be rejected from the epistemic premise that it is impossible to *ascertain* a correspondence between thought and mind-independent reality: thoughts (judgments, beliefs, statements, concepts, cognitions, ideas, etc.) can only be compared with other thoughts and never with mind-independent reality. The argument has to be formulated with some care. For it is in constant danger of deteriorating into a version of "Berkeley's Gem": Since I can only have cognitive access to a thing *as cognized* by me, I cannot have access to the thing *in itself*. The premise is surely true, if it reminds us that we cannot perceive a thing without perceiving it. From this tautology, however, it hardly follows that we cannot perceive a thing that exists independently of our perceiving it, or that we cannot perceive any qualities it has independently of our perceiving that it has those qualities. The premise must amount to more than a mere tautology. Appropriately stronger versions are vigorously opposed by realists on the grounds that they make the premise false or the argument question begging.[53]

Correspondence theorists of realist persuasion will object to the use the Kantian argument makes of metaphors like "comparing" and "accessing". In the basic case of observational beliefs our "accessing" reality amounts to reality causally impinging *on us*. The physical properties of a material object cause changes in my perceptual mechanism – changes that lead to my having various sensations and eventually to the formation of an observation belief, say, the belief that there is something black in front of me. Reliabilists, foundationalists, and coherentists have different views about what it takes for such an observation belief to be justified. The reliabilist holds, roughly, that my perceptual mechanism has to function reliably and that I should not have any defeating background beliefs (i.e., beliefs indicating there is nothing black in front of me after all, or beliefs indicating that my perceptual mechanism is not functioning reliably in this case). The foundationalist may hold that my being-appeared-blackly-to is prima facie justification for my observation belief and that this prima facie justification constitutes adequate justification provided there are no defeating background beliefs. The coherentist will hold that to be justified the observation belief has to cohere with the maximally coherent subsystem of my total belief set. Now, assume my observation belief satisfies whatever conditions of justification are deemed to be the right ones (and that my justification is not "Gettierized"): if my belief that there is something black in front of me is also true, i.e., if it corresponds to reality, then it constitutes knowledge. It is hard to see at what point the metaphors of "comparing" and "accessing" can gain a real foothold in any of these accounts. Not even epistemological coherentism leaves much room for these metaphors. For, according to the coherentist, its coherence with the system *is* what justifies the belief: add correspondence to reality and knowledge ensues (unless there are Gettier problems). A defender of the Kantian argument might object that all these forms of justification fail to *guarantee* that the belief conforms to the external situation – but this is just the dispute between infallibilism and fallibilism all over again. A defender of the Kantian argument may also point out that the observation belief is not a cognitively uncontaminated presentation of raw data but a *cognition* that results from the application of concepts (*black*, *object*, etc.) to sensory input. But there is nothing here with which the correspondence theorist is likely to disagree. Sure enough, we cognize an object *by cognizing* it – the point is: we are cognizing *an object*. Let us take a case that seems more vulnerable to the Kantian argument, one that involves a general belief, e.g., the belief that all swans are white. When that belief is refuted by the observation that *this* swan is black, is this more than a mere comparison between beliefs? Again, realist correspondence theorists of all epistemic persuasions will say Yes. My observation belief that this swan is black is a perceptual response to the causal influences of my environment (e.g., to a black swan). No doubt, being a perceptual response means that it is a *cognitive* response, resulting from the application of concepts and background knowledge about swans to my sensory input from the world. Now, to the extent that the language of "comparing" is appropriate here at all, the correspondence theorist will hold that the observation belief *mediates* the "comparison" between the general belief that all swans are white and the black swan. The observation belief is itself a cognitive response to reality; and the general belief is "compared" with that reality *by* "comparing" it with the observation belief – that is how "comparing" works in such cases (i.e. inference). The defender of the Kantian argument will insist that such a *mediated* comparison between cognition and reality does not count as a *real* comparison, especially since the mediator is itself a cognition: only *immediate* comparisons are real comparisons. But the correspondence theorist will demur. She will hold that mediated comparisons between cognitions and reality are still comparisons, especially when the mediator is itself a cognition *of reality*. Compare: "You cannot really see a black swan; all you can *really* see is your idea of a black swan." – "I can see a black swan *by* having an idea of a black swan that is properly caused by a black swan."[54]

The Kantian argument can be divided into two parts. Part (A) says that there is something we cannot do, namely ascertain the correspondence between thought and reality. Part (B) says that we have to do it, on a realist correspondence theory of truth, if we are to obtain knowledge. With respect to part (A), one should ask whether the epistemic competitors of the correspondence theory actually enjoy any significant advantage when held to the same standards. Consider an account like (E_4), according to which a belief is true iff it is in principle possible for someone to have indefeasible justification for it. Is it really any easier to ascertain that a belief satisfies this condition than to ascertain that it corresponds to a fact? Consider an account of truth in terms of coherence. How easy is it to ascertain that a system of beliefs does not harbor some hidden inconsistencies – some belief which, by some devious route, entails the falsehood of some other belief of the system? How easy is it to ascertain which of two very comprehensive systems is the simpler and more conservative? Putnam's proposal, (IC), hardly needs separate mention. Since we are not in epistemically ideal circumstances, there is little hope of ascertaining what we would be justified in believing, if we were in epistemically ideal circumstances.

The line of reasoning underlying part (B) of the Kantian argument faces two difficulties. Consider again the rough version: If knowledge requires truth and truth is correspondence, then we have to know that our beliefs corresponds to reality, if we are to know the truth. There are two assumptions implicit in this argument – both are dubious:

(i) S knows *x* only if S knows that *x* is true;
(ii) if truth = F, then S knows that *x* is true only if S knows that *x* is F.

Recall the classical definition of knowledge, (K), from section 6.1. It tells us that S knows something only if S believes it and it is true. The definition does not make the requirement described in (i). It only requires S's belief to *be* true; it does not require S to *know* that her belief is true. The latter would be a requirement for *knowing that one knows.* One might want to defend assumption (i) with the "KK" thesis: S knows that p only if S knows that she knows that p. But this thesis is highly contentious and is rejected by a good number of contemporary epistemologists. Alternatively, one might want to defend (i) with the claim that knowing that p is the same as knowing that it is true that p. But what if S does not have the concept of truth? The claim would entail the thesis that it is not possible to know anything – to know any truth – without possessing the concept of truth. The thesis is not very plausible. Moreover, if knowing that p and knowing that it is true that p are the same state, then (i) by itself cannot lead to any epistemological difficulties. For one can then simply know that it is true that p *by* knowing that p. The weight of the argument must rest on assumption (ii).

Assumption (ii) is highly implausible. This comes out best when considering comparable requirements. Water = H_2O. By the standards of (ii), nobody who does not know that water is H_2O can know that the Nile contains water – which means, of course, that until fairly recently nobody knew that the Nile contained water. Similarly, until fairly recently, nobody knew that there were stars in the sky, whales in the sea, or that the sun gives light. All of this is quite absurd (or simply presupposes skepticism). Note also, even if one knows that Water is H_2O, one's strategy for finding out whether the liquid in one's glass is water does not have to involve chemical analysis. Tasting it and/or remembering that it came from the bottle labeled 'Water' can be quite sufficient. This problem with assumption (ii) throws serious doubt on the Kantian argument. It shows that the truth of the correspondence theory does not entail that we have to know that a belief corresponds to a fact in order to know that it is true – much less does it entail that we have to *first* know that a belief corresponds to a fact *before* we can know that it is true. Of course, if truth is correspondence to a fact, then obtaining knowledge amounts to obtaining a belief that corresponds to a fact. But our strategy for how to go about obtaining knowledge does not have to be a strategy of "comparing" beliefs with facts. Our strategy can be one of making observations and experiments, of deducing logical consequences from what we already know, of listening to reliable testimony, or of doing our best to make our system of beliefs coherent and comprehensive, etc. Assumption (ii) fails; hence, the basic line of reasoning that is essential to the Kantian argument is fallacious and has to be abandoned.[55]

7. ALETHIC RELATIVISM

The central thesis of relativism about truth – *alethic* relativism – is that truth is perspectival: a proposition (statement, belief) is not true or false *tout court* but true or false *for* a person, *for* the members of a society, *in* a tradition, or *in* a conceptual framework. As this characterization already indicates, there are many different versions: there is personal relativism, there is cultural or societal relativism, and there are various versions that relativize truth to less "tangible" perspectives – to traditions, conceptual frameworks, or forms of life. Alethic relativism should be distinguished from ethical and epistemological relativisms which maintain that the standards of ethical or epistemic evaluation are relative. These forms may be implied by, but do not imply, relativism about truth. Sometimes alethic relativism is associated with skepticism. In particular, some people seem to entertain skeptical worries based on their relativism about truth. If all truth is relative, they reason, then there cannot really be any knowledge, for knowledge requires absolute truth. Although this line of reasoning has the ring of plausibility, it is hard to see how a relativist about truth could have any use for it. If the very notion of truth is relative, then knowledge requires merely relative truth. If anything, relativism would seem to make it easier, rather than harder, to acquire knowledge, at least insofar as the acquisition of truth is concerned.

The essence of alethic relativism lies in its power to function as a *universal conflict solvent.* Any apparent conflict between two parties – one advancing the claim that p and the other advancing the claim that not-p – can be dissolved by relativizing truth so that both parties can be said to be "right." To put this slightly differently, the alethic relativist is able to hold that one and the same proposition can be both true and false (in a sense) without violating the law of contradiction: the proposition is true for S and false for S*.[56]

According to the relativist, 'true' is a contraction of 'true for' or 'true in', and the first question to ask is what these expressions are supposed to mean. Often '*x* is true for S' is just a paraphrase of 'S believes *x*' – the transition from the second to the first being mediated by 'S believes that *x* is true'. On this construal, relativism about "truth" turns out to be a misleading expression of the widely held view that different people can, and do, believe different things. The aim of this sort of verbal relativism is peaceful coexistence: although we disagree wildly, everyone gets to keep the honorific 'true' – maybe that will keep us from getting at each others' throats. But what if the relativist insists that 'true' really means nothing more than that? What if he holds that when we say '*x* is true' we just mean 'I believe *x*'? The obvious response is that this is clearly not what we mean. In addition, one can offer a diagnosis. It is indeed a crucial feature of 'true' that I am disposed to call *x* 'true' if and only if I believe *x*. But it is rash to infer that when I affirm '*x* is true' what I mean is 'I believe *x*'. The inference founders on my disposition to affirm: "Some, even many, of the things I believe are not true" – which would be inconsistent on this interpretation of 'true'. Sure enough, relativizing is at times an appropriate strategy for settling disagreements. It is appropriate to dissolve a quarrel over the question whether vanilla ice-cream tastes good by coming to the conclusion that there is nothing more to it than that it tastes good to me and bad to you. Moreover, due to the behavior of 'true', this *can* be paraphrased as 'it is true for me that vanilla ice-cream tastes good' and 'it is true for you that vanilla ice-cream tastes bad'. But this sort of relativism "from below" can motivate general truth relativism only if one is prepared to hold that *all* disputes about specific subject matters can be resolved like disputes about the taste of vanilla ice-cream. Note also that, once we have relativized tastes to tasters, we treat the result as absolute; we think that it is simply true that vanilla ice-cream tastes good to me and bad to you.[57]

Relativism is a position with perennial appeal, and considerations like the ones above have not been very effective in undermining this appeal. In the *Theaetetus* – the classic treatment of relativism – Plato sought a more forceful refutation; he sought to prove that relativism is inconsistent. But it is doubtful whether he was ultimately successful – and it seems fair to say that the last 2300 and odd years have not added all that much to the discussion. Plato faced Protagoras' *global subjectivist* relativism: "As each thing appears to me, so it is for me, and as it appears to you, so it is for you" (*Theaetetus* 152^a). He attacked this position on the grounds that it is self-refuting: If everything is relative, then it is true that everything is relative; but then one thing is not relative, namely the truth of global relativism – *reductio*. While this argument may have some effect on a young, inexperienced relativist, an old hand like Protagoras will be unmoved: of course, if truth is exempted from relativism, global relativism is incoherent, which just goes to show that truth is relative too. But Plato's argument does achieve something. It shows that Protagoras cannot coherently assert global/alethic relativism as simply true. Instead, he has to hold that *truth is relative* is true *for him*, and he must allow that it may be false for others. This diminishes the bite of relativism considerably, for it seems that no one's beliefs are actually threatened, if relativism is only true for the relativist. Still, this very popular "self-refutation" argument falls short of a proof that Protagoras' position is inconsistent. Plato has an additional argument, one involving beliefs about the future. According to relativism, if it seems to me today that tomorrow I will be feverish, then it is true for me today that tomorrow I will be feverish. But also, it is true for me tomorrow that I am feverish iff I feel feverish tomorrow. What if I don't feel feverish tomorrow? What happens to today's truth in that case? But Protagoras would have a response: relativize truth to persons *and* times. Finally, Plato makes the point – a point that has been frequently reiterated in one form or another – that relativism runs into deep difficulties with our ordinary understanding of a large number of concepts. Are not experts people who possess truth where others are wrong? What is going on in teaching and learning? in discussion and debate? What is disagreement? What is refutation? and perhaps most fundamentally: Isn't to assert something to put it forward as true, period? Can a relativist really make sense of such activities? This is a serious issue, and it is unlikely that the relativist can give satisfactory relativistic reconstructions of all these truth-linked concepts. Still, the advertised proof that global/alethic relativism is *inconsistent* is hard to come by.[58]

Since utility and justification are both relative to persons, alethic relativism has come up repeatedly during the discussions of pragmatic and epistemic accounts of truth; cf. sections 5, 6.2, 6.3, and 6.4. In each case, a proposed account was confronted with the objection that it allowed for logically conflicting beliefs of different persons to come out true, thereby effectively disqualifying itself as an account of *truth*. In each case, relativism might offer a tempting escape from this lethal difficulty because it allows for S's belief that p and S*'s belief that not-p to both be true – in a sense. And taking this escape route would no doubt be much easier than the more involved proposals discussed in the preceding sections. But relativism about truth is not an acceptable doctrine. The objections mentioned above are serious, and they are readily adapted to pragmatist and epistemic versions of alethic relativism. Admittedly, Plato's objections did not succeed in proving relativism inconsistent, and it is doubtful whether any such proof is to be had. But a position can be false, even if it is not inconsistent (philosophers are apt to forget that).

An additional objection can be raised. The central promise of alethic relativism – the promise to dissolve apparent logical conflicts – is a vulnerable point at which a purely internal objection can get a foothold. It is not clear that any relativistic *analysis* of truth actually guarantees the dissolution of all conflicts; for it is not clear whether any such analysis excludes the possibility that *x* may be both *true for S* and *false for S*. Consider Protagoras' subjectivist relativism. It is not really logically impossible that S believes *x* and also believes the negation of *x*. A person might have conflicting beliefs because she has compartmentalized her thoughts on, say, science and religion. Consider a pragmatist truth-relativist. It is not obvious that it might not be useful for S to believe *x* and useful for S to believe the negation of *x*. The one belief might be useful with respect to one set of goals and the other with respect to a set of conflicting goals. A coherentist truth-relativist – *x* is true for S iff *x* coheres with S's coherent belief system – is secure only if coherence is defined so that it requires relativized (to S) consistency. But consistency for S will be defined in terms of truth for S, so that the security is bought at the price of circularity.

As a final objection, one might wield (T) – it is true that p iff p – or one of its relatives, against relativism about truth. If the relativist cannot uphold (T), or something like it, then he is not talking about truth to begin with. Consider now:

(1) The proposition that a is F *is true for S* iff a is F;
(2) The proposition that a is F *is true for S* iff a is F *for S*.

The relativist cannot accept (1). For he wants to hold that one and the same proposition can be true for S and false for S*, which would entail that something can be both F and not-F and that *would* make relativism inconsistent. But the relativist can accept (2). Moreover, he can plausibly maintain that (2) must be the schema that is the right one for relativized truth. What else could it be? If truth is relative, predication must be relative too. This seems logically consistent; but it makes the position hard to understand. After all, the predicate 'is true for S' can itself be substituted for 'is F'; and since (2) has instances for persons S, S*, S**, and so on, we get multiply relativized truth predicates, like 'is true for S for S*', and 'is true for S for S* for S**', and so on. As Putnam puts it, "our grasp on what the position even means begins to wobble."[59]

Serious relativism about truth – the position discussed until now – should be distinguished from *trivial* relativism about truth. Trivial truth-relativism is a rather ordinary thesis that can take on the semblance of serious relativism. Consider the following seemingly relativistic claims involving truth applied to *sentences*:

(3) The sentence 'Socrates is snub-nosed' is true for S;
(4) The sentence 'Socrates is snub-nosed' is false for S*.

This could be an expression of serious relativism. But it could also indicate that the sentence 'Socrates is snub-nosed' does not mean for S what it means for S*. Maybe in S's language it means that Socrates is snub-nosed, whereas in S*'s language it means that Socrates is not snub-nosed, or that Socrates is a spy from Persia. In the latter case, (3) and (4) merely remind us that *sentence-truth* is relative to *meaning* – not a thesis that creates much stir among absolutists about truth. According to this trivial version of relativism, when a sentence does mean the same in S's language as it does in S*'s, then it is true for S iff it is true for S*. This is exactly what the serious relativist must deny. According to the serious relativist, a sentence can be true for S and false for S*, even if it has the same meaning for both.

Positions according to which truth is relative to conceptual frameworks, or conceptual schemes, or forms of life, often hover uncomfortably between serious and trivial relativism. They usually present themselves as genuine alternatives to absolutism about truth. But if S and S* have radically different conceptual schemes, then it is far from clear whether they can be used to make a case for more than trivial relativism. To make a case for serious relativism, one has to hold that S and S* can utter sentences with different truth values that mean the same thing. Yet, conceptual-framework relativists like to deny that S and S* can mean the same thing if they have radically different conceptual schemes. Consequently, this kind of conceptual-framework relativism is not a form of serious relativism, even though it is often presented as if it were.

8. DEFLATIONARY VIEWS

"Truth has no nature." This slogan expresses the spirit behind deflationary views of truth. It is directed against the traditional assumption that the predicate 'is true' stands for a genuine property – the property common to all and only the true beliefs, the property in virtue of which true beliefs are true. According to deflationists, the superficial grammatical similarity between 'is true' and other predicates of our language leads us into thinking that there ought to be a genuine explanatory theory of the nature of truth, like there is a genuine explanatory theory of the nature of water and magnetism. But 'is true' functions differently than most other predicates. It does not stand for a substantive property to be described or explained by a substantive theory. There is no substance to truth for such a theory to be a theory *of*; hence, an account of truth will not go beyond an account of our *notion* of truth, of the role played by the *word* 'true' in our language. According to deflationists, this is the lesson to be learned from an elementary observation: To say that it is true that snow is white amounts to (about) the same thing as to say that snow is white.[60]

8.1 Redundancy

The 13th century logician William of Sherwood may have been an early advocate of deflationism: "It is the same thing to say 'Socrates is running' and 'it is true that

Socrates is running'" (*Introduction*, chap. 1.23). Similar remarks were made a bit later by Frege: "One can, indeed, say: 'The thought that 5 is a prime number is true'. But closer examination shows that nothing more has been said than in the simple sentence '5 is a prime number'" (1892, 34); Ramsey: "It is evident that 'it is true that Caesar was murdered' means no more than that Caesar was murdered" (1927, 38); and A.J. Ayer: "We find that in all sentences of the form "p is true", the phrase "is true" is logically superfluous. When, for example, one says that the proposition "Queen Ann is dead" is true, all that one is saying is that Queen Ann is dead" (1936, 88f.).

Consider the biconditionals under (1) below and their *schema* (T$_P$):

(1) The proposition that 5 is a prime number is true iff 5 is a prime number.
The proposition that Queen Anne is dead is true iff Queen Anne is dead.
The proposition that Socrates is running is true iff Socrates is running.

(T$_P$) The proposition that p is true iff p

The schema displays the pattern, or form, of the biconditionals. They are substitution instances of the schema, that is, each biconditional is obtained from (T$_P$) by substituting a declarative sentence for the schematic sentence-letter 'p'. The authors quoted above advocate an *equivalence thesis*, saying that all biconditionals of the form (T$_P$) hold in virtue of an equivalence of some sort between what is said to the left of the 'iff' and what is said to the right of the 'iff'. The equivalence thesis suggests the *basic deflationary idea*: there is nothing more to the notion of truth than what is given by the substitution instances of (T$_P$). A glance at the biconditionals makes clear why such a view might be summarized in the slogan that truth has no nature. The biconditionals firmly *fail* to point to any property that is shared by all and only the true propositions which could be said to be the property in virtue of which they are true. If *they* specify all there is to truth, then truth has no nature and substantive accounts of truth are misguided.

The above characterization of deflationism is deliberately vague, for deflationists have held different views about the kind of equivalence underlying biconditionals of the form (T$_P$). According to a very strong deflationary view, these biconditionals hold because the sentences to the left of the 'iff' mean nothing over and above the sentences to the right: 'the proposition that p is true' means no more than 'p'. This is a radical *redundancy* or *disappearance* theory of truth. It maintains that the term 'is true' is redundant; that it could simply be stricken from our vocabulary without any significant loss (and without first defining it in other terms). Of course, deleting 'is true' in the items under (1) results in nonsense. But the radical redundancy theorist will hold that the reference to propositions in the items under (1) is spurious. Expressions of he form 'the proposition that p' do not function as subjects; they do not refer to anything at all. Since 'is true' is not really a predicate of anything, truth bearers are to be discarded. This means that the schema

(T) It is true that p iff p

will take center-stage. (T_P) will be regarded as nothing more than a long-winded version of (T) with 'it is true that' understood in the "bearerless" operator sense (cf. section 1.5). According to the redundancy theorist, truth is the bearerless operator-notion *it is true that*, and he maintains that this notion is redundant: 'it is true that' can always be erased without loss.[61]

Deflationists generally subscribe to some version of the equivalence thesis – to the very strong version mentioned above, or to a weaker version according to which biconditionals of the form (T_P) are true "in virtue of meaning" (analytic) understood in a way that does not require that their sides have the same meaning. Deflationists who are worried about putting too much stress on the difficult notion of meaning tend to emphasize epistemic versions of the equivalence thesis, saying that the biconditionals are platitudinous, trivial, self-evident, or a priori.

It is noteworthy that all such claims go much better with the instances of (T) than with the instances of (T_P). The former really do appear analytic, platitudinous, trivial, self-evident, or a priori. Although the instances of (T_P) are often treated as if they were as trivial as the instances of (T), this is a dubious practice. After all, the left-hand sides of the biconditionals in (1) imply existence claims not implied by their right-hand sides, i.e., claims of the form 'the proposition that p exists'. Whether one wants to uphold a meaning-equivalence thesis or merely a "triviality thesis," it seems (T_P) had better be existentially amended to something like: The proposition that p is true iff the proposition that p exists and p. One might try to avoid this complication with the claim that (T_P) is nothing more than a long-winded version of (T). But that would mean joining radical redundancy theorists and giving up on truth bearers altogether. A friend of truth bearers has to be more moderate than a radical redundancy theorist; he has to say, on the contrary, that (T) is to be analyzed in terms of (T_P).

Any deflationary approach to truth (strong or moderate) faces the *generality problem.* As it stands, the approach seems unable to account for generalizations and related "blind" and "indirect" ascriptions of truth, like 'Everything he asserts is true', 'What John just said is true', 'Gödel's theorem is true', 'That is true'. Logicians and philosophers are especially fond of generalizations involving truth, e.g., 'Every proposition that is entailed by a true proposition is true'; 'There are arithmetical sentences that are true but not provable'; 'The epistemic goal is to believe what is true and to avoid believing what is false'; 'A belief constitutes knowledge only if it is true'; 'A belief is justified only if it is produced by a belief-forming process that tends to produce true beliefs rather than false ones.'

To take a simple example, consider the generalization (2) below. (3) is a partial translation of (2) into standard predicate logic (with the variable ranging over propositions), (3a) is a paraphrase, and (3b) is one of is instances:

(2) Everything he asserts is true;

(3) $(\forall x)$(if he assert x, then x is true);

(3a) For every proposition, if he asserts it, then it is true;

(3b) If he asserts the proposition that Socrates is running, then the proposition that Socrates is running is true.

Evidently, simply eliminating 'is true' would produce nonsense. This observation already refutes the radical redundancy theory. Ramsey (1927, 39) and Ayer (1936, 89) proposed a modification. They wanted to employ schematic sentence-letters to paraphrase (2) without having to mention 'is true'. The natural idea is to turn to 'If he asserts that p, then p'. But this is just a schema. Although its instances are perfectly grammatical, the schema itself does not express anything at all. What seems to be needed is some way of generalizing what this schema attempts to convey, something like

(4) For every p, if he asserts that p, then p.

But it is questionable whether this makes any sense. The difficulty is to understand how the quantifier phrase 'for every p' interacts with the occurrences of the schematic letter 'p' in the rest of (4). Clearly, (4) cannot be interpreted in the way generalizations are ordinarily interpreted, i.e., as in (3a)-(3b), for that would yield nonsense again:

(4a) For every proposition, if he asserts that it, then it;
(4b) If he asserts the proposition that Socrates is running, then the proposition that Socrates is running.

Rather, the idea must be to so interpret (4) that it has (5) as one of its instances:

(5) If he asserts that Socrates is running, then Socrates is running.

There is indeed an interpretation of (4) that will yield just this result, but it is an unfriendly interpretation according to which (4) is a symptom of confusing talk about language with talk about the world. On this interpretation, (4) is a deeply confused attempt to express the well-formed metalinguistic generalization

(4c) Every substitution instance of the schema 'If he asserts that p, then p' is true,

which leads readily to (5), provided we recognize that 'Socrates is running' can be substituted for 'p', and provided the notion of truth is *already available*. The upshot is a dilemma for the deflationary approach: (4) is nonsense, see (4a) and (4b), or it presupposes the notion of truth, see (4c).[62]

The generality problem is recognized as posing a crucial challenge for redundancy theories and for all deflationary views about truth. Prima facie, generalizations (and other blind ascriptions) involving truth indicate that the "nature" of truth is not exhausted by the totality of biconditionals whose pattern is depicted by schemata like (T_P) or (T) or one of their analogues for sentences or beliefs. If deflationists cannot handle generalizations involving truth, then the deflationary thesis would appear to be undermined.

8.2 Prosentences

Grover, Camp, and Belnap's (1975) *prosentential theory of truth*, which builds on Prior (1971), construes truth-talk anaphorically. Consider the use of the pro*noun* in 'Mary was hungry, but *she* did not want to eat'. The pronoun is said to have "anaphoric reference" because it points to its antecedent, the noun 'Mary', from which it inherits its "real" reference – it does not, as it were, have any real reference of its own. At times we use 'so' in a similar way, but as a pro*sentence*, e.g., 'Einstein said that the speed of light is the limit; if *so*, then we won't ever get very far'. Grover, Camp, and Belnap propose that 'it is true' and 'that is true' function as prosentences in the deep-structure of English. Thus 'It is true that Socrates is running' can be construed as 'Socrates is running. That is true'. And 'If it is true that Mary is hungry, then she should eat' becomes 'Mary is hungry. If that is true, then she should eat'. Note that prosentences, like all pro-forms, inherit their semantic value anaphorically from their antecedents. Since the antecedent of a prosentence is a sentence, and since sentences do not refer to anything, the prosentence 'that is true' does not refer to a truth bearer. The prosentential theory promises a "bearerless" approach to truth. Consequently, 'is true' does not function as a real predicate and does not express a property of anything at all; it is merely a fragment of a prosentence. The thesis now is that all apparently predicative uses of 'is true' in English can be eliminated in favor of prosentential uses. Thus Grover, Camp, and Belnap advocate a conservative extension of the radical redundancy theory: although 'is true' is not redundant as fragment of a prosentence, it never functions as a predicate expressing a property.

Anaphoric pro*nouns* play an important role in generalizations, e.g., 'For every number, if *it* is even, then *it* is divisible by two'. Since the prosentential theory construes truth-talk as anaphoric, it can take such generalizations as the model for dealing with generalizations that involve truth. So, with the help of prosentences, the problematic quantification (4) is interpreted in the light of (6), i.e., by replacing the schematic sentence-letter 'p' with the prosentence 'it is true' and modifying the quantifier phrase:

(2) Everything he asserts is true;
(4) For every p, if he asserts that p, then p;
(6) For every proposition, if he asserts that it is true, then it is true.

In this way, the prosentential theory attempts to resolve the dilemma posed by (4). Sentence (6) makes sense. Moreover, (6) is not metalinguistic, for the prosentence 'it is true' is not *about* anything. (6) employs 'is true' as part of a prosentence, but not as a predicate. Hence, so the proposal, there is no reason for thinking that generalizations involving 'true' involve a substantive notion of truth.

A common reaction to this proposal is: puzzlement. The problem is, of course, that (6) contains all these occurrences of 'is true' and the quantifier phrase 'for every proposition' and generally looks exactly like it talks about propositions and involves attributions of the property of being true to propositions. The difference is all in the stage setting. Grover, Camp, and Belnap *say* that the quantifier phrase does not refer to genuine truth bearers and that 'is true' is merely a fragment of a prosentence, but one wonders whether saying it makes it so. It is unclear how much weight the distinction between prosentences and predicates can bear. Maybe what the prosentential theory indicates is not that truth is not a property, but rather that there are prosentences that express properties. 'It is true' in (6) could be a prosentence attributing to the propositions expressed by its antecedents the property of corresponding to a fact. (Compare: 'Marian is hungry, but *he* does not want to eat', where the pronoun attributes the property of being male in addition to doing its job as a pronoun.) Unfortunately, it is not at all clear how one would go about deciding this issue.

One might also be worried about the claim that (6) is an analysis of (2). Remember that (4) only came in as an attempt to capture (2). If (2) is analyzed as (6), then (2) has grown an additional 'true'. Switching from assertion to belief will help bring out why this is problematic. On the prosentential analysis, 'S believes something that is true' amounts to 'for some proposition, S believes that it is true, and it is true'. It follows that S could not possibly have a true belief without having the concept of truth (or the concept *it is true*) – according to the analysis, it is incoherent to even entertain the idea that S could believe something that is true without having the concept of truth. But intuitively it does seem that there could be believers who lack the concept of truth.

The stance taken by the prosentential theory towards schema (T) raises a somewhat related worry. (Since the theory offers a "bearerless" approach to truth, (T) is the most directly relevant schema.) Remember that the apparatus of prosentences allows us to interpret quantifications like (4) by replacing the schematic sentence-letter 'p' with the prosentence 'it is true'. This means that (T) can be turned into the quantification (7) to be interpreted by (7a):

(T) It is true that p iff p;
(7) For every p, it is true that p iff p;
(7a) For every proposition, it is true that it is true iff it is true;
(7b) It is true that Socrates is running iff Socrates is running;
(7c) Socrates is running. That's true iff Socrates is running.

The substitution instances of (T) can all be said to be self-evident, necessary, and self-evidently necessary. The same can be said, *or so it seems*, of (7) interpreted as (7a). Now, (7b) is supposed to be an instance of (7a); and according to the prosentential theory, 'It is true that Socrates is running' is to be analyzed as 'Socrates is running. That's true' (see Grover, et al. 1975, sec. 2.4). This analysis leads to (7c). But (7c) is not a necessary truth, for it affirms that Socrates is running. Something has gone wrong somewhere – a necessary principle cannot entail a contingent claim without hidden contingent premises. It looks like the prosentential analysis of truth-talk does not quite capture the deflationary spirit of (T).

In spite of these worries, one can surely say that the prosentential theory offers an intriguing approach to truth. Moreover, it has been put to interesting use by Robert Brandom (1994, chap. 5). Brandom aims to present a worked out proper-use theory of meaning and content, a global alternative to standard semantic theorizing.

In this alternative, the traditionally central concepts of truth and reference are replaced by epistemic/inferential concepts and ultimately by *normative* concepts, pertaining to the normative commitments we undertake ourselves and ascribe to others in our social-communicative practices. In other words, Brandom is trying to work out in detail what the theory of meaning and content looks like from the point of view of the primacy of the normative over the descriptive (cf. section 6.5). Now, from this point of view, there can be no fundamental descriptive dimension to meaning/content/truth-talk – the descriptive must dissolve into norm-governed ascriptions and expressions of attitudes and commitments. So the approach requires some form of *performative* account of truth, i.e., an account of truth in terms of what we are *doing* – what speech-act we are performing – when we *assert* that something is true. The guiding idea is that truth-talk is characterized by redundancy of expressive force: To assert 'it is true that p' is to assert that p. But, as Geach (1965) has shown, simple performative accounts suffer from crippling defects. They cannot account for uses of 'true' in generalizations or even for uses within logically complex assertions, for one can assert the conditional 'if it is true that p, then q' without asserting the antecedent (cf. section 1.4). This is the point at which the prosentential theory is put to use. Brandom argues that the theory makes possible an extended performative (ultimately, normative) account of truth-talk, because it vindicates the view that truth-talk is not descriptive: 'it is true' merely inherits its content from its anaphoric antecedents; it does not contribute any content of its own. So the prosentential theory may solve an obstacle that stands in the way of developing the perspective of the primacy of the normative. At the same time, it may itself gain considerably from being embedded in this much larger philosophical enterprise. The marriage may provide this initially rather puzzling proposal with motivational support.[63]

8.3 Tarski

Contemporary deflationism is heavily influenced by some aspects of Tarski's (1933, 1944) work on truth in formalized languages. Tarski treated meaningful *sentences* as the primary truth bearers and thought of a truth definition as defining a truth predicate for (the sentences of) a given language. He usually began with a reference to correspondence formulations (which he criticized for not being sufficiently precise), singling out Aristotle's definition as the most promising and using it as a starting point from which to introduce the biconditional

(8) 'Snow is white' is a true sentence iff snow is white.

Although at times Tarski was inclined to say that (8) holds in virtue of a meaning equivalence (cf. 1969, 64), he usually expressed himself somewhat differently, saying that (8) can be regarded as a *partial definition* of the term 'true', as a definition of 'true' with respect to one particular sentence. He would then go on to say that a complete definition of truth (for sentences) will be, in a sense, the *infinite conjunction*, or the *logical product*, of such partial definitions: "Not much more in principle is to be demanded of a general definition of true sentence than that it should satisfy the usual conditions of methodological correctness and include all partial definitions of this type as special cases; that it should be, so to speak, their logical product" (Tarski 1933, 187). To make this idea more precise, he formulated a condition of adequacy for truth definitions. The following is a fairly close paraphrase of his formulation.

Convention T. A formally correct definition of the symbol 'Tr', formulated in the metalanguage, will be called an *adequate definition of truth* for an object-language, L, if it has the following logical consequences: (*a*) all sentences obtained from *Tarski's schema*

x is Tr if and only if p

by substituting for '*x*' a quotation-name of a sentence of L and for 'p' the translation of this sentence into the metalanguage; and (*b*) the sentence 'for all *x*, if *x* is Tr, then *x* is a sentence of L' (cf. Tarski 1933, 187f.).

This requires a comment. Tarski held that truth is not definable for natural languages, because natural languages are too unwieldy in structure and because they contain their own truth predicates, thus giving rise to the paradox of the liar (see section 2). He held that we can define truth predicates only for restricted and formalized (or formalizable) object-languages, and only in essentially richer metalanguages: a correctly defined truth predicate, 'Tr', will *apply* only to the sentences of the object-language *for* which it is defined but *belong* to the metalanguage *in* which it is defined; hence, it will not apply to sentences containing itself, thus avoiding the paradox. (Note that the object-language/metalanguage distinction is an artifice and will often not match the way one ordinarily distinguishes between different languages. An object-language simply is any set of sentences for which truth is to be defined; it can itself be contained within the metalanguage – indeed, this is the preferred case). Tarski tended to be rather strict about formalizability, but in more broadly philosophical contexts one often thinks of the relevant object-language, a bit vaguely, as comprising some sizable fragment of a natural language, say, of English – a fragment purged of semantic vocabulary to avoid liar-problems and (hopefully) more-or-less formalizable in first-order predicate logic.

It is debatable whether Tarski's work on truth is best understood as promoting the cause of deflationism. However, it is uncontested that the ideas sketched above, together with Tarski's procedure for defining truth predicates for formalized languages (sketched below), have been instrumental to the formation of contemporary deflationism. In what follows, I will concentrate on those aspects of Tarski's treatment of truth that seem to lend themselves to the cause of deflationism: the tantalizing characterization of (8) as "partial definition," the suggestion that a complete definition would be little more than the "logical product" or the "infinite conjunction" of such partial definitions, and finally Tarski's adequacy condition, Convention T – they are naturally taken as successively more precise ways of capturing the deflationary idea that there is nothing more to truth than what is exhibited by platitudinous (8) and its kin.[64]

Laying out a Tarskian truth definition for a reasonably rich language in a precise manner involves quite a bit of technical apparatus. Only a brief sketch of the procedure will be given here to convey some idea of the features that are especially relevant to deflationary thinking. Imagine first an object-language, L_0, consisting of a finite number of sentences of English. The following truth definition for L_0 employs quite literally the logical product of the relevant instances of Tarski's schema:

(9) *x* is Tr $=_{df}$ (*x* = 'Helen is tall' and Helen is tall) or (*x* = 'Peter is tall' and Peter is tall) or (*x* = 'Susan is a bat' and Susan is a bat) or (*x* = 'Snow is white' and snow is white) or... and so on for every sentence of the language.

By Convention T, this is an adequate definition of truth for L_0. It is just that L_0 is not a very exciting language. Interesting languages will comprise infinitely many sentences. This shows, according to Tarski, that a construction like (9) is not sufficiently general. For most languages, it would have to be infinite, which would prohibit us from deriving the instances of Tarski's schema: "We cannot arrive at a more general definition simply by forming the logical conjunction of all partial definitions. Nevertheless, what we eventually obtain is in some intuitive sense equivalent to the imaginary infinite conjunction" (Tarski 1969, 68f.). Tarski is hinting at the use of recursive methods to capture infinitely many sentences by finite means. Imagine an object-language, L_1, that results from adding 'not', 'and', and 'or' to L_0. Although L_1 is infinite, it is easy to see that its truth predicate can be finitely defined by adding recursive clauses to (9); e.g., the clause for 'and' will say that any complex sentence consisting of a sentence *x*, followed by 'and', followed by a sentence *y* is true iff *x* and *y* are both true, thereby defining truth for complex sentences by showing how they inherit their truth conditions from their simpler constituents via the logical structure of the language. However, this particular recursive definition works only because L_1 is based on a finite stock of sentences. Interesting languages are not based on a finite stock of sentences. But they are based on a finite stock of primitive predicates and primitive names; hence, we *can* define the notion of *satisfaction* for primitive predicates (open sentences) and the notion of *reference* for primitive names by *finite* lists exactly analogous to (9):

(10) A predicate *f* is satisfied by an object *o* iff (*f* = '*x* is tall' and *o* is tall) or (*f* = '*x* is a bat' and *o* is a bat) or (*f* = '*x* is white' and *o* is white) or... and so on for all primitive predicates of the language.

(11) A name *n* refers to an object *o* iff (*n* = 'Helen' and *o* = Helen) or (*n* = 'Peter' and *o* = Peter) or (*n* = 'Susan' and *o* = Susan) or (*n* = 'snow' and *o* = snow) or... and so on for all names of the language.[65]

For a simple sentence, consisting of a name and a primitive predicate, truth could be defined right away: 'Susan is a bat' is Tr iff there is an object *o* such that 'Susan' refers to *o* and *o* satisfies '*x* is a bat'. But most sentences are not that simple. Two modifications need to be mentioned. First, Tarski suggests to dissolve (11) into (10), which can be done by adding to (10) disjuncts like: (*f* = '= Helen' and *o* = Helen). This is not obligatory; it merely simplifies the construction by reducing reference to satisfaction so that there is only one basic notion (cf. Tarski 1933, 194). Second,

(10) has to be expanded by adding disjuncts covering n-place predicates. This introduces a complication. Since the *order* of objects is important when it comes to satisfying n-place predicates, the satisfaction relation turns out to be a relation between predicates and *sequences* (ordered n-tuples) of objects (e.g., 'x_1 is taller than x_2' is satisfied by a sequence $\langle o_1, o_2\rangle$ iff o_1 is taller than o_2). With these modifications in place, (10) will define *primitive satisfaction* (and reference) in a finite manner merely in terms of a long list. Now Tarski adds recursive clauses detailing how logically complex predicates inherit their satisfaction conditions from their logically simpler constituents via the logical structure of the language (and especially, detailing how quantified relational predicates, like 'for some x_2, x_1 is taller than x_2', inherit their satisfaction conditions). Putting all this together yields a recursive definition of the general notion of satisfaction that applies to arbitrarily complex predicates. Finally, Tarski treats sentences as limiting cases of predicates (as zero-place predicates) and takes care to formulate the definition of general satisfaction so that it covers this limiting case. This allows him to define truth for an infinite and relatively rich object-language on the basis of satisfaction simply like this:

(12) A sentence x is Tr $=_{\mathrm{df}}$ x is satisfied by all sequences of objects.[66]

Although no more than a sketch, this should illustrate why the Tarskian approach has been taken to underwrite deflationism about truth. By the lights of Convention T, (9) is a perfectly adequate definition of truth for L_0. Now, (9) is certainly deflationary; it is nothing more than the logical product of (8) and its kin. But the definition of truth for more complex languages, i.e., the definition that is based on the definition of satisfaction, although logically complex, does not seem to go significantly beyond a definition of the form (9). It merely adds logical (recursive) machinery, exploiting the structure of language to capture an infinite logical product by finite means. Although this constitutes a qualification of the original deflationary idea, it seems in line with the spirit of deflationism: there is no more genuine content to the notion of truth than what is given by (8) and its kin – logico-structural "content" is regarded as not "genuine" in the sense relevant to the issue whether truth is a substantive notion. Note especially the base clauses, (10) and (11), on which the whole construction rests. A deflationist will maintain that the manner in which they define primitive satisfaction (and reference) makes clear that there are no substantive notions lurking "behind" the notion of truth. The account of truth is deflationary *because* the account of satisfaction (and reference) is deflationary. And the account of satisfaction (and reference) is deflationary, because it rests on (10), which defines satisfaction in terms of the logical product of claims like 'o satisfies 'x is white' iff o is white' and its kin.[67]

A Tarskian truth definition helps clarify the semantic status of a truth predicate: 'Tr' is like other predicates in that it has an extension – a set containing all and only the sentences of the object language to which it applies. But it is unlike other predicates in that there is no intuitively unifying feature shared by the members of that set. All they have in common is that they are generated by the same rules from the same list and that they belong to the set in question. This gives some sense to the deflationary idea that the notion of truth does not pick out a genuine property. The Tarskian approach also addresses the generality problem. Since 'Tr' is a predicate, there is no in-principle problem with making sense of generalizations; e.g., Tarski (1933, §3) proves the semantic laws of non-contradiction and bivalence, and the law that every consequence of a true sentence is true. Of course, the generalizations are restricted to the sentences of the object-language for which 'Tr' is defined – unrestricted generalizations involving truth lead into the liar paradox.

Consider the T-schema below and note the slight difference between it and Tarski's schema as given above in Convention T.

T-schema: *x* is true if and only if p

A *T-sentence* is any sentence obtained from the T-schema by replacing '*x*' with a quotation-name of a declarative sentence and replacing 'p' with a translation of that sentence into the language of the schema (English) – the best known T-sentence is Tarski's paradigm (8): 'Snow is white' is true iff snow is white. Convention T is designed so that 'Tr' is a truth predicate for a given object-language, L, just in case its definition yields all the T-sentences for L. This insures that a predicate 'Tr' that complies with the convention is, as one says, *extensionally adequate* (relative to L). That is, it will apply to just those sentences of L to which 'true' applies: *x is true in L if and only if x is Tr in L.* Note the restriction of *our* predicate 'true' to L. The concept expressed by 'Tr' cannot coincide with our concept of truth; it can only coincide with a *restriction* of our concept to the language for which 'Tr' is defined. For one thing, 'Tr' is inevitably more restricted than our predicate, because our predicate leads into paradox – but this particular "limitation" should not speak against 'Tr'. In addition, 'Tr' is more restricted than our predicate because it is definable only for however much of our language is in principle formalizable – this restriction may be more disturbing.

As Etchemendy (1988) and Soames (1999, chap. 4) point out, given the restrictions put on the truth predicate, the claim that a Tarskian truth definition is an account of *truth* can only amount to the proposal that a Tarskian truth predicate could serve as a reconstructive analysis, or a theoretical explication, or possibly as a theoretical replacement, of our pretheoretical notion of truth (for the restricted domain over which 'Tr' can operate). The question is, then, whether a Tarskian truth predicate offers an adequate explication of (some plausible restriction of) our notion of truth. Two problems seem especially pertinent with an eye to deflationary views of truth.

One cannot understand 'true' without knowing that it is a (the) truth predicate. But one can understand a Tarskian truth predicate without knowing that it is a truth predicate; that is, one can understand the definition of a predicate 'Tr' without knowing that the definition does in fact comply with Convention T. Imagine a foreign language, L_3, that contains the sentence 'Ivica es mavagai'. Imagine you learn the truth definition for L_3. Maybe L_3 is finite and you find the following clause in the L_3-version of (9): (*x* = 'Ivica es mavagai' and Peter is a rabbit). Or maybe L_3 is infinite and you find the following clauses in the L_3-versions of (10) and (11) respectively: (*f* = '*x* es mavagai' and *o* is a rabbit); (*n* = 'Ivica' and *o* = Peter). Given your knowledge that Peter is a rabbit iff Peter is a rabbit, you can immediately infer (13) and/or (13a) and hence (13b), but not (13c):

(13) ('Ivica es mavagai' = 'Ivica es mavagai' and Peter is a rabbit) iff Peter is a rabbit;

(13a) (There is an object *o* such that 'Ivica' = 'Ivica' and *o* = Peter, and '*x* es mavagai' = '*x* es mavagai' and *o* is a rabbit) iff Peter is a rabbit;

(13b) 'Ivica es mavagai' is Tr in L_3 iff Peter is a rabbit;

(13c) 'Ivica es mavagai' is true in L_3 iff Peter is a rabbit.

Note that the only information relevant to 'Ivica es mavagai' conveyed by (13b) is just the information given in (13) and/or (13a). Only confusing 'Tr' with 'true' could lead to the impression that (13b) conveys what (13c) conveys. The latter gives you at least an inkling of what 'Ivica es mavagai' might mean, whereas (13)-(13b) tell you virtually nothing. One might share Davidson's (1973) hope that more information like (13c) will eventually enable one to come up with (something like) a theory of meaning for L_3. More "information" of the sort given by (13)-(13b) must surely be useless for any such enterprise. They, and their kin, shed no light whatever on the question whether (i) 'Tr' is a truth predicate for L_3 and 'Ivica es mavagai' means something in the ballpark of 'Peter is a rabbit' or whether (ii) 'Ivica es mavagai' means something entirely different and 'Tr' is not a truth predicate for L_3. The problem arises because a Tarskian truth definition does not define the notion of truth as restricted to L *for anyone* who does not understand L. The diagnosis suggests a possible remedy. A deflationist might propose to modify Convention T so that it requires the object-language to be contained in the metalanguage (the home language). Assuming the metalanguage to be English, the suggestion is that a definition of 'Tr' will be an adequate definition of (restricted) truth for speakers of English just in case it logically implies the *homophonic T-sentences* of (restricted) English, i.e., T-sentences like Tarski's paradigm (8), as opposed to T-sentences like (13c). This seems in line with the original intuition, for the idea that T-sentences can serve as partial definitions of truth arose only with respect to homophonic T-sentences to begin with. Of course, the proposal leaves the notion of truth as it occurs in diaphonic T-sentences in limbo – at least for the moment.[68]

The second point is simply that T-sentences express contingent truths. Consider again Tarski's paradigm (8): 'Snow is white' is a true sentence iff snow is white. Whether the sentence 'Snow is white' is true depends in part on its meaning, whereas the color of snow does not depend at all on the meaning of 'Snow is white'. If the sentence 'Snow is white' meant that snow is green then it would be false, but the color of snow would remain unchanged. Moreover, 'Snow is white' could be true even if snow were not white; for example, if it meant that blood is red. The contingency of (8) might escape notice at first, because we tend to *presuppose* that 'Snow is white' has its ordinary meaning. We are tempted to think: "*Assuming* 'Snow is white' means what it ordinarily means, it is necessary that 'Snow is white' is true iff snow is white." But making a contingent assumption cannot *make* a contingent claim come out necessary. The thought seems to rest on confusing (14), which is false, with (15), which is true:

(14) If sentence x means only that p, then $\Box(x$ is true iff p);
(15) $\Box$(If sentence x means only that p, then x is true iff p).

This observation gives rise to an objection to deflationary views about *sentence truth*. T-sentences are mere material biconditionals, and it is hard to see in what sense a material biconditional could serve as a partial *definition* of truth – even partial definitions should offer more than that. The point affects the Tarskian definition (for restricted English) of the alleged truth predicate 'Tr'. If the Tarskian "definition" is merely contingent, then it is not really a definition at all. If, on the other hand, the definition is necessary, then it cannot have the contingent T-sentences as logical consequences. Since, however, the corresponding Tr-sentences *are* logical consequences of the definition of 'Tr', they must then be necessary truths. It follows, so the objection goes, that the Tarskian truth predicate, 'Tr', does not offer an adequate explication of truth – Has the Tarski-inspired deflationist confused (14) with (15)?[69]

If a Tarskian predicate satisfies Convention T, then it is correctly defined: it cannot, as it were, fail to do what it is defined to do. Consequently, the objections mentioned above are really objections against Convention T to the effect that it is not an adequate adequacy condition for truth definitions. This is important to the debate over deflationism. For it is really Convention T that has all the deflationary import: if having the T-sentences as logical consequences is *sufficient* for being an adequate definition of truth, then there is no more genuine content to the notion of truth than what is contained in the T-sentences. According to the first objection, the convention is too weak, because a definition of 'Tr' may satisfy it even though 'Tr' does not express the notion of truth. According to the second objection, the convention is not only too weak it is quite besides the point as well. A definition of 'Tr' that satisfies the convention will thereby have the *Tr-sentences* as logical consequences. But that just proves that Tr is not truth, because a definition of truth for sentences *should not* have the *T-sentences* as logical consequences, since they are contingent. Note that no correspondence definition for sentence-truth will satisfy Convention T. And according to correspondence theorists, this is how it ought to be. A correspondence definition for sentence-truth is a necessary truth. It must be combined with a substantive theory of content (semantics) before it will yield T-sentences. Since sentences have their contents contingently, such a theory will issue contingent claims.[70]

8.4 Disquotationalism

Truth, according to Quine, is disquotation. This, he says, "is explicit in Tarski's paradigm: 'Snow is white' is true if and only if snow is white. Quotation marks all the difference between talking about words and talking about snow. The quotation is a name of a sentence that contains a name, namely 'snow', of snow. By calling the sentence true, we call snow white. The truth predicate is a device for disquotation" (Quine 1970, 12). So far, *disquotationalism* looks just like a simple redundancy theory of truth for sentences. The claim seems to be that all there is to the notion of truth is what is given by the substitution instances of the *disquotation schema*:

(DS) 'p' is a true sentence iff p

and the substitution instances are just the homophonic T-sentences – Tarski's partial truth-definitions. But the redundancy theory was seen to suffer from a serious illness. It could not give an intelligible account of generalizations involving truth without presupposing the notion of truth: truth is not redundant. And Quine concurs. Indeed, he emphasizes that the truth predicate is very much needed precisely for generalizations, i.e., precisely when we want to assent to sentences that we cannot list individually and are forced to resort to indirection. We cannot affirm the law of excluded middle by affirming each instance of 'p or not-p'. Instead, we engage in "semantic ascent" and talk about sentences: Every *sentence* of the form 'p or not-p' is *true*. The function of the truth predicate is to cancel the effect of semantic ascent, thus enabling us to affirm infinitely many disjunctions in one breath: "We may affirm the single sentence just by uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. We need it to restore the effect of objective reference when for the sake of some generalization we have resorted to semantic ascent" (Quine 1970, 12). The point of course extends beyond logical examples to cases like 'Everything he says is false', and 'My goal is to affirm only what is true', and to other blind uses, like 'Something Tarski said is false' and 'His favorite sentence is true'. In all such cases, 'true' ('false') is needed as a device for effecting assent to, or dissent from, possibly infinite object-language conjunctions and disjunctions.[71]

Note the way in which this approach attempts to resolve the dilemma posed for deflationism by the generality problem described in section 8.1. The point is that the generality problem does indeed refute the redundancy theory (and any simple deflationary theory) but not in a way that underwrites the thesis that truth is a substantive notion in need of a substantive theory. On the contrary, that truth is essential for making generalizations just proves Tarski's point that the truth predicate is our means of forming the infinite logical product of the T-sentences. Truth is a logical, or quasi-logical, notion. It does not belong with 'water', 'knowledge', or 'justice'; it belongs with 'and', 'or', 'every', and 'some'. The idea that the truth predicate functions as a kind of disquotational quantifier could be exhibited by transforming the disquotation schema (DS) into the following formulation:

(DT) x is a true sentence iff, for some p, x = 'p' and p.

This makes use of the schematic quantifier-phrase 'for some p' to "encode" the infinite-logical-product account of truth considered by Tarski; see (9) of section 8.3. The right-hand side of (DT) is supposed to indicates the infinite disjunction that 'x is true' allows us to abbreviate: (x = 'Helen is tall' and Helen is tall) or (x = 'Peter is tall' and Peter is tall) or (x = 'Susan is a bat' and Susan is a bat) or (x = 'Snow is white' and snow is white) or... Admittedly, we (probably) do not have such a 'for some p'-quantifier in English – so (DT) should (probably) not be regarded as a definition. Nevertheless, (DT) shows us that the truth predicate does the work that would be done by such a quantifier, if we had one – or better, the disquotationalist should point out that we *do* have such a quantifier, namely the truth predicate '*x* is true'.[72]

Disquotationalism is obviously much indebted to Tarski's work on truth. But there are also significant differences. The disquotational approach to truth is noticeably less precise than a Tarskian truth definition. This is because disquotationalism attempts to offer an account of the *function* of *our* truth predicate. The disquotationalist may well hold that a precisely defined Tarskian truth predicate serves as a partial technical reconstruction of our truth predicate. But a Tarskian truth predicate is very restricted; its greatest asset is also a great liability. Its definition nicely captures infinitely many T-sentences by finite means. But it depends on antecedent structural analyses of the sentences of the relevant object-language. If the object-language is a natural language, we find many constructions for which we do not (and may never) have such analyses. Consequently, we do not have, and maybe could not have, a nicely finite Tarskian definition of a predicate with a reach close to our truth predicate (even if we think of our predicate as restricted to some paradox-free fragment of our language). A Tarskian truth predicate is simply too restricted to serve as a realistic candidate for an analysis or explication of our notion of truth. The disquotational account, although it does not really capture the infinitely many T-sentences by *finite* means, seems more realistic on this score. (DT) could be said to illustrate the function of our notion of truth. Unlike a Tarskian truth definition, it only presupposes that the language it applies to *has* sentences. It is quite insensitive to their internal structure; consequently, it does not depend on the availability of an analysis of sentential structure.[73]

Disquotationalism inherits some problems from its Tarskian ancestor. Tarski's T-sentences hold generally only for what Quine calls "eternal sentences"; i.e., for univocal sentences without demonstratives or indexicals or other context sensitive elements. Context sensitive sentences can generate *false* instances of (DS): the sententence ''He is hungry' is true iff he is hungry' will be false whenever the 'he's are used to refer to different persons (consider also: ''This sentence contains five words' is true iff this sentence contains five words'). The generally accepted solution is to treat the T-sentences as applying to individual sentence-*tokens* as understood by the speaker in question. Such a restriction will be necessary in any case. According to the disquotationalist, the notion of truth encodes the homophonic T-sentences – the infinite expansion of (DT) consists of the (slightly rearranged) instances of the disquotation schema (DS). But what are the homophonic T-sentences? At first one wants to respond that the homophonic T-sentences for English are the instances of (DS); and the homophonic T-sentences for German are the instances of the German version of (DS), and so on. But this would be at once too broad and too narrow. Too broad because a national language like English is an idealization. No one actually understands the whole vocabulary of English. Too narrow, because most people understand at least some "mixed" T-sentences, like: ''Schnee ist weiß' is true if and only if Schnee ist weiß'. If disquotational truth were thought of as encoding the homophonic T-sentences of a national language, it would not capture anyone's notion of truth. So a disquotationalist must restrict the account of truth to all and only the sentences that are understood by an individual speaker, i.e., to the sentences of a speakers idiolect. Field calls this *pure disquotational truth* and characterizes it with an equivalence thesis: "A person can meaningfully apply "true" in the pure disquotational sense only to utterances that he has some understanding of; and for such an utterance *u*, the claim that *u* is true (true-as-he-understands-it) is cognitively equivalent (for the person) to *u* itself (as he understands it)."[74] This has the immediate consequence that, unless there are linguistically identical twins, no two speakers will have the same notion of truth. After all, pure disquotational truth is nothing but a device for abbreviating all and only the sentences a given speaker understands, and no two speakers understand exactly the same sentences. Moreover, pure disquotational truth is inapplicable to any other language but one's own idiolect. Since, however, *our* notion of truth *is* applicable to languages we do not understand, the disquotationalist has to introduce some notion of *extended disquotational truth* to cover the range of our notion of truth (cf. Field 1994, sec. 8). The extended notion will be defined in terms of the pure notion plus the concept of synonymy (or correct translation): A foreign utterance u^* is $\text{true}_e =_{df} u^*$ is synonymous with some home utterance *u*, and *u* is true_{pd}. Of course, no two speakers can have the same extended notion of truth either, since it is defined in terms of the pure notion. And applications of truth to foreign sentences that have no translation into a speaker's idiolect (or whose translation into his idiolect a speaker would not understand) are still not covered by this account. In addition, extended disquotational truth raises the worry that the concept of synonymy (or correct translation) may not be intelligible without recourse to the notion of sameness of truth conditions. Finally, since disquotationalism offers a theory of truth (however deflationary) and since disquotational truth is nothing but an encoding of a speaker's T-sentences, disquotationalism is committed to the view that the T-sentences of a speaker's idiolect as understood by him are necessary: □('Snow is green' is true_{pd} iff snow is green). So, if I had used 'Snow is green' as 'Snow is white' is actually used by me now, then snow would have been green (cf. section 8.3). Opinion on all these consequences of disquotationalism is divided. Proponents take them in stride as features that are constitutive of disquotational truth. Critics tend to regard them as *reductio* of the thesis that disquotational truth has anything whatever to do with truth.[75]

One feature that has given rise to objections to disquotationalism is actually a family trait that it shares with all deflationary approaches to truth. The T-sentences do not distinguish between falsehood and the absence of truth – neither do the instances of (T_P) nor the instances of (T). Once one conceives of truth in terms of a schema like ''p' is true iff p', falsehood must be conceived in terms of ''p' is false iff not-p', which collapses falsehood into untruth. The result will be that deflationary views enforce *bivalence*, i.e., they cannot tolerate truth-value gaps, since the claim that 'p' is not true and not false would come out as the contradiction that not-p and not not-p. But some ethicists have maintained that moral claims are neither true nor false; and the same is sometimes said about claims that presuppose the existence of things that do not exist – e.g., the claim that the present King of France is lazy. Arguments that such claims should be regarded as "gappy" have been turned into arguments against the deflationary approach. To put the issue the other way round, it seems a theory of truth should not simply dictate that, say, emotivism is an inconsistent meta-ethical position. The liveliest discussion in this area concerns what may be the strongest candidate for generating truth-value gaps: *vague* sentences applied to borderline cases, say 'Jones is bald'. To give an example of just one aspect of this discussion, Field (1986, 69) suggests that deflationists should handle vagueness by introducing a 'determinately' operator. So instead of being forced to treat the view that 'Jones is bald' is neither true nor false as a contradiction ('Jones is bald' is not true and not not true), the deflationist can say that 'Jones is bald' is not determinately true and not determinately not true – no contradiction. Of course, by the very idea of deflationism, the latter must be equivalent to the claim that Jones is not determinately bald and not determinately not bald, which brings out a questionable feature of this attempt to handle vagueness: deflationism will be committed to blaming all vagueness, without exception, on the world rather than on language. Correspondence theories can be more judicious; they can blame some vagueness on indeterminacies in the world and some on indeterminacies in the semantic relations between language and the world.[76]

8.5 Minimalism

According to Paul Horwich, traditional theories of truth grow out of the misconception "that truth *has* some hidden structure awaiting our discovery"; but "unlike most other predicates 'is true' is not used to attribute to certain entities (i.e. statements, beliefs, etc.) an ordinary sort of property – a characteristic whose underlying nature will account for its relations to other ingredients of reality. Therefore, unlike most other predicates, 'is true' should not be expected to participate in some deep theory of that to which it refers – a theory that goes beyond a specification of what that word means" (Horwich 1990, 2). The deflationary theory of truth proposed by Horwich, the *minimal theory*, *MT*, consists of all propositions of the form:

(T_P) The proposition that p is true if and only if p.

MT is a very large theory. It has infinitely many axioms: any proposition expressed by the result of replacing the letter 'p' in schema (T_P) by a declarative sentence of English, or of any possible extension of English, is an axiom of MT (see Horwich 1990, 16-22).

Like disquotationalism, minimalism denies the redundancy theory and holds that the truth predicate serves an important logical need. We need it to endorse (or reject) claims when we are not in a position to express them directly, either because we do not know what exactly the proposition is that we want to endorse, or because we want to generalize and endorse indefinitely or infinitely many propositions. Moreover, like disquotationalism, minimalism holds that the truth predicate exists *solely* to serve this logical need and cannot play any explanatory role in, for example, logic, or epistemology, or in theories of mind and language. But there are also some important differences between minimalism and disquotationalism. First, minimalism is formulated primarily as theory of truth for propositions. Second, it is more unabashedly infinite than disquotationalism. Third, Horwich emphasizes that he regards truth as a property of some sort – not, of course, as a substantive property, but as a logical, or quasi-logical property.[77]

Horwich (1990, 34-38) claims for MT not only that it accounts for the property of being true (of being a true proposition), he also claims that it accounts for our *concept* of truth. MT is our definition of the truth predicate. It is an implicit definition: our understanding of the truth predicate, i.e., our grasp of the concept of truth, consists in the disposition to accept, without evidence, any substitution instance of (T_P), that is, any sentence that results from substituting a declarative sentence of English (including any possible extension of English) for 'p' in schema (T_P). This account of our concept of truth faces some difficulties. Note that MT is a vast theory. Due to their complexity, large numbers of propositions constituting MT, as well as the sentences expressing them, will be beyond our grasp. And for each person, there will be many relatively simple members of MT that she does not understand, namely, members that are composed of propositions belonging to specialized sciences of which the person has no more than the most fragmentary understanding. But a person who does not understand Gödel's theorem can understand and assert the proposition *that Gödel's theorem is true* and thereby commit herself to the truth of Gödel's theorem. Minimalism cannot account for her grasp of the notion of truth in that proposition, because Gödel's theorem is not among the fragment of MT that constitutes the person's grasp of truth. Moreover, each person's truth concept will likely be different from everyone else's because it is unlikely that any two persons understand exactly the same propositions. It seems, according to minimalism, we do not really *grasp* truth – we are just holding on to the tip of its tail-end.

But maybe Horwich's account is not meant to require that our disposition to accept substitution instances of (T_P) must involve our understanding of the propositions expressed by them. Maybe all that is required is that we assent to the sentences that results from substituting 'p' in (T_P), never mind whether we actually understand 'p'. Consider:

(16) The proposition that context-sensitive selection restrictions are essential to generative grammar is true if and only if context-sensitive selection restrictions are essential to generative grammar.

A person who has never heard Chomskyan linguistics and has no clue what the embedded sentence might mean may indeed have a disposition to assent to (16). But that disposition will derive from her recognition that (16) is a substitution instance of (T_P), combined with her recognition that every substitution instance of (T_P) is *true*. If minimalism merely requires assent to the sentences that are instances of (T_P), its account of our grasp of truth is in serious danger of presupposing our grasp of truth.

Horwich claims that MT, in combination with relevant background theories, explains all the facts about truth. For example, MT explains why the proposition that snow is white follows from the propositions that Tarski's favorite proposition is true and that Tarski's favorite proposition is the proposition that snow is white. That the proposition that snow is white is true follows from logic alone (logic being the default background theory). The proposition that snow is white can then be deduced from the relevant member of MT (see Horwich 1990, sec. 2). But Horwich also needs to show that *generalizations* involving truth can be explained in this manner. He attempts a proof of the law: If one proposition (materially) implies another, and the first one is true, then so is the second. As Tarski pointed out (in advance, as it were), this attempt must fail. The minimal theory (with logic again functioning as the default background theory) can prove every instance of the schema:

(17) (The proposition that p is true & the proposition that p implies the proposition that q) → the proposition that q is true.

But the desired generalization remains beyond reach. A universal generalization is not deducible from the totality of its instances without the additional premise that they are all the instances. The same problem arises with respect to other generalizations, e.g., the law of non-contradiction, and the law that a conjunction is true iff both its conjuncts are true. Nor does the problem seem to be restricted to logical generalizations: MT will deliver the instances of a generalization like 'A proposition is known only if it is true', but the generalization itself will be out of reach. Gupta (1993a, sec. 2) argues that the fault lies with Quine's original attempt to overcome the generality problem: affirming all the instances of a universal generalization is not the same as affirming the generalization itself.[78]

Finally, Gupta (1993, 365) points out that MT is, in an important respect, the *maximal* theory of truth. Compare MT to a correspondence theory of truth for propositions, e.g.: A proposition is true iff it corresponds to a state of affairs that obtains. To be sure, MT is ontologically more parsimonious than the correspondence theory, since it is not committed to an ontology of states of affairs. But MT contains an axiom for each and every proposition that is expressible in possible extensions of English. Consequently, it employs each and every property expressible in possible extensions of English; no property (expressible in possible extensions of English) is exempted. Compared to this, the correspondence theory, albeit ontologically more demanding, appears relatively deflationary. It employs merely the relation of correspondence, the property of being a state of affairs, and the property of obtaining. And of course, MT employs these properties too. After all, somewhere in MT there will be a fragment that says, for example: ...the proposition that there are states of affairs is true iff there are states of affairs; the proposition that some objects correspond to other objects is true iff some objects correspond to other object; the proposition that some objects obtain or fail to obtain is true iff some objects obtain or fail to obtain...

These difficulties for minimalism also arise, in slightly modified form, for disquotationalism. They indicate that these deflationary views have never really overcome the generality problem. A theory that aims to account for truth in terms of the instances of a schema like (T_P) or (DS) is not strong enough to yield generalizations about truth because the theory itself does not consist in a generalization. For the same reason, such a theory seems much too demanding as an account of our concept of truth: our understanding of the concept of truth does not require the massive conceptual resources required for understanding the substitution instances of these schemata. With respect to the latter problem, the deflationist may want to turn away from the infinite array of *instances* of the schema and reconsider the role of the schema itself:

(T_P) The proposition that p is true if and only if p.

The schema is nicely finite; and it is simple – inviting deflationary intuitions about truth. The idea would be that it is the schema itself, rather than its instances, that explains our concept of truth, or the meaning of 'true' (as applied to propositions). This is an interesting approach, but it has to overcome a prima facie obstacle. A schema is just a pattern, a mere frame for a sentence; it does not express a proposition and does not *say* anything (cf. note 62). The "schematic theory of truth" needs an account of how our concept of truth could be explained by a mere schema; it needs an account of what our *understanding* of the schema could consists in. There is one account, but it is not available to an advocate of the "schematic theory": our understanding of (T_P) consists in our recognition that its substitution instances are true.[79]

Marian David
University of Notre Dame

NOTES

[1] Remarks on notation. The schema 'Her belief that p is true' abbreviates a *sentence* and must be parsed accordingly as '[Her belief that p] is true'; similarly for '[The belief that p] is true', and '[The proposition that p] is true'. The letters 'p' and 'q' will be used as schematic sentence-letters—mere dummies always replacable by complete declarative sentences. They will never be used as genuine variables, like '*x*' and '*y*', which range over objects and can only be instantiated by names or other singular terms referring to objects (including singular terms referring to propositions and sentences). At times 'iff' will be used as short for 'if and only if'.

[2] For the Stoics see Long and Sedley 1987, sec. 33. Medieval authors referred to the Stoic-type proposition as the *dictum* or the *complexe significabile*—the medieval *propositio* always indicates a verbal or mental sentence or act; see Nuchelmans 1973.

[3] The recognition that logical complexity creates a strong need for unbelieved, unjudged, unstated, and unasserted truth bearers is mainly due to Frege; see, e.g., Frege 1891, 21f., and 1918. Compare also: Meinong 1910, chaps. 2 and 6; Geach 1965; and Fodor 1978.

[4] See Vision 1988, chap. 2, and Kirkham 1992, chap. 1, for further discussion of the various projects that have been pursued under the label "theory of truth." Note that a Davidsonian "truth theory" is not a theory of truth at all; it is a theory of *meaning* that attempts to capture meaning in terms of truth; cf. Davidson 1973 and 1977.

[5] For some treatments of the liar paradox see: Tarski 1933; Quine 1970, chap. 3; Kripke 1975; Martin 1984; McGee 1991; Gupta and Belnap 1993; Soames 1999, chaps. 5 & 6.

[6] "Whereas the true statement is in no way the cause of the actual thing's existence, the actual thing does seem in some way the cause of the statement's being true; it is because the actual thing exists or does not that the statement is called true or false" (Aristotle, *Categories* 14^b15). "None of the things underlying an affirmation or negation is a statement. These are, however, said to be opposed to one another as affirmation and negation are...For in the way an affirmation is opposed to a negation, for example 'he is sitting'—'he is not sitting', so are

opposed also the actual things underlying each, his sitting—his not sitting" (Aristotle, *Categories* 12^{b}5).

[7] The Stoic passage is cited from Kneale and Kneale 1962, 152. For the other passages see: John Buridan, *Sophisms*, II.B.14; Thomas Aquinas, *De veritate*, Q1, A1&3; Descartes 1639, ATII 597; Spinoza 1677, axiom vi; Locke 1700, IV.v.i; Leibniz 1765, IV.v.ii; and Kant 1787, B82. It should be noted that Aquinas's attribution of the *adaequatio*-definition to Isaac Israeli appears to be mistaken; cf. Woleński 1994.

[8] Moore was writing in the years 1910-11. Russell had already been talking about facts and correspondence in 1904 under the influence of Meinong, and in section 3 of 1906-07. The latter is a response to H. H. Joachim who defended a coherence theory and attacked the "correspondence-notion of truth" which, at one point, he described thus: "A judgement e.g. is true, if the thoughts whose union is the judgement 'correspond' to the facts whose union is the 'real' situation which is to be expressed" (Joachim 1906, 19). For more on the history of correspondence to facts, see Woleński 1994a.

[9] Although Russell (1918, 223) is committed to taking beliefs as primary truth bearers, it is deeply unclear how recursive accounts could be made to work for beliefs taken as primary: since complex beliefs do not in general *have* simpler *beliefs* as their constituents, their truth values cannot in general be accounted for recursively in terms of the truth values of simpler constituent beliefs; see section 1.4. This suggests that advocates of recursive approaches should turn to the language-of-thought analysis of belief; see section 1.2. Note that, in any case, recursive accounts of truth work best (exclusively?) for sentences; after all, they rely heavily on what appears to be the *logico-syntactic* structure of truth bearers.

[10] Fact-terminology varies greatly. Wittgenstein calls only conjunctive facts 'facts' and uses 'state of affairs' (*Sachverhalt*) to refer to atomic facts. Armstrong calls conjunctive *and* atomic facts 'states of affairs'. I reserve the term 'state of affairs' for a different use.

[11] Intuitively, facts are only *relatively* abstract, because many facts depend for their existence on concrete entities: the fact that Clinton is president could not have existed if Clinton had not existed (in this respect facts are like impure sets). Indeed, on an Aristotelian view on which universals exist only when instantiated by at least one particular, all facts would depend for their existence on concrete objects. Full-fledged (Platonic) abstract entities are not existence-dependent on any concrete particulars.

[12] Moments in general were known as *modes* in early modern philosophy. Individual-accident moments are now often referred to as *tropes*; see Armstrong 1997, chap. 2.

[13] Cf. Olson (1987) who argues that facts are needed to account for relations; see also Armstrong 1997, chap. 8, and Wittgenstein 1921. As Mulligan et al. point out themselves, their factless atomism faces a number of additional difficulties; see 1984, 300-02, 314-18. Facts are a tricky topic, involving substantive as well as terminological issues; see: Ramsey 1927, 36f.; Strawson 1950; Austin 1961; Vendler 1967, chap. 5; Fine 1982; Bennett 1988, chaps. 1 and 2; Olson 1987; Armstrong 1997.

[14] A strictly Tarskian truth-definition treats satisfaction in a rather deflationary manner; see section 8.3. Of course, this does not mean that correspondence theorists cannot make use of the technical machinery made available by Tarski; cf. Field 1972. However, the correspondence theorist must address a potential obstacle. If a clause like ''p or q' is true iff 'p' is true or 'q' is true' is to be used in a recursive account of *our* notion of *truth*, as opposed to some other notion, it has to be presupposed that 'or' expresses *disjunction*—one cannot define 'or' and 'true' at the same time. To avoid circularity, a recursive correspondence theory (be it atomic or subatomic) must hold that the logical connectives can be understood without reference to correspondence truth.

[15] For advocates of the subatomic correspondence-as-reference-*cum*-satisfaction theory, see: Field 1972, and 1978 (but he discards this approach in his 1994); Devitt 1982, 1984; Schmitt 1995; cf. also Kirkham 1992, chaps. 5 and 6. These authors envision some form of causal theory of "correspondence", i.e., of reference and satisfaction. Also, they prefer a more

nominalist brand of base-clause for satisfaction than the one given in the text, i.e., they aim for an account of satisfaction that does not invoke the idea that predicates express properties or relations. The issue whether predicate-satisfaction can be handled without such ontological commitments is the linguistic version of the problem of universals. It turns out that *n*-place predicates require talk of satisfaction by ordered *sequences* of objects: ‘Rx_1, x_2, x_3’ is satisfied by the sequence ⟨o_1, o_2, o_3,⟩. This brings in relatively abstract set-theoretic objects independently of the issue whether an account of satisfaction requires recourse to properties and relations. Davidson, who argues that satisfaction by *sequences* is all that remains of the traditional idea of correspondence to *facts* (cf. 1969), seems to regard reference and satisfaction as “theoretical constructs” not in need of causal, or any, explanation; see his 1977.

[16] Cf. Plato, *Theaetetus* 188^{c}-189. The problem of falsehood is the theme of pp. 54-76 of Russell’s (1904) discussion of Meinong. Joachim (1906, chap. 4) raises it as an essential part of his attack on the correspondence theory. The problem receives the most sustained discussion by far in Moore 1953, 249-287; see also: Russell 1912, chap. 12; Prior 1967; and Williams 1976, chap. 5.

[17] This issue arises in much the same form for the subatomic version of the correspondence theory on which (1) is true iff the belief relation is satisfied by (the ordered pair containing) the object referred to by ‘Susan’ and the object referred to by ‘that p’. The latter has to be an object that can function as a false maker, but the subatomists’ framework does not seem to provide for such things.

[18] Davidson (e.g., 1973) wants to give a truth-conditional account of meaning roughly along the following lines: a theory explains (4) if it entails the right instances of ‘*s* is *true* iff p’ and satisfies a number of substantive logical and empirical constraints. This approach to (4) will be useless in the present context. As Davidson realized, an account of meaning in terms of truth must presuppose the notion of truth. Yet, as we have seen, the problem of how to handle (4) comes up *within* the theory of truth.

[19] Quine tends to agree with the propositionalist that (1)-(4) can in the end only be accounted for by invoking propositions. But since he holds that propositions are creatures of darkness and are not to be admitted into serious science, he infers that there are no (serious) truths about meaning or belief: (1)-(4) don’t need to be accounted for; they are merely a “dramatic idiom” that has no room in science; cf. Quine 1960, §45. Less radical attempts to bypass the argument from falsehood to propositions often lead into Meinongianism about facts—an instance of Meinong’s (1910, 79) general thesis that there are objects that do not exist. Moore, desperately trying to find a way out of Plato’s problem, was strangely drawn to Meinongianism about facts: “I think, therefore, that the most essential point to establish about truth is merely that every belief *does refer*...to *one* fact and one fact only and that to say of a belief that it is true is merely to say that *the* fact to which it refers *is*; while to say of it that it is false is merely to say that the fact to which it refers, *is not*—that there is no such fact” (Moore 1953, 269). The idea seems to be to provide facts, albeit non-existent ones, for false beliefs to correspond to so that the need for propositions does not arise in the first place.

[20] For accounts of truth along the lines of (CS), see: Austin 1950; Chisholm 1976 and 1977; Taylor 1976; Bealer 1982; Barwise and Perry 1983; Forbes 1986; and David 1994. Barry Smith (1994, chaps. 4 and 6) shows that for the early 20th-century history of states of affairs (*Sachverhalte*) one has to look at the students of Brentano, especially Carl Stumpf, Kasimir Twardowski, Anton Marty, Husserl, and Meinong. An early formulation of (CS) can be found in the German original of Wittgenstein’s *Tractatus* (1921, 4.25): “Ist der Elementarsatz wahr, so besteht der Sachverhalt; ist der Elementarsatz falsch, so besteht der Sachverhalt nicht.” However, Wittgenstein’s states of affairs (*Sachverhalte*) are atomic *facts*; and facts cannot fail to obtain (*bestehen*) and still exist. It appears that Wittgenstein was a Meinongian about atomic facts (states of affairs).

[21] The term 'proposition' is often used to refer to content, whatever content turns out to be. On this terminology, the view that contents should be identified with states of affairs is expressed as the thesis that *propositions* are constituted by worldy objects and properties. Authors who focus on the role of proper names and indexicals tend to put the view as the thesis that contents are "singular" or "Russellian" propositions, rather than "Fregean" propositions. The view has also been expressed as the thesis that contents are "wide" rather than "narrow". Besides the three seminal works mentioned in the text, see McGinn 1989, and the essays collected in Woodfield 1982, and Salmon and Soames 1988.

[22] For advocates of an identity theory, see: Moore 1899 and 1901-02; Meinong 1910, chap. 3; Ducasse 1940; Chisholm 1976, chap. 4, and 1977, chap. 5. Russell's discussion of Meinong ends with a tentative and short-lived endorsement of the theory; see Russell 1904, 74-76. Frege's wording of (PI) actually suggests an account of facts rather than truth: "What is a fact? A fact is a thought that is true" (1918, 74). The theory has recently received some renewed attention. Candlish (1989), who introduces the label "identity theory", and Baldwin (1991) discuss Bradley's version. Baldwin also offers a nice quote from Hegel (1830, §213): "Truth in the deeper sense consists in the identity between objectivity and the notion." Extended discussion of a version of the identity theory can be found in Hornsby 1997. For indefinability claims concerning truth, see: Moore 1899, 5; Russell 1904, 75f.; Frege 1918, 60; cf. also Cartwright 1987.

[23] Take '$\Box$' as short for 'it is necessary that'. Note that both parties agree on: $\Box$(if x = S's belief that p, then x has the content that p). What the (token) physicalist denies and his opponent affirms is rather: If x = S's belief that p, then $\Box$(x has the content that p).

[24] Only for the *metaphysics* of knowledge. The identity theory does not imply that we get knowledge somehow for free. Say, S believes that p, and that p is a fact. It does not even begin to follow that S knows that p—to think otherwise would be to confuse knowledge with true belief.

[25] Moore, in his 1953, p. 308, raises the following objection. Assume the proposition that p is (contingently) true. According to the identity theory, the proposition exists whether it be true or false. But the *fact* that p would not have existed, if the proposition had been false. Hence, the fact that p cannot be identical with the proposition that p. As Cartwright (1987, 76-8) points out, the argument is question-begging. It assumes that 'the fact that p' is a "rigid designator," designating the proposition that p in every world in which it exists. But the identity theorist will hold that 'the fact that p' is non-rigid, designating the proposition that p only in those worlds in which it is *true*. (Compare: If John had not married Mary, then Mary's husband would not have existed; it does not follow that John is not identical with Mary's husband.) Another objection one might raise is this: the identity theory commits one to the view that facts are true. The identity theorist will have to take this in stride. She will have to say that "facts are true" is literally true; it merely sounds odd because it amounts to the redundant claim that true propositions are true.

[26] James had a great many seemingly different things to say about truth. Besides numerous passages like the ones cited in the text, one can find endorsements of a verifiability theory of truth, a coherence theory, an ideal-consensus theory, and especially of a correspondence theory; cf. 1907, 96-103; 1909, 90-98, 104-107, 112, 117. Whether they can all be fit into a consistent whole is a matter of some debate among James scholars. The rough outlines of the picture are fairly clear however. James held that different features are critical for the truth of different types of beliefs: for observational beliefs the critical feature is the "copying" of sensations; for theoretical beliefs it is verifiability; for metaphysical beliefs (like the belief in God) it is emotional satisfaction—coherence with other beliefs and the potential for leading to successful action are relevant throughout. The only common trait, James thought, that underlies and unites all these different ways of being true is utility: "Our account of truth is an account of truths in the plural...having only this quality in common, that they

pay" (1907, 104). He maintained that his account *explains* what "truth is agreement with reality" really amounts to. See Kirkham 1992, chap. 3.3, for a more detailed interpretation.

[27] Since pragmatists tend to be opposed to serious propositions, I will talk of beliefs and statements as truth bearers. It should be remembered though that they are problematic candidates for the role of primary truth bearers; see section 1.4.

[28] At one point James seems inclined to deny (heroically) that his account is committed to substitutability in (T); see 1909, 150. One may wonder whether it is quite fair to wield (T) in the manner described. For, as was pointed out in section 1.5, all theories that construe 'true' as a predicate of truth bearers have difficulties accommodating (T) due to the essential use (T) makes of the "bearerless" operator form 'it is true that'. Can one rely on (T) for an objection specifically against the pragmatic theory? But note that the Russell-Moore argument does not seem to hinge on any issue involving truth bearers. The argument will go through, even if the pragmatic definiens is interpreted as existence-neutral, say, along the lines of 'if it were believed that p, then that would be useful'. Alternatively, the objection could make use of a more awkward principle in place of (T), e.g.: S's belief that p is true iff S believes that p, and p. This would complicate the argument but would not introduce any substantial changes..

[29] James's stance towards anti-realism is not easy to determine. At times his writings suggest that he sees reality as largely constructed by us; see James 1907, lecture 7. His response to Russell will not inspire a realist with much confidence either; see James 1909, 146-150. But when he explicitly addresses the objection that errors are often satisfactory, he responds that utility is not sufficient for truth and seems to say that "leading to reality" is a necessary condition for the truth of a belief in *addition* to its utility; see 1909, 106.

[30] James (1909, 57): "Theoretic truth, truth as passive copying, sought in the sole interest of copying as such, not because copying is *good for something*...seems, if you look at it coldly, to be an almost preposterous idea." See also James 1907, 109-113, and Rorty's introduction to his 1982.

[31] Nietzsche (1886, 11-12): "The falseness of a judgment is for us not necessarily an objection to a judgment...The question is to what extent it is life-promoting, life-preserving, species-preserving, perhaps even species-cultivating. And we are fundamentally inclined to claim that the falsest judgments...are the most indispensable for us...To recognize untruth as a condition of life—that certainly means resisting accustomed value feelings in a dangerous way; and a philosophy that risks this would by that token alone place itself beyond good and evil."

[32] Brand Blanshard argues from infallibilism—the "test" of truth must provide a *proof* of truth—to the conclusion that truth must be epistemic (1941, 268): "If you place the nature of truth in one sort of character and its test in something quite different, you are pretty certain, sooner or later, to find the two falling apart. In the end, the only test of truth that is not misleading is the special nature or character that is itself constitutive of truth. Feeling that this is so, the adherents of correspondence sometimes insist that correspondence shall be its own test. But...if truth does consist in correspondence, no test can be sufficient." He concludes that the character that actually provides our test of truth (he favors coherence) must also be the character that is constitutive of truth: "truth *consists* in coherence" (1941, 269).

[33] Dummett's rejection of non-epistemic theories of truth tends to focus on EC (1976, 75): "The notion of truth, when it is introduced, must be explained, in some manner, in terms of our capacity to recognize statements as true, and not in terms of a condition which transcends human capacities"; see also Dummett 1978. Putnam tends to emphasize the connection between EC and INF (1978, 125): "The most important consequence of metaphysical realism is that *truth* is supposed to be *radically non-epistemic*—we might be 'brains in a vat' and so the theory that is 'ideal' from the point of view of operational utility, inner beauty and elegance, 'plausibility', 'simplicity', 'conservatism', etc. *might be false*. 'Verified' (in any

operational sense) does not imply 'true', on the metaphysical realist picture, even in the ideal limit."

[34] It seems not all advocates of an epistemic notion of truth would be much troubled by the charge of redefinitionism. Dewey (1938) has been interpreted as aiming to *replace* truth with warranted assertibility. Rorty (1979) might have a similar replacement/redefinition in mind; and James could be construed as purposely redefining truth in terms of a multifaceted conception of rational belief.

[35] Dummett (1978, 155): "An understanding of...a statement consists in knowing what counts as evidence adequate for the assertion of the statement, and the truth of the statement can consist only in the existence of such evidence." As Dummett has often pointed out, any such approach must vigorously oppose the idea that falsehood could be identified with the absence of truth. For this would allow the "magical conversion" of *absence* of evidence *for* an evidence-transcendent hypothesis into evidence *against* it.

[36] The original verificationists often characterized verification as truth-entailing. Many of Dummett's characterizations are also most naturally interpreted in this way, e.g.: "An understanding of a statement consists in a capacity to recognize whatever is counted as verifying it, i.e. as conclusively establishing it as true" (Dummett 1976, 71f.).

[37] Evidential defeat is usually defined as follows: *d* defeats *e* as evidence for *x* iff *e* is evidence for *x* but *e* in conjunction with *d* is not evidence for *x*. Note that there is some temptation to read indefeasibility in a factual manner, i.e., as requiring that there be no *fact* such that, if S were apprised of that fact, S would have evidence that defeats his evidence for *x*. Advocates of epistemic truth have to resists the temptation of this reading: the notion of a *fact* is not available for an epistemic definition of truth. Crispin Wright advises the advocates of epistemic truth to make use of indefeasibility which he calls *superassertibility* (1992, 48): "A statement is superassertible...if and only if it is, or can be, warranted and some warrant for it would survive arbitrarily close scrutiny of its pedigree and arbitrarily extensive increments to or other forms of improvement of our information."

[38] Note that justification is person-relative even if the *standards* of justification are absolute. Absolute standards will have different results when applied to persons with different experiences and different background beliefs.

[39] The definitions could be strengthened by using a notion of *absolute* indefeasibility: S has absolutely indefeasible evidence for *x* iff S has justifying evidence for *x* and nobody ever had, has now, or ever will have (or could have) any evidence which, when added to S's total evidence, would defeat S's evidence for *x*. This yields a very strong notion of indefeasible justification—one wonders whether justification with this kind of immunity to counter-evidence is available for anything but elementary *a priori* beliefs and beliefs about one's own conscious states.

[40] See the remarks about epistemic truth and anti-realism in section 6.6. For discussion of some further variants of the epistemic approach see Goldman (1986, chap. 7.2) who has influenced my presentation here; see also Schmitt 1995, chap. 4.

[41] Clear-cut advocates of coherence theories of truth (epistemological or metaphysical) are not too easy to find, especially not prior to the 20th century. Walker (1989, chap. 3) enlists Spinoza as the earliest advocate. But closer scrutiny of Walker's evidence suggests that Spinoza may better be seen as holding a coherence theory of reality in conjunction with an identity theory of truth. Note also that Spinoza explicitly subscribed at least once to the correspondence theory; cf. section 3.1. Walker (1989, chaps. 4-6) also attributes the theory to Hegel and Kant, which seems certainly plausible given their overall views. Still, Kant "grants" explicitly that correspondence is the "nominal definition" of truth (cf. section 3.1); he goes on to argue that the real issue is the question of the "criterion" or "test" of truth; see Kant 1787, B82, and 1800, intro. vii. Bradley is often interpreted as a truth-coherentist: "Truth is an ideal expression of the Universe, at once coherent and comprehensive" (1994, 313). Yet, at times he is very explicit about proposing coherence merely as a "test" of truth; and Candlish

(1989) and Baldwin (1991) argue that Bradley was an identity theorist. Bradley's overall position may be similar to Spinoza's—much the same holds for Joachim 1906. Turning to the 20th century, it is often hard to tell whether a proposed coherence theory is intended as a theory of of justification or whether is also intended as a theory of truth—e.g., despite their titles, Rescher (1973) and Davidson (1986) seem to offer only coherence theories of justification. Blanshard is one of the earliest advocates of coherence-truth who clearly distinguishes the question of the nature of truth from the question of the test of truth: he *argues* that truth must be identified with epistemic coherence; see 1941, chap. 21, and the quotation in note 32.

[42] Blanshard's attempts to resolve this difficulty are uncharacteristically obscure. At one point he says the coherence theory of truth "holds that one system only is true, namely the system in which everything real and possible is coherently included" (Blanshard 1941, 276). To solve the difficulty, 'everything real' would have to mean 'everything true' or 'all the facts'; but a coherence theory of truth is not entitled to making essential use of these notions. At one stage of the evolution of Logical Positivism, Neurath and Carnap advocated a coherence theory. Hempel describes their view in a well known passage (1935, 57): "The system of protocol statements [observation reports] which we call true and to which we refer in every day life and science, may only be characterized by the historical fact, that it is the system which is actually adopted by mankind, and especially by the scientists of our culture circle; and the "true" statements in general may be characterized as those which are sufficiently supported by that system of actually adopted protocol statements." Note that this does not even begin to address the uniqueness problem. However, the position described by Hempel may not be intended as a coherence theory of *truth*. The scare-quotes around 'true' and the phrase 'which we *call* true' may indicate that it is intended solely as a coherence theory of justification. Neurath himself was more explicit—and more radical—espousing linguistic anti-realist pluralism (1934, 102): "We call a content statement "false" if we cannot establish conformity between it and the whole structure of science...we reject the expression that a statement is compared with 'reality', and more so, since for us 'reality' is replaced by several totalities of statements that are consistent with themselves but not with each other."

[43] Peirce was strongly drawn towards the thesis that a proposition is true iff it reflects reality. So he had to make up his mind whether to hold that a proposition would be agreed upon at the limit of inquiry *because* it reflects reality, or whether it reflects reality *because* it would be agreed upon at the limit of inquiry—he seems to have taken the second option; see Kirkham 1992, chap. 3.2. Note that Putnam assumes a coherentist theory of epistemic justification. This is not essential to the proposal as such. In principle this type of subjunctive approach could be coupled with other theories of justification. Brentano offered a foundationalist version. He said that a judgment is true iff it would be made by one who judged with evidence; and with respect to *evidence* he said that a judgment is either immediately evident or made evident by a proof based solely on judgments that are immediately evident; see Brentano 1915. It should be mentioned that Putnam has since given up the proposal quoted in the text but still thinks that truth is epistemic in some manner; see Putnam 1994, p. v.

[44] The claim that anything can be evidence *for anything* is a bit of an exaggeration. Self-evident logical and mathematical beliefs as well as beliefs about one's own experiences are by and large exempt: it is not the case that anything can be evidence for them. This is of little help to (IC), though, because it means that humans can be in epistemically ideal conditions only with respect to these special beliefs.

[45] Similarly, the definition in terms of God would allow one to prove that God exists: simply put 'God exists' for '*x*'. For a more rigorous version of the argument see Plantinga 1982, 64-67. For additional discussion of Putnam's proposal see: Putnam 1981, chaps. 3 and 5; Field 1982; Devitt 1984, chap. 12; Wright 1992, chap. 2; and Alston 1996, chap. 6.

[46] Dummett announces the primacy of the epistemic most concisely (1990, 190): "The concept of truth is born from a more basic concept, for which we have no single clear term, but for which we may here use the term "justifiability""; see also: Dummett 1959, 1976, and 1978; Price 1988; and Ellis 1990. Putnam says that "*truth is not the bottom line*: truth itself gets its life from our criteria of rational acceptability" (1981, 130). But his overall view seems to be that truth and rationality are interdependent notions. The primacy of the *normative* is most explicit in Brandom: "Assessments of truth, no less than assessments of rationality, are normative assessments. Truth and rationality are both forms of correctness. To ask whether a belief is true is to ask whether it is in some sense proper...The business of truth talk is to evaluate the extent to which a state or act has fulfilled a certain kind of responsibility" (1994, 17); and: "Being true is then to be understood as being *properly* taken-true (believed)" (1994, 291); cf. section 8.2. For further objections to the primacy of the epistemic/normative see: Kirkham 1992, chap. 2; Schmitt 1995, chap. 7; Alston 1996, chap. 8.

[47] A sociological observation may not be out of place here. Contemporary analytic philosophers who specialize in epistemology—reliabilists, virtue theorists, foundationalists, and coherentists alike—are by and large opposed to an epistemic approach to truth.

[48] See Wright 1992, chap. 3. The term is Wright's, the contrast itself is from Plato's *Euthyphro*: Are pious acts pious because they are loved by the gods; or are they loved by the gods because they are pious?

[49] For more discussion of realism and anti-realism in relation to truth see: Dummett 1976, 1978; Putnam 1978, 1981, 1994, parts 4 and 5; Field 1982; Blackburn 1984; Devitt 1984; Vision 1988; Kirkham 1992; Wright 1992; and Alston 1996.

[50] Dummett's view of truth may be an exception; it may have *inter*-subjectivist anti-realism as an *immediate* consequence. For he tends to say things like: "A verificationist theory represents an understanding of a sentence as consisting in a knowledge of *what counts as* conclusive evidence for its truth" (1976, 88, my emphasis). If this is meant seriously, then the nature of dinosaur metabolism would, for Dummett, depend directly on what we *believe to be* conclusive evidence about dinosaur metabolism. See also the quotation in note 36, and his 1978, pp. 161f., where he seems to be saying that in many cases inductive evidence *is* conclusive because "in practice we treat a great deal of inductive evidence as conclusive."

[51] Kant is surely the most famous anti-realist correspondence theorist; see section 3.1 and note 41. James may also fall into this category; see note 26. Kirkham mentions McTaggart and Sellars as further candidates; see Kirkham 1992, chap. 4.6.

[52] The "Kantian argument" surely originates with Berkeley who tried to establish idealism by showing that Descartes's realist representationalism must lead into skepticism. Davidson gives a contemporary version (1986, 307): "If meanings are given by objective truth conditions there is a question how we can know that the conditions are satisfied, for this would appear to require a confrontation between what we believe and reality; and the idea of such a confrontation is absurd." For a small sample of other versions see: Neurath, 1934; Hempel 1935, 50f.; Blanshard 1941, 226-35; Rorty 1979, chap. 6 and passim; Rorty 1982, introduction and essays 1 and 9.

[53] Stove 1991, chaps. 5 and 6, offers a nice discussion of Berkeley's Gem—the reference is, of course, to Berkeley's argument that one cannot conceive of things that are unconceived; see Berkeley 1734, secs. 23-24.

[54] According to an argument sometimes called the "space of reasons argument", causal input from the world cannot "generate" a *rational* relation. The idea seems to be that x can be evidence *for* a belief (proposition, judgment), only if x is *itself* a belief (proposition, judgment)—only if x is a potential truth *bearer*. Only truth bearers can be *reasons for* truth bearers: the relation of evidence cannot break out from the space of reasons. For clear statements of this argument, which derives from Sellars, see, e.g., Davidson 1986, and BonJour 1985, chap. 4 (BonJour now rejects it). Contemporary foundationalism and reliabilism are *built on* the rejection of this argument. This issue, which is maybe *the* most

fundamental issue in epistemology, is surely related to what I have here called "the Kantian argument." But how exactly it bears on the argument—which is, after all, about *truth*—is not altogether clear. For some more discussions of various versions of the epistemological argument see: Goldman 1986, chap. 2; Schmitt 1995, 84-86 and chap. 7; Alston 1996, chap. 3 and chap. 7, sec. xi.

[55] Frege (1918, 60) gives a variation on the Kantian argument: "But could we not maintain that there is truth when there is correspondence in a certain respect? But which respect? For in that case what ought we to do so as to decide whether something is true? We should have to inquire whether it is *true* that an idea and a reality, say, correspond in the specified respect. And then we should be confronted by a question of the same kind, and the game could beging again." Frege noticed that this regress argument would speak against any definition of truth. He concluded that truth is indefinable. Since the argument clearly presupposes assumptions (i) and (ii), it should be rejected.

[56] The thesis that truth is relative is sometimes interpreted as the claim that every proposition is true for someone and false for someone else, so that it would be an expression of absolutism to say that some proposition is true for everyone. This interpretation does not comport well with the intentions of known relativists. They tend to try to convince *everyone* of relativism; and it is hard to see why they should want to deny that there are some issues on which everyone agrees. It is more plausible to interpret the relativist as advancing the conceptual or semantic thesis that the logical form of 'x is true' has to be represented as 'x is true for S', which makes it possible for x to be true for someone and false for someone else without entailing that every x actually is true for someone and false for someone else.

[57] Analogous considerations apply to societal or cultural relativism: 'x is true for a society' means something like: all, or most, or the grown-up, or the sane, members of the society believe x, or tend to believe x, or are somehow implicitly committed to believing x, etc. Again, cultural relativism seems designed for peaceful coexistence and seems to be an overreaction to the undoubted fact that many differences between cultures are indeed appropriately dissolvable through relativization. *How many* is of course a sticky issue—once we get to ethical disagreements the pressure to keep the peace by embracing relativism may become strong indeed. In general, personal and cultural relativisms tend to raise exactly analogous issues. The main difference between the two is that it is much harder to specify precisely what 'true for' means for a cultural relativist.

[58] For Plato's version of the self-refutation objection see *Theaetetus* 170^d-171^c. Plato drops the qualifier 'for S' at a crucial stage of the argument. A valiant attempt to make the refutation work with the qualifier restored is undertaken in Burnyeat 1976. For Plato's argument that, according to Protagoras, we have within ourselves "the criterion of what will happen tomorrow", see *Theaetetus* 177^c-179^b. The argument that the relativist cannot make sense of expertise, etc., and cannot genuinely *assert* his own position can be found in *Theaetetus* 160^e-170^d.

[59] Cf. Putnam 1981, 120-21. Note that the objection does not complain about 'it is true for S that it is true for S* that it is true for S** that p'. This is just the ordinary iteration of truth: 'it is true that it is true that it is true that p'. The Protagorean relativist handles this easily with 'S** believes that S* believes that S believes that p', which certainly does make sense. But the iterated truth-predicates in the text are different, and it is not easy to see what they mean. Putnam claims that relativism is inconsistent; but this argument does not quite bear out his claim. Although the problem is certainly grave, "wobbling" is not exactly the same as inconsistency.

[60] Deflationary views and correspondence views share a common ancestor, namely Aristotle's minimal correspondence definition of truth (cf. section 3.1). Deflationism is best understood as aiming to deflate correspondence theories. To put it a bit paradoxically:

according to deflationists, the right theory of truth is a correspondence theory—but without correspondence and without facts.

[61] Although Ramsey (1927) and Ayer (1936, chap. 5) both start out with remarks suggesting such a redundancy theory, they immediately modify it, offering *paraphrases* for truth-talk instead of simply erasing the word 'true'. Since paraphrases paraphrase the meaning of truth-talk, this strategy amounts to a more moderate view according to which 'true' contributes a meaning to the sentence in which it occurs and is, broadly speaking, definable (in some deflationary manner). Frege, though drawn to a redundancy theory, did not accept it and held that 'true' has a meaning which is *sui generis* and indefinable; see Frege 1918, 60.

[62] Philosophers are so used to 'p's and 'q's, they sometimes forget that a schema like 'if he asserts that p, then p' does not say anything, but merely amounts to 'if he asserts that   then   ', combined with the instruction to fill the blanks with the same sentence. At times 'p's are used inconsistently, e.g., one can find formulations like 'John knows that p only if p is true'. Note that there is no coherent way of substituting for 'p' in this formulation without making other changes. Usually, such small inconsistencies don't make much difference. Unfortunately, when it comes to theories of truth, they can make all the difference. For more on redundancy theories see: Davidson 1969; Prior 1971; Williams 1976; Forbes 1986; Field 1986, sec. 1; Kirkham 1992, chap. 10; and Soames 1999, chaps. 2 & 8. For the questions raised by quantifications like (4), see sections 8.2 and 8.4, note 72.

[63] But Brandom (1994, 303-305) modifies the theory in a way that aggravates one of its problematic features. He construes 'is true' as a "prosentence-forming operator", i.e., as an operator that yields a prosentence when applied to a singular term, like 'that'. This modification threatens to make the prosentence/predicate distinction inscrutable: predicates are often defined as operators that yield sentences when applied to singular terms. Given such a definition, 'is true' simply *is* a predicate, no matter what else it is called. For more on the prosentential theory see: Grover 1992; Forbes 1986; and Kirkham 1992, chap. 10.6. For the performative approach see Strawson 1949; Price 1988; Kirkham 1992, chap. 10; and Soames 1999; chap. 8.

[64] To my knowledge, only one philosopher has officially renounced (8) and its kin—and even he did not quite *deny* them—Otto Neurath (1944, 12): "We should be doubtful even in admitting a definition of 'true' which implies that the saying, 'There is an elephant here,' may be called true if and only if there is an elephant here. Even this sounds like an absolute expression...which we do not know how to fit into a framework based on observation statements." For a brief survey of the dispute over the question whether Tarski's work should be seen as promoting the cause of deflationism or rather the cause of the correspondence approach to truth, see Kirkham 1992, sec. 5.8.

[65] Note that 'snow', being a *mass*-noun, fits only uneasily into the category of names; but it wouldn't fit easily into the category of predicates either. Difficulties like this are among the reasons why Tarski was leery of defining truth for natural languages.

[66] To see why this must work, one can momentarily think of the sentence 'Snow is white' as if it were the degenerate predicate '*x* is such that snow is white'. Obviously: 'Snow is white' is true iff '*x* is such that snow is white' is satisfied by all objects. This is not exactly what Tarski did, but it works by the same principle. I should point out that talk of "primitive predicates" and "primitive satisfaction" is non-Tarskian; it derives from Field 1972. For Tarski's original construction, see Tarski 1933, §3. A detailed introductory exposition can be found in Kirkham 1992, chap. 5. For various alternative methods of constructing Tarski-style truth definitions plus illuminating discussions see, e.g.: Quine 1970, chap. 3; Etchemendy 1988; McGee 1991, chap. 3; and Soames 1999, chaps. 3 & 4.

[67] Compare section 3.2, where I indicated how correspondence theorists might want to make use of recursive constructions to avoid (complex) facts. But note that a correspondence theorist will *not* base the recursions on (10) and (11): according to the correspondence theorist, there *are* substantive notions of satisfaction and reference behind the notion of truth;

cf. Field 1972. Turning a recursive definition into an explicit definition requires additional resources from set theory. Since it is rather dubious that set theory can claim logical status, a deflationist would have to admit that the qualification of the original deflationary idea is actually more serious than is suggested in the text; see, e.g.: Tarski 1933; Quine 1970, chap. 3; and McGee 1991, chap. 3. Jan Woleński reminds me that one may well question the legitimacy of the deflationist move to simply discount logico-structural (and set-theoretic) content as irrelevant to the issue whether truth is a substantive notion. To what extent Tarski's work is seen as promoting the cause of deflationism will depend on the stance taken towards these issues.

[68] Tarski gives homophonic versions of Convention T in 1944, sec. 4, and in 1969. The point raised in the text has been made repeatedly and in a number of different ways by different authors, e.g., by Dummett (1959, 7) and by Davidson (1973, 321) against his own former views. My presentation is indebted to Etchemendy 1988, and Soames 1999, chap. 4.

[69] Inspection of (13) and (13a) reveals that Tr-sentences express necessary truths provided the syntactic identities they contain are necessary—Tarskian truth definitions presuppose "logical syntax" which seems to carry existential commitments to syntactic objects. Note that taking the syntactic identities in (13) and (13a) to be contingent (on the grounds that they imply existence claims) would not solve the problem, because the contingency of the T-sentences is additional to the existence claims implied by their left-hand sides: it is due to the contingency of meaning. This can be seen from the fact that the following is still contingent: If 'snow is white' exists, then 'snow is white' is true iff snow is white.

[70] See Gupta and Belnap 1993, chap. 1, for more discussion of the deflationary import of Convention T. Note that the official version of Convention T gives only a sufficient condition of adequacy. Tarski's remarks about Convention T suggest he also intended it as a necessary condition (cf. Tarski 1933, 187, and 1944, sec. 4.); and this is how the convention is usually interpreted. Of course, if T-sentences are contingent, then Convention T will face the objection that it is also inadequate as a necessary condition of adequacy because it is too strong.

[71] Field (1986, 57) stresses the importance of 'false' for allowing us to reject a theory on the grounds that it has unacceptable consequences without knowing which part of the theory is to blame. Next to Quine, the main advocates of disquotationalism are Leeds 1978; Field 1994 and 1994a; and McGee 1993. Field 1986 is an important contribution to the discussion of disquotationalism, although at the time Field was still in opposition.

[72] This somewhat paradoxical way of characterizing (DT) is inspired by Field 1986, 57f. (DT) was considered by Tarski but rejected in favor of his recursive approach. See Tarski 1933, §1, and Soames 1999, 86-92. The schematic quantifier-phrases—'for some p' and 'for every p'—are the ones found dubious at the end of section 8.1, giving rise to the dilemma with respect to (4). There are, broadly speaking, two views about the use of such phrases. (*A*) According to the "orthodox" Quinean view, the phrase 'for some p' is a pseudo-quantifier not fit for regular meaningful use. It is employed in (DT) only as an aide for the eye to gesture at the infinite disjunction which the truth predicate allows us to abbreviate. Note that a Quinean disquotationalist would *not* use the 'for some p'-locution when recasting (4) into canonical notation: (4) is interpreted in terms of standard first-order predicate logic in combination with the indispensable truth predicate, i.e.: $(\forall x)$(if he utters x, then x is true). This seems to be the very point of Quine's version of disquotationalism. (*B*) It is sometimes held that 'for some p' can be taken as a genuine quantifier, the ***substitutional quantifier***, which is (sort of) available in English and can be used to ***define*** truth by way of (DT) ***and*** to make sense of (4). A problem with this "iconoclastic" view is that a substitutional quantifier with schematic sentences letters as substituends does not seem to make much sense. What, for example, could the following mean: 'For some p, p'? The worry is that we tacitly read 'for some p, ...p...' as saying that the schema '...p...' has some *true* substitution instances—see section 8.1. For more

on these issues see: Quine 1970, 92-94; Forbes 1986; Horwich 1990, secs. 5 & 6; Grover 1992; Field 1994; David 1994, chap. 4; and Soames 1999, 39-49.

[73] There is a deeper agenda to this. In their need for structure, Tarskian truth theories will analyze belief and meaning as relations—'*x* believes *y*' and '*x* means *y*'—where '*y*' has to range over *objects*. If the problem of falsehood is to be avoided (cf. section 3.3), these objects will likely turn out to be propositions, which would render deflationary theories of *sentence-truth* pointless. The primary motivation for such theories lies in trying to do without propositions and other "content-objects". The structural analyses required for Tarskian truth definitions threaten to invite propositions through the back door, making it difficult to see why one should not let them in at the front and accept propositions as primary bearers of truth. Since (DT) is not interested in sentential structure, it promises to avoid this trap.

[74] Field 1994, 250. Field qualifies this equivalence thesis to take care of the existential commitment involved in the claim that *u* is true which might keep it from being cognitively equivalent to *u* for some persons. Field's formulation does *not* introduce a relativized notion of truth, '*x* is true for *y*', where '*y*' would range over different speakers. It introduces different restricted absolute notions; each restricted to the utterances a speaker understands. Note that Tarskian truth predicates are sometimes misunderstood as relativized because they are often talked about in terms of 'true in L'. But 'L' is not a variable in this expression; it indicates an absolute notion of truth restricted to a specific language L. There cannot be a purely disquotational or a Tarskian notion of *true in y* for variable *y*, since each notion of truth is characterized with reference to all and only the sentences of one specific language: truth is immanent—as Quine would say.

[75] McGee 1993 argues that disquotationalism can avoid absurd modal consequences, provided disquotational truth is construed as containing hidden indexicals: 'is true' would then not just be *restricted* to my actual idiolect, it would *mean* 'is a true sentence of the idiolect I actually speak'. For more discussion and criticism of disquoationalism see: Putnam 1978, lectures 1 &2; 1994, chaps. 13 & 17; Devitt 1984, chap. 6; Gupta 1993a; David 1994; and Schmitt 1995, chap. 5. See also section 8.5.

[76] For more discussion of vagueness in relation to deflationism about sentence-truth see: Field 1994a; David 1994, chap. 5.8; Wright 1992, 61-64. Horwich (1990, secs. 23-29) inists that, unlike sentences, propositions cannot be gappy and adapts Field's suggestion to his deflationary view about proposition-truth; see Schmitt 1995, chap. 5, for discussion.

[77] Cf. Horwich 1990, 38. Horwich's also offers an acount of sentence-truth similar to extended disquotationalism; see 1990, chap. 6.

[78] Horwich's failed attempt at deriving the general law can be found on p. 23 of his 1990. Gupta (1993a, sec. 3) argues that the same mistake underlies Williams' (1986, 232) and Horwich's (1990, 24) claim that deflationism can explain the "law" that true beliefs engender successful action. It appears that Tarski (1933, §5) was the first—but not very enthusiastic—advocate of minimalism (applied to sentence-truth). Believing that his method of explicitly defining truth via an antecedent recursive definition would not work for "languages of infinite order," he considered adding as axioms to the metatheory all instances of his schema '*x* is Tr if and only if p', where 'p' translates the sentence whose name is substituted for '*x*'. He observed that the resulting theory "would be a highly incomplete system, which would lack the most important and most fruitful general theorems" (1933, 257); cf. also McGee 1991, 72.

[79] Schemata play an important role in other theories as well. Sentential logic and first-order predicate logic are usually presented with schematic sentence-letters and predicate-letters. Peano Arithmetic contains the schematic Induction Axiom, and set theory contains the schematic Replacement Axiom. If schemata are acceptable in these basic theories—even as parts of our theories of numbers and sets—why should it be objectionable to formulate the theory of truth in terms of a schema? Again, there is a ready reply. We understand, say, the Induction Axiom schema, because we recognize that all its substitution instances are *true*: our understanding of number depends on our grasp of truth.

REFERENCES

Alston, W. P.: 1996, *A Realist Conception of Truth*, Cornell University Press, Ithaca and London.

Aristotle: *The Complete Works of Aristotle: The Revised Oxford Translation*, J. Barnes (ed.), Princeton University Press, Princeton, 1984.

Armstrong, D. M.: 1997, *A World of States of Affairs*, Cambridge University Press, Cambridge.

Austin, J. L.: 1950, 'Truth', *Proceedings of the Aristotelian Society*, suppl. vol. **24**, reprinted in *Philosophical Papers*, 3d ed., J. O. Urmson and G. J. Warnock (eds.), Oxford University Press, Oxford, 1979, pp. 117-33.

Austin, J. L.: 1961, 'Unfair to Facts', in *Philosophical Papers*, 3d ed., J. O. Urmson and G. J. Warnock (eds.), Oxford University Press, Oxford 1979, pp. 154-74.

Ayer, A. J.: 1936, *Language, Truth, and Logic*, 2d ed., 1946, Dover, New York.

Baldwin, T.: 1991, 'The Identity Theory of Truth', *Mind* **100**, 35-52.

Barwise, J. and J. Perry: 1983, *Situations and Attitudes*, MIT Press, Cambridge, Mass.

Bealer, G.: 1982, *Quality and Concept*, Clarendon Press, Oxford.

Bennett, J.: 1988, *Events and their Names*, Hackett Pub. Comp., Indianapolis.

Berkeley, G.: 1734, *A Treatise Concerning the Principles of Human Knowledge*, K. Winkler (ed.), Hackett Pub. Comp., Indianapolis, 1982.

Blackburn, S.: 1984, *Spreading the Word: Groundings in the Philosophy of Language*, Clarendon Press, Oxford.

Blanshard, B.: 1941, *The Nature of Thought*, vol. 2, The Macmillan Company , New York.

BonJour, L.: 1985: *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge, Mass.

Bradley, F. H.: 1994: *Writings on Logic and Metaphysics*, J. W. Allard and G. Stock (eds.), Clarendon Press, Oxford.

Brandom, R. B.: 1994, *Making it Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press, Cambridge, Mass.

Brentano, F.: 1915, 'Über den Satz: veritas est adaequatio rei et intellectus', in *Wahrheit und Evidenz*, Felix Meiner Verlag, Hamburg, 1974, pp. 137-39.

Burnyeat, M.F.: 1976, 'Protagoras and Self-Refutation in Plato's *Theaetetus*', *Philosophical Review* **85**, 172-95.

Candlish, S.: 1989, 'The Truth About F. H. Bradley', *Mind* **98**, 331-48.

Cartwright, R. L.: 1987, 'A Neglected Theory of Truth', in *Philosophical Essays*, MIT Press, Cambridge, Mass., pp. 71-93.

Chisholm, R.M.: 1976, *Person and Object*, George Allen & Unwin, London.

Chisholm, R.M.: 1977, *Theory of Knowledge*, 2nd ed., Prentice Hall, Englewood Cliffs, N. J.

David, M.: 1994, *Correspondence and Disquotation: An Essay on the Nature of Truth*, Oxford University Press, New York.

Davidson, D.: 1969, 'True to the Facts', *The Journal of Philosophy* **66**, 748-64.

Davidson, D.: 1973, 'Radical Interpretation', *Dialectica* **27**, 313-28.

Davidson, D.: 1977, 'Reality Without Reference', *Dialectica* **31**, 247-53.

Davidson, D.: 1986, 'A Coherence Theory of Truth and Knowledge', in E. LePore (ed.), *Truth and Interpretation*, Blackwell, Oxford, pp. 307-19.

Descartes, R.: 1639, 'Letter to Mersenne: 16 October 1639', in J. Cottingham, R. Stoothoff, D. Murdoch, and A. Kenny (eds.), *The Philosophical Writings of Descartes*, vol. 3, Cambridge University Press, Cambridge 1991, 138-40.

Devitt, M.: 1982, *Designation*, Columbia University Press, New York.

Devitt, M.: 1984, *Realism and Truth*, 2d ed., Blackwell, Oxford, 1991.

Dewey, J.: 1938: *Logic: The Theory of Inquiry*, reprinted in *John Dewey: The Later Works 1925-1953*, vol. 12, Southern University Press, Carbondale, 1986.

Ducasse, C. J.: 1940, 'Propositions, Opinions, Sentences, and Facts', *The Journal of Philosophy* **37**, 701-11.

Dummett, M.: 1959, 'Truth', *Proceedings of the Aristotelian Society*; reprinted in *Truth and Other Enigmas*, Harvard University Press, Cambridge, Mass. 1978, pp. 1-24.

Dummett, M.: 1976, 'What Is a Theory of Meaning (II)', in G. Evans and J. McDowell (eds.), *Truth and Meaning*; reprinted in M. Dummett, *The Seas of Language*, Clarendon Press, Oxford 1993, pp. 34-93. (Page references are to the reprint.)

Dummett, M.: 1978, 'Realism', in *Truth and Other Enigmas*, Harvard University Press, Cambridge, Mass. 1978, 145-65.

Dummett, M.: 1990, 'The Source of the Concept of Truth', in G. Boolos (ed.), *Meaning and Method: Essays in Honor of Hilary Putnam*; reprinted in M. Dummett, *The Seas of Language*, Clarendon Press, Oxford 1993, 188-201. (Page references are to the reprint.)

Ellis, B.: 1990, *Truth and Objectivity*, Blackwell, Oxford.

Etchemendy, J.: 1988, 'Tarski on Truth and Logical Consequence', *The Journal of Symbolic Logic* **53**, 51-79.

Field, H.: 1972, 'Tarski's Theory of Truth', *The Journal of Philosophy* **69**, 347-75.

Field, H.: 1978, 'Mental Representation', *Erkenntnis* **13**; reprinted in N. Block (ed.), *Readings in Philosophy of Psychology*, vol. 2, Harvard University Press, Cambridge, Mass. 1981, pp. 78-114.

Field, H.: 1982, 'Realism and Relativism', *The Journal of Philosophy* **79**, 553-67.

Field, H.: 1986, 'The Deflationary Concept of Truth', in G. Macdonald and C. Wright (eds.), *Fact, Science and Morality: Essays on A. J. Ayer's 'Language, Truth & Logic'*, Basil Blackwell, Oxford, pp. 55-117.

Field, H.: 1994, 'Deflationist Views of Meaning and Content', *Mind* **103**, 249-85.

Field, H.: 1994a, 'Disquotational Truth and Factually Defective Discourse', *The Philosophical Review* **103**, 405-52.

Fine, K.: 1982, 'First-Order Modal Theories III—Facts', *Synthese* **53**, 43-122.

Fodor, J.A.: 1975, *The Language of Thought*, Crowell, New York.

Fodor, J.A.: 1978, 'Propositional Attitudes', *The Monist* **61**; reprinted in *Representations: Philosophical Essays on the Foundations of Cognitive Science*, MIT Press, Cambridge, Mass., 1983, pp. 177-203.

Forbes, G.: 1986, 'Truth, Correspondence and Redundancy', in G. Macdonald and C. Wright (eds.), *Fact, Science and Morality: Essays on A. J. Ayer's Language, Truth & Logic*, Basil Blackwell, Oxford, pp. 27-54.

Frege, G.: 1891, 'Funktion und Begriff', Herman Pohle, Jena. English translation: 'Function and Concept', in Frege, 1984. (Page references are to the original pagination.)

Frege, G.: 1892, 'Über Sinn und Bedeutung', *Zeitschrift für Philosophie und philosophische Kritik* **100**, 25-50. English translation: 'On Sense and Meaning', in Frege, 1984.

Frege, G.: 1918, 'Der Gedanke: Eine logische Untersuchung', *Beiträge zur Philosophie des deutschen Idealismus* **1**, 58-77. English translation: 'Thoughts', in Frege, 1984. (Page references are to the original pagination.)

Frege, G.: 1984, *Collected Papers on Mathematics, Logic, and Philosophy*, B. McGuinness (ed.), Basil Blackwell, Oxford.

Geach, P. T.: 1965, 'Assertion', *Philosophical Review* **74**, reprinted in *Logic Matters*, Basil Blackwell, Oxford 1972, pp. 254-69.

Gettier, E.: 1962, 'Is Justified True Belief Knowledge?', *Analysis* **23**, 121-23.

Goldman, A.: 1986, *Epistemology and Cognition*, Harvard University Press, Cambridge, Mass.

Grover, D. L.: 1992, *A Prosentential Theory of Truth*, Princeton University Press, Princeton.

Grover, D. L., J. C. Camp, and N. D. Belnap: 1975, 'The Prosentential Theory of Truth', in *Philosophical Studies* **27**, 73-125.

Gupta, A.: 1993, 'Minimalism', *Philosophical Perspectives* **7**, 359-69.

Gupta, A.: 1993a, 'A Critique of Deflationism', *Philosophical Topics* **21**, 57-81.

Gupta, A., N. and Belnap: 1993, *The Revision Theory of Truth*, MIT Press, Cambridge, Mass.

Harman, G.: 1973, *Thought*, Princeton University Press, Princeton.

Hegel, G. W. F.: 1830, *The Science of Logic*, in *Hegel's Logic*, translated by W. Wallace, Clarendon Press, Oxford, 1975.

Hempel, C. G.: 1935: 'On the Logical Positivists' Theory of Truth', *Analysis* **2**, 49-59.

Hornsby, J.: 1997, 'Truth: The Identity Theory', *Proceedings and Addresses of the Aristotelian Society* **97**, 1-24.

Horwich, P.: 1990, *Truth*, Basil Blackwell, Oxford.

James, W.: 1907: *Pragmatism*; reprinted in *Pragmatism and the Meaning of Truth*, Harvard University Press, Cambridge, Mass., 1975.

James, W.: 1909, *The Meaning of Truth*, reprinted in *Pragmatism and the Meaning of Truth*, Harvard University Press, Cambridge, Mass. 1975. (References are to bracketed page numbers.)

Joachim, H. H.: 1906, *The Nature of Truth*, 2d ed., Oxford University Press, Oxford 1936.

John Buridan: *Sophisms On Meaning and Truth*, translated by T.K. Scott, Appleton-Century-Crofts, New York, 1966.

Kirkham, R. L.: 1992, *Theories of Truth: A Critical Introduction*, MIT Press, Cambridge, Mass.

Kant, I.: 1787, *Critique of Pure Reason*, translated by N.K. Smith, St. Martin's Press, New York, 1929.

Kant, I.: 1800, *Logic*, translated by R. S. Hartmann and W. Schwarz, Dover, New York, 1974.

Kaplan, D. 1977, 'Demonstratives', in J. Almog, J. Perry, and H. Wettstein (eds.), *Themes on Kaplan*, Oxford Unversity Press, New York, pp. 481-563.

Kneale, W., and M. Kneale: 1962, *The Development of Logic*, Clarendon Press, Oxford.

Kripke, S.: 1972, *Naming and Necessity*, Harvard University Press, Cambridge, Mass., 1980.

Kripke, S.: 1975, 'Outline of a Theory of Truth', *The Journal of Philosophy* **72**, 690-716.

Leeds, S.: 1978, 'Theories of Reference and Truth', *Erkenntnis* **13**, 111-29.

Lehrer, K.: 1990, *Theory of Knowledge*, Westview, Boulder.

Leibniz, G. W.: 1765, *New Essays on Human Understanding*, translated and edited by P. Remnant and J. Bennett, Cambridge University Press, Cambridge, 1982.

Locke, J.: 1700, *An Essay Concerning Human Understanding*, P. H. Nidditch (ed.), Clarendon Press, Oxford, 1975.

Long, A. A., and D. N. Sedley: 1987, *The Hellenistic Philosophers*, vol. 1, Cambridge University Press, Cambridge.

McGee, V.: 1991, *Truth, Vagueness, and Paradox*, Hackett Publishing Company, Indianapolis and Cambridge, Mass.

McGee, V.: 'A Semantic Conception of Truth?', *Philosophical Topics* **21**, 83-111.

McGinn, C.: 1989, *Mental Content*, Blackwell, Oxford.

Mackie, J. L.: 1973, *Truth, Probability, and Paradox*, Clarendon Press, Oxford.

Martin, R.L. (ed.): 1984, *Recent Essays on Truth and the Liar Paradox*, Clarendon Press, Oxford.

Meinong, A.: 1910, *Über Annahmen*, 2d ed. (1st ed. 1902), in A. Meinong, *Gesamtausgabe*, Band iv, R. Haller and R. Kindinger (eds.), Akademische Druck- u. Verlagsanstalt, Graz, 1977.

Moore, G. E.: 1899, 'The Nature of Judgment', *Mind* **8**; reprinted in *Selected Writings*, edited by T. Baldwin, Routledge, London and New York, 1993, pp. 1-19.

Moore, G. E.: 1901-02, 'Truth and Falsity', in the *Dictionary of Philosophy and Psychology*; reprinted in *Selected Writings*, T. Baldwin (ed.), Routledge, London and New York, 1993, pp. 20-22.

Moore, G. E.: 1908, 'William James' 'Pragmatism'', *Proceedings of the Aristotelian Society* **8**; reprinted in *Philosophical Studies*, Littlefield, Adams & Co., Paterson, N.J., 1959, pp. 97-146.

Moore, G. E.: 1953, *Some Main Problems of Philosophy* (lectures given in 1910-11), George Allen & Unwin, London.

Mulligan, K., P. Simons, and B. Smith: 1984, 'Truth makers', *Philosophy and Phenomenological Research* **44**, 287-321.

Neurath, O.: 1934, 'Radical Physicalism and the "Real World"'; reprinted in *Philosophical Papers 1913-1946*, Reidel, Dordrecht, 1983. (Page references are to the reprint.)

Neurath, O.: 1944, 'Foundations of the Social Sciences', *International Encyclopedia of Unified Science*, vol. 2, University of Chicago Press, Chicago.

Nietzsche, F., 1886, *Beyond Good and Evil: Prelude to a Philosophy of the Future*, translated by W. Kaufmann, Vintage Books, New York, 1989.

Nuchelmans, G.: 1973, *Theories of the Proposition: Ancient and Medieval Conceptions of the Bearers of Truth and Falsity*, North-Holland, Amsterdam.

Olson, K. R.: 1987, *An Essay on Facts*, CSLI lecture notes, Stanford.

Peirce, C. S.: 1878, 'How to Make Our Ideas Clear', in *Philosophical Writings of Peirce*, J. Buchler (ed.), Dover Publications, New York 1955, pp. 23-41.

Peirce, C. S.: (1901-02), 'Truth and Falsity', in J. M. Baldwin (ed.), *Dictionary of Philosophy and Psychology*, MacMillan, New York, pp. 718-20.

Plantinga, A.: 1982, 'How to be an Anti-Realist', *Proceedings and Addresses of the American Philosophical Association* **56**, 47-70.

Plato: *Complete Works*, J. M. Cooper (ed.), Hackett, Indianapolis, 1997.

Price, H.: 1988, *Facts and the Function of Truth*, Basil Blackwell, Oxford.

Prior, A. N.: 1967, 'Correspondence Theory of Truth', in P. Edwards (ed.), *The Encyclopedia of Philosophy*, vol. 2, Macmillan Publishing Co. Inc. & The Free Press, New York, pp. 223-32.

Prior, A. N.: 1971, *Objects of Thought*, edited by P.T. Geach and A.J.P. Kenny, Clarendon Press, Oxford.

Putnam, H.: 1975, 'The Meaning of Meaning', in *Mind, Language, and Reality: Philosophical Paper*, vol. 2, Cambridge University Press, pp. 215-71.

Putnam, H.: 1978, *Meaning and the Moral Sciences*, Routledge & Kegan Paul, Boston.

Putnam, H.: 1981, *Reason, Truth and History*, Cambridge University Press, Cambridge.

Putnam, H.: 1994, *Words and Life*, Harvard University Press, Cambridge, Mass.

Quine, W. V. O.: 1960, *Word and Object*, MIT Press, Cambridge, Mass.

Quine, W. V. O.: 1970, *Philosophy of Logic*, 2d ed., Harvard University Press, Cambridge, Mass., 1986.

Quine, W. V. O.: 1987, *Quiddities: An Intermittently Philosophical Dictionary*, Harvard University Press, Cambridge, Mass.

Ramsey, F. P.: 1927, 'Facts and Propositions', *Proceedings of the Aristotelian Society*, suppl. vol. 7; reprinted in *Philosophical Papers*, D. H. Mellor (ed.), Cambridge University Press, Cambridge, 1990, pp. 34-51.

Rescher, N.: 1973, *The Coherence Theory of Truth*, Oxford University Press, Oxford.

Rorty, R.: 1979, *Philosophy and the Mirror of Nature*, Princeton University Press, Princeton.

Rorty, R.: 1982, *Consequences of Pragmatism*, University of Minnesota Press, Minneapolis.

Rorty, R.: 1986, 'Pragmatism, Davidson and Truth', in E. LePore, *Truth and Interpretation*, Blackwell, Oxford, pp. 333-55.

Russell, B.: 1904, 'Meinong's Theory of Complexes and Assumptions', *Mind* **13**; reprinted in *Essays in Analysis*, D. Lackey (ed.), Braziller, New York 1973, 21-76.

Russell, B.: 1906-07, 'On the Nature of Truth', *Proceedings of the Aristotelian Society* **7**, 28-49.

Russell, B.: 1908, 'William James's Conception of Truth', *Albany Review*; reprinted in *Philosophical Essays*, Simon and Schuster, New York 1966, pp. 112-30.

Russell, B.: 1912, *Problems of Philosophy*, London; reprinted at Oxford University Press, Oxford, 1971. (Page references are to the reprint.)

Russell, B.: 1918, 'The Philosophy of Logical Atomism', in *Logic and Knowledge: Essays 1901-1950*, R. C. Marsh (ed.), George Allen and Unwin, London, 1956, 177-281.

Salmon, N., and S. Soames (eds.): 1988, *Propositions and Attitudes*, Oxford University Press, Oxford.

Schiller, F. C. S.: 1907, *Studies in Humanism*, Macmillan, London.

Schmitt, F. F.: 1995, *Truth: A Primer*, Westview Press, Boulder.

Smith, B.: 1994, *Austrian Philosophy: The Legacy of Franz Brentano*, Open Court, Chicago and LaSalle.

Soames, S.: 1999, *Understanding Truth*, Oxford University Press, New York.

Spinoza, B.: 1677, *The Ethics*, translated by R.M.H. Elwes, Dover, New York, 1955.

Stove, D.: 1991, *The Plato Cult and other Philosophical Follies*, Blackwell, Oxford.

Strawson, P. F.: 1949, 'Truth', *Analysis* **9**, 83-97.

Strawson, P. F.: 1950, 'Truth', *Proceedings of the Aristotelian Society*, suppl. vol. **24**; reprinted in G. Pitcher (ed.), *Truth*, Prentice-Hall, Englewood Cliffs, 1964, pp. 32-53.

Strawson, P. F.: 1964, 'A Problem About Truth—A Reply to Mr. Warnock', in G. Pitcher (ed.), *Truth*, Prentice-Hall, Englewood Cliffs, pp. 68-84.

Tarski, A.: 1933, 'Pojęcie prawdy w językach nauk dedukcyjnych', Warsaw; English translation: 'The Concept of Truth in Formalized Languages', in *Logic, Semantics, Metamathematics*, 2nd ed., translated by J.H. Woodger, edited by J. Corcoran, Hackett Publishing Company, Indianapolis 1983, pp. 152-278. (Page references are to the translation.)

Tarski, A.: 1944: 'The Semantic Conception of Truth,' *Philosophy and Phenomenological Research* **4**, 341-75.

Tarski, A.: 1969, 'Truth and Proof', *Scientific American* **220**, June, 63-77.

Taylor, B.: 1976, 'States of Affairs', in G. Evans and J. McDowell (eds.), *Truth and Meaning: Essays in Semantics*, Clarendon, Oxford, pp. 263-84.

Thomas Aquinas: *De veritate*, Leonine edition, 1970, XXII.

Vision, G.: 1988, *Modern Anti-Realism and Manufactured Truth*, Routledge, London & New York.

Vendler, Z.: 1967, *Linguistics in Philosophy*, Cornell University Press, Ithaca.

Walker, R. C. S.: 1989: *The Coherence Theory of Truth: Realism, Anti-Realism, and Idealism*, Routledge, London and New York.

William of Sherwood, *Introduction to Logic*, translated by N. Kretzman, University of Minnesota Press, Minneapolis, 1966.

Williams, C. J. F.: 1976, *What is Truth?*, Cambridge University Press, Cambridge.

Williams, M.: 1986, 'Do We (Epistemologists) Need a Theory of Truth?', *Philosophical Topics* **14**, 223-42.

Wittgenstein, L.: 1921, *Logisch Philosophische Ahandlung: Tractatus Logico-Philosophicus*, in *Annalen der Naturphilosophie*; reprinted in *Werkausgabe*, Band 1, Suhrkamp, Frankfurt am Main 1984; English translation by D. F. Pears and B. F. McGuinness, Routledge, London, 1961.

Woleński, J.: 1994, 'Contributions to the History of the Classical Truth-Definition', in D. Prawitz, B. Skyrms, and D. Westerståhl (eds.), *Logic, Methodology and Philosophy of Science*, IX, Elsevier Science B. V., pp. 481-95.

Woleński, J.: 1994a, 'A Controversy over the Concept of Correspondence (Bradley, Joachim, Russell)', in J. Hintikka and K. Puhl (eds.), *The British Tradition in 20th Century Philosophy*, Austrian Ludwig Wittgenstein Society, pp. 537-43.

Woodfield, A. (ed.): 1982, *Thought and Object: Essays on Intentionality*, Clarendon Press, Oxford.

Wright, C.: 1992, *Truth and Objectivity*, Harvard University Press, Cambridge, Mass.

SUSAN HAACK

REALISM

INTRODUCTION

'Realism' refers, not to a single, simple thesis, but to a whole family of positions; and so it contrasts, not with a single, simple, opposing thesis, but with another whole family of non-realist positions – idealism, nominalism, instrumentalism, relativism, irrealism, etc., etc.

REALISMS	NON-REALISMS
perceptual realism	representative theory of perception
physicalism, dualism, neutral monism	idealisms: subjective, theological, objective
truth as possibly outrunning us	pragmatic maxim
truth-condition theory of meaning	Verification Principle
	assertability-condition theory of meaning
realism about universals, the reality of generals	nominalism
scientific realisms:	
theoretical statements as genuine statements	instrumentalism
truth as the goal of science	constructive empirism
cumulative realism	
convergent realism	'the natural ontological attitude'
optimistic realism	
explanatory realism	
dual-level realism	social constructivism
metaphysical realism	relativity of truth to theory
internal realism (??? ——>)	cultural relativism
	Goodmanian irrealism
	conceptual relativity
innocent realism	

TABLE 1:Varieties of Realism and Non-Realism

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 415-436.
© 2004 *Kluwer Academic Publishers.*

Very roughly, the common theme that unites the many and various members of the realist family is that something – the world, truth, universals, numbers, moral values, etc., etc. – is independent of human beings and their beliefs, concepts, cultures, theories, or whatever. What distinguishes the different members of the realist family from each other is exactly what it is that each holds to be independent, in exactly what way, of exactly what about us.

Roughly, again, the key issue in the many and various disputes between realists and their non-realist opponents is how the world/truth/universals/etc. can be at once independent of us, and yet knowable by us. Those disputes all focus in one way or another on how much of what we know of the world is properly thought of as the world's contribution and how much as our contribution, on where the line runs between what we discover and what we construct.

Realists, stressing discovery, the world's contribution, have often succumbed to the temptation to articulate the independence of the world/truth/etc. in ways that, compromising the accessibility of reality to our knowledge, lead to scepticism. Non-realists, stressing construction, our contribution, have often succumbed to the temptation to articulate the accessibility of the world/truth/etc. in ways that compromise the independence of the reality we can sometimes, partially and fallibly, come to know.

Since a key concern is the balance of independence and accessibility, the integration of metaphysics and epistemology, one quite good way to begin is with realist versus non-realist theories of perception.

PERCEPTUAL REALISM VERSUS THE REPRESENTATIVE THEORY OF PERCEPTION

Ironically enough, Locke, the founder of empiricism, seems to have taken for granted the Cartesian principle that what is most immediately accessible to a subject is the contents of his own mind; and so assumed that the immediate objects of perception are ideas, mental images, not the physical things and events of which these ideas are representations. Empiricism thus set off on the 'way of ideas', a path which led inevitably to the question: how, if what we perceive is ideas, can we ever know the objects we take them to represent? – and then, after Locke's unsuccessful struggles to establish the reality of perceptual ideas, to Berkeleian idealism and Humean scepticism.

The realist position that contrasts with the representative theory of perception is perceptual realism, anticipated in some passages of Book IV of Locke's *Essay* incompatible with the representationalist thrust of Book II, and articulated in Thomas Reid, sharp critic of the way of ideas; as, also, in C. S. Peirce, who (though his critical common-sensism combines elements of Kant's with this element of Reid's response to Hume) writes that 'it is the external world that we directly observe';[1] and, in our times, perhaps most articulately defended by psychologist J.J. Gibson.

Though perceptual realism has sometimes been, and is often described by its opponents as, 'naïve', there is no need for a perceptual realist to hold that our perceptions are infallible. The essential point, to put it as Gibson might, is that we interact with the world by means of sensory organs which are competent to detect (some of the) information afforded by the things and events around us; which by no means implies that our senses, or our perceptual judgments, are perfect. Thus modestly construed, perceptual realism can acknowledge that all our perceptual judgments depend to some extent on background beliefs as well as on sensory input, explaining our susceptibility to perceptual illusions and mistakes; and so can accommodate the considerations that lead Richard Gregory, for example, to hold that we must construe perception as entirely inferential, as 'hypothesis'.

From the perspective of the empiricists' representative theory, however, physical things and events lie behind a 'veil of perception', and their very existence appears to be a large and doubtfully justifiable inference from the ideas of which we are immediately aware. But, as Berkeley realized, the inference would not be large or doubtful if those physical things and events were reconstrued as consisting, simply, of collections of ideas. Berkeley grounded the continuing existence of material objects in the mind of God. Mill resorts instead to a subjunctive analysis in terms of the perceptions an observer would have if he were present, construing physical objects as 'permanent possibilities of sensation'; and Russell, armed with the apparatus of modern logic, construes them as logical constructions out of sense-data. Hence (in Berkeley's case) the connection of the representative theory of perception with theological idealism, and (in more recent manifestations) of a sense-datum theory of perception with a phenomenalist, subjective idealism.

REALISM VERSUS IDEALISM, ETC.

An idealist holds that everything there is, is mental: that the world is a construction out of our, or, in the case of the solipsist, his own, ideas – subjective idealism; or is constituted by God's ideas – theological idealism; or that the world is itself of a mental or spiritual character – objective idealism, as in Hegel. The realist, denying this, affirms that not everything is mental.

So there are several distinct forms of realism-as-opposed-to-idealism. A physicalist denies that everything is mental because he maintains that everything is physical; and so, if he is not an eliminativist like Feyerabend, Rorty or Churchland, but admits that there are mental states, he construes those mental states as reducible to something physical – whether, like Quine at his most Skinnerian, to dispositions to overt behaviour, or, like Smart or Armstrong, to states of the brain or central nervous system (or, like less strictly Skinnerian time-slices of Quine, he may take dispositions to behave to *be* neurophysiological states).

Dualists deny that everything is physical because they maintain that, besides physical objects and events, there are irreducibly mental states and processes; some holding that the mental and the physical interact, others that there is rather a pre-established harmony between the two, epiphenomenalists that causation goes only one way, etc.[2] Popper and Eccles are perhaps the best contemporary examples of interactionist, Cartesian dualists – though there is a case for categorizing Popper as a pluralist rather than simply as a dualist, since he also acknowledges a quasi-Fregean third realm of abstract objects such as numbers, propositions, problems and theories.

Neutral monists deny that everything is physical because they maintain that everything is constituted of neutral stuff itself neither mental nor physical: stuff called, rather misleadingly, 'experience' in James's radical empiricism, the precursor of Russell's development of the classical neutral-monist position. (How is this compatible with James's *Pluralistic Universe*? – his is a 'mosaic' philosophy according to which there are many instances of this one kind of stuff.)

Peirce, as so often, resists easy classification. His panpsychism, expressed in that far-from-transparent observation that 'matter is effete mind',[3] might perhaps be interpreted as less Hegelian than it initially appears, as intimately related to his agapism, the thesis that the universe is gradually evolving from an initial chaos into increasingly lawful orderliness. All the same, his position might perhaps be characterised, using a label suggested by his father (the Harvard mathematician Benjamin Peirce), as 'ideal-realism'. For Peirce holds that the real, though independent of how you or I or anybody takes it to be, is the object of the hypothetical final representation, the ultimate opinion that would be agreed were inquiry to continue long enough.

The relation of pragmatism to the various kinds and styles of realism and non-realism is as subtle and complex as this double-aspect character of Peirce's conception of reality already suggests; more so, in fact, for there are relevant disagreements within pragmatism, not to mention relevant shifts and ambiguities in the thinking of various individual pragmatists.

PRAGMATISM AND REALISM

When Peirce describes his form of pragmatism, pragmaticism, as 'prope-positivism' the 'prope-' indicates that, unlike positivism, pragmatism does not eschew metaphysics (Peirce had in mind the positivism of Comte; but there is the same difference between pragmatism and the later, logical, positivism).

At the heart of Peirce's pragmaticism, intended as a means both of clarifying hard concepts and of filtering out the pragmatically meaningless, is the pragmatic maxim: the meaning of a concept is given by its potential pragmatic, i.e., *pragmatische*, experiential, consequences. The maxim will, Peirce thinks, reveal much of traditional metaphysics to be pragmatically meaningless, 'gibberish' as he once puts it; but 'all such rubbish being swept away, what will remain of philosophy will be a series of problems capable of investigation ... by the true sciences'. He envisages a purified, scientific, metaphysics, conducted with the genuine truth-seeking attitude, and, like natural-scientific inquiry, using the method of experience and reasoning. So 'instead of merely jeering at metaphysics' the pragmaticist 'extracts from it a precious essence'.[4]

Peirce develops a whole range of metaphysical theories including, besides the panpsychism and agapism mentioned earlier, his categories, his synechism, and his tychism (the thesis that there is real chance in the universe). And unlike Carnap, for example, who dismisses the issue of nominalism versus realism as a pseudo-question, Peirce comments that this traditional metaphysical question is 'still as pressing as ever it was'[5] – and comes down firmly, if obliquely, on the realist side.

Truth is one of those hard concepts the meaning of which the pragmatic maxim should illuminate. So Peirce is led to his account of truth as the opinion that would be agreed were inquiry to be continued indefinitely – the ultimate opinion referred to earlier. The truth, Peirce writes, 'is SO, whether you or I or anybody thinks it is so or not'; it is not, however, independent of the hypothetical agreement of the community of inquirers were inquiry indefinitely pursued.

The price of thus ensuring the accessibility of truth at least appears to be some compromise of independence – as is revealed in Peirce's discussion of what he calls 'the problem of buried secrets'. Propositions about the past that would not be settled however far inquiry were to be pushed must, apparently, be deemed neither true nor false. We are too quick to prejudge what future inquirers might be able to find out, Peirce replies, and at the same time too quick to presume that propositions that really would never be settled are nevertheless meaningful. Since tychism implies that some information about past events is bound to be irrecoverably lost, it seems Peirce must rely on denying pragmatic meaning to at least some statements about past events – and hence is committed to repudiating the realist intuition that any well-formed, linguistically meaningful statement is either true or else false, that 'There were exactly n dinosaurs', for example, is true for one value of n and false for others even if inquirers could never find out how many dinosaurs there were.

Whether or not Peirce's account of truth may be [re-]interpreted, as some scholars have proposed, so as to avoid this apparent compromise of independence, James, who writes that 'the trail of the human serpent is ... over everything',[6] is surely less realist than he. Unlike Peirce, James thinks of 'pragmatism' as deriving from *praxis*, 'action', is not so careful to express the pragmatic maxim subjunctively, and seems temperamentally leery of abstractions.

James distinguishes abstract Truth and concrete truths; and he characterizes abstract Truth, in a manner at least vaguely reminiscent of Peirce, as the opinion on which our present formulations will eventually converge. But, though he acknowledges that abstract Truth is the prior concept, he urges that the pragmatist focus on specific, concrete truths: propositions not merely verifi*able*, but verifi*ed*, made true; and when he says that what is true today may turn out to be false tomorrow, reveals that he is using 'verified' to mean no more than 'confirmed' – so that concrete 'truths' may not be true at all.

So perhaps it is understandable why Russell feared that pragmatism would lead to 'cosmic impiety', or at least to fascism; and was inclined, like many critics, to read James's comment that 'the true is only the expedient in the way of belief' as saying that whatever satisfies the believer to believe, is true. But it is often forgotten that James continued: 'expedient in the long run and on the whole of course; for experience has ways of *boiling over*, and making us correct our present formulas'.[7] On the tangle of questions about independence and accessibility that concerns us here, perhaps the best assessment is, not that James is thoroughly non-realist, but that he is thoroughly ambivalent.

The same might be said of Dewey, who, though he once describes Peirce's as 'the best definition of truth',[8] is more comfortable working with the concept of warranted assertibility. But it surely could not be said of Schiller, who, apparently reading James's remarks about concrete truths as if they constituted – as James himself recognised they could not constitute – a complete account of truth, is a relativist who aligns himself unabashedly with Protagoras. Observing how the etymology of 'fact' (from the past participle of the Latin verb *facere*) relates it to what is made or done, Schiller relativizes truth to human purposes and values.

Self-styled neo-pragmatist Richard Rorty, often accused of relativism, denies the charge. The issue is far from straightforward; for in epistemology, at least, Rorty seems to shift from a contextualist view which does qualify as relativist ('S is justified in believing that p iff S satisfies the epistemic standards of his community'), to a version better characterized as tribalist ('S is justified in believing that p iff S satisfies the epistemic standards of our community').

Relativist or not, on the question of truth Rorty sometimes sounds very radical indeed. Pragmatism holds, he tells us, that truth is not the kind of thing we should expect to have an interesting theory about.[9] (One's first reaction may be to protest that Peirce certainly *has* an interesting theory of truth; one's second, perhaps, that Rorty's remark is somewhat reminiscent of James's urging us to focus on concrete truths rather than abstract Truth.) A true statement, Rorty says, is just one you can defend against all conversational objections; to call a statement true is just to say that it is a statement we can agree about.[10] (This sounds a lot like the result of stripping Peirce's characterization of truth of everything that ties it to the world.) There are two senses of 'true', Rorty tells us, the homely sense – of which he approves – in which it just means 'what you can defend against all comers', and the Philosophical sense – of which he disapproves – in which it is designed precisely to stand for the Ideas of the Unconditioned.[11] (Setting off from this false dichotomy, Rorty foists his homely sense of 'true' – which needless to say is not a sense of 'true' at all – onto Tarski and Davidson.) Anyhow, Rorty tells us, he doesn't 'have much use for notions like ... "objective truth"';[12] to call a statement 'true' is merely to give it 'a rhetorical pat on the back'.[13]

But one also finds in Rorty statements to the effect that the pragmatist should not take a position on the question of realism, that this is a question better repudiated than answered. And so, at times, Rorty sounds a lot like an old logical positivist urging that the traditional territory of metaphysics be abandoned and not re-occupied.

POSITIVISM AND REALISM

Unlike the classical pragmatists, logical positivists/logical empiricists are committed to denying the intelligibility of metaphysical questions, the issues of realism versus non-realism among them. Thus Schlick, repudiating the question of the reality of the external world, writes that 'The world of the non-metaphysician is the same world as that of all other men ... . He merely avoids adding meaningless statements to his description of the world'. And again: 'The empiricist does not say to the metaphysician "what you say is false", but "what you say asserts nothing at all!"'.[14] And Carnap, pointing out that we can deduce no perceptual statements either from the assertion of the reality of the physical world or from the opposite assertion, writes that 'both have no empirical content – no sense at all'.[15]

But those supposedly illegitimate metaphysical questions often seem to sneak back in under the guise of questions about language. Carnap, for example, officially denies that such 'framework questions' as 'are there physical objects?', 'are there abstract propositions and properties?', have true or false answers; they are pragmatic questions to be decided by reference to the relative convenience of this or that linguistic framework. And yet the quantification over propositions required by Carnap's semantics, or his choice of a basis of 'elementary experiences' and set-theoretical tools for the 'logical construction of the world', certainly look metaphysically consequential (indeed, Goodman criticises Carnap's construction in part because of what he perceives as its objectionably platonist character).

A further complication is that, as with pragmatism, there are shifts and ambiguities in the thinking of various individual positivists, and relevant disagreements within positivism – internecine disputes about, for example, the relative merits of correspondence and coherence theories of truth. Still, a commitment to the Verification Principle is shared; and, as with the pragmatic maxim, this does have some tendency to pull against a full-blooded realism about truth.

Notoriously, the Verification Principle – a statement is empirically meaningful if and only if it is verifiable by experience – is multiply ambiguous: the 'can' implicit in 'verifi*able*' may be read as 'can in principle' or as 'can in practice'; and verifiability may be taken as requiring that the statement can be conclusively shown true or false, or only that it can be confirmed or disconfirmed, shown more or less probable, by experience.

With the more restrictive interpretations of the VP, some logical positivists were drawn to phenomenalism, to operationalism, or to instrumentalism with respect to scientific theories – the fear motivating instrumentalism being that, if acknowledged as genuine statements, scientific theories would be in danger of being ruled unverifiable, and hence empirically meaningless. As more hospitable readings of the VP emerged, however, logical positivism sometimes took on rather the aspect of, as Feigl puts it, an empirical or critical realism.[16]

Nevertheless, however it is interpreted, the VP makes truth and falsity essentially accessible. A statement cannot be true or false unless it is meaningful. So, if a statement is meaningful if and only if it is verifiable, a statement cannot be true or false unless it is verifiable. The VP thus precludes the possibility that truth might outrun us, that there could be truths or falsehoods inaccessible to us.

Though the heyday of logical positivism is now long past, there are echoes of some of its themes in debates in (fairly) recent philosophy of language about the relative merits of truth-condition versus assertibility-condition theories of meaning.

TRUTH-CONDITION VERSUS ASSERTIBILITY-CONDITION THEORIES OF MEANING

Dummett's championship of, as he variously calls it, an 'anti-realist', 'idealist' or 'constructivist' theory of meaning seems to be most directly motivated, however, less by positivist sympathies than by the influence of mathematical intuitionism. The meaning of non-mathematical statements lies in the conditions of their warranted assertibility, as, according to the intuitionist, the meaning of mathematical statements lies in the conditions of their provability; and Dummett suggests we might think of reality more generally, as an intuitionist thinks of mathematical reality, not as independently pre-existing, but as coming into existence as we probe.

Dummett contrasts his conception of meaning as assertibility-conditions with a realism which construes meaning as truth-conditions. So in this context, 'realism versus anti-realism' means, in effect, 'Davidson versus Dummett'.

Davidson's identification of the meaning of a sentence with its truth-conditions – the conditions in which, were they to obtain, the sentence would be true – is in the tradition of Wittgenstein's *Tractatus*, but distinguished by a reliance on Tarski's theory of truth to articulate the structure of truth-conditions and hence meaning.

From the perspective of what came to be called 'the Davidson programme', the main issue was how to extrapolate Tarskian methods, apparently applicable only to languages which are formally specifiable, to natural languages; hence the preoccupation with quirks of natural languages at least *prima facie* unsuitable for Tarskian treatment, such as adverbs, attributive adjectives, *oratio obliqua* and the like. (Tarski himself judged this project impossible, and Davidson seems more recently to have come to have doubts about its feasibility himself).

From the perspective of anti-realism about meaning, however, the main issue appears, rather, as the potential inaccessibility of truth-conditions. A truth-condition theory, as an anti-realist sees it, makes a mystery of how we are able to learn and understand language. For it implies that to understand a sentence is to understand the conditions in which, if they obtained, the sentence would be true; but these are conditions for a 'transcendent' truth-value, the conditions in which the sentence would be true, regardless of whether or not we knew them to obtain. But when we learn a language, the anti-realist argues, we learn to assert or assent to sentences in circumstances under which we are warranted in asserting them. If there were sentences the meanings of which involve conditions we could never know to obtain, we could not learn their meaning, nor understand them.

Dummettian anti-realists are apt to have doubts about bivalence; as intuitionist logicians do (and as some have thought the problem of buried secrets might lead a pragmatist to do). As this reveals, disputes between assertibility-condition and truth-condition theories echo older questions raised by pragmatist and verificationist approaches to meaning.

Similarly, contemporary disputes about the semantics of natural-kind terms echo older disputes between nominalists and realists.

REALISM VERSUS NOMINALISM

The nominalist (the term comes from the Latin word for 'name' or 'word') holds that unlike particular, individual things and events such as this rose or that explosion, universals – properties and relations such as redness or loudness or resemblance – are not in the world independent of us and our vocabularies or conceptual schemes; generality lies in our words or in our concepts. The realist, denying this, affirms that there are universals independent of us.

One form of realism-about-universals is platonism, which conceives of universals as abstract particulars, neither mental nor physical, somehow instantiated by individual things and events. Another, sometimes called 'moderate realism', is rather Aristotelian than Platonic, conceiving of universals not as separate abstractions, but as somehow realized in particulars.

A third and quite distinctive form of realism-about-universals, articulated by Peirce (who seems as a young man to have inclined towards nominalism, but later, under the influence of Scotus, came to describe himself as 'a scholastic realist of a somewhat extreme stripe'),[17] maintains that there are real generals, but denies that generals exist. Peirce presses the adjective 'general' into service as a noun referring to natural kinds/laws/potentialities; and distinguishes the real, characterized by its independence of how you or I or anybody thinks it to be, from the existent, characterized by its capacity to interact (a key difference between nominalists and realists, he holds, is that they have different conceptions of the real). Whatever exists is real, but not vice versa; and Peirce can distinguish his, as he believes, genuine realism from the view that universals are existent, abstract, particulars – 'nominalistic platonism', he unkindly calls it. *What* generals are real, Peirce adds, noting that here he departs from Scotus, is not to be settled simply by reference to language, but is an empirical question for the sciences to settle. Unless there were kinds and laws independent of how we think them to be, he argues, prediction, induction, explanation, science itself, would be impossible.

In our times the old issues of nominalism versus realism have sometimes played out, not straightforwardly as a dispute about the ontological status of universals, but rather as a dispute about the status of abstract objects, entities neither mental nor physical – as with Church's semantic platonism, or Field's mathematical nominalism. And sometimes, as with Goodman, the issue has come to focus on the legitimacy or otherwise of sets; or, as with Quine, on the legitimacy or otherwise of intensional objects.

And where the issue is more directly focussed on the status of universals, properties, kinds, much contemporary debate proceeds as if on the unstated assumption that, as Peirce might have put it, nominalism and nominalistic platonism exhaust the alternatives. Peirce would, for example, doubtless categorize as 'nominalistic platonism' the contemporary style of realism about natural kinds associated with Kripke's and Putnam's construal of natural kind terms as rigid designators, which treats terms like 'gold' or 'tiger' as working, semantically, as they believe proper names work, purely denotatively.

Quinean realism about natural kinds, though very different from this Kripke-Putnam style of realism, is informed by the presumption that multiplication of senses of 'to be' should be avoided, and that the quantifiers must be interpreted objectually. This exerts a strong nominalizing influence, so that the 'there are' of 'there are natural kinds and similarities', must be construed as the same 'there are, there exist' as in 'there are tigers in India'. So, though Quine shares with Peirce the idea that the possibility of successful induction depends on there being natural kinds and similarities, he too is rather a nominalistic platonist than a real realist in Peirce's sense.

Lewis's realism-about-possibilities, though very different from Quinean or Kripkean realism, is again recognizably of the nominalist-platonist stripe: the intelligibility of subjunctive conditionals, Lewis holds, requires the existence of abstract particulars in the form of possible worlds and their possible inhabitants – possible worlds construed as things of the same kind as the actual world.

Armstrong acknowledges universals as independent of the mind; furthermore, he holds that what universals there are is a matter for scientific discovery, and sees this *a posteriori* realism as playing a key role in the explication of the notions of causation and nomic connection. He always writes of universals as 'existing', as 'entities', but since he holds that universals are *in* particulars, perhaps his view might be best categorized as nominalistic aristotelianism.

The question of the reality of universals, as previous paragraphs indicate, has by now come to focus most specifically on the entities, kinds, and stuff of scientific theories, and old debates between nominalists and realists are reprised within the philosophy of science.

SCIENTIFIC REALISM VERSUS INSTRUMENTALISM, ETC.

But in philosophy of science as elsewhere, 'realism' is multiply ambiguous. In one of its several uses in this context, it contrasts with instrumentalism regarding scientific theories. According to the instrumentalist view, theoretical terms are not really referential, and what appear to be theoretical statements are not really statements at all, and so are neither true nor false; they are instruments, tools for predicting particular observational consequences – sometimes likened to an abacus, or construed as disguised inference-rules.

The instrumentalist position presupposes a distinction of observational versus theoretical statements; so reservations about the robustness of that distinction are one motivation for realism-as-opposed-to-instrumentalism – the thesis that theoretical statements in science are genuine, referential, true or false statements.

The position van Fraassen calls 'constructive empiricism', according to which the goal of science is, not true theories, but empirically adequate theories, theories with true observational consequences, though quite distinct from instrumentalism – since it acknowledges scientific theories to be, as they appear to be, literally true or false statements about the world – nevertheless has some affinities with the instrumentalist branch of the non-realist family. In the sense in which it contrasts with constructive empiricism, 'realism' refers to the idea that the goal of science is, not just empirical adequacy, but truth.

Declaring realism dead, citing Mach, sometimes sounding a bit like Schlick holding metaphysical questions at arms' length, Fine proposes, not a non-realism on a par with Van Fraassen's constructive empiricism, but what he calls the 'natural ontological attitude': we should accept the results of science as true, and scientific statements as referential, with whatever degree of confidence the evidence permits, but without 'the progressivism that seems inherent in realism'.[18]

As this reveals, realism with respect to science has sometimes been construed as requiring not only (as against instrumentalism) that scientific theories be genuine, true or false statements which, when true, refer to real theoretical entities, and (as against constructive empiricism) that the goal of science is to discover true theories, but also that science progresses by the accumulation of true theories – cumulative realism. Or sometimes it has been construed as involving the thesis that, as science proceeds, its theories gradually get closer to the truth – convergent realism. Sometimes, again, scientific realism is construed more strongly yet, as committed to the thesis that currently accepted theories in mature sciences are true, or at least approximately true – optimistic realism; and sometimes to the thesis that only by reference to the truth of these theories is it possible to explain the success of science – explanatory realism. Each of these stronger variants of scientific realism presupposes the weaker thesis required by scientific realism as defined by the contrast with instrumentalism, that scientific theories are *bona fide*, true or false, statements.

Those who urge that the entities referred to in true scientific theories are real usually have in mind the theoretical entities of the natural sciences – genes, electrons, quarks and so forth. But the question of the status of theoretical entities in the social sciences is less straightforward. Social institutions such as money or marriage don't seem to be independent of human beings and their beliefs, intentions, etc., in the same way as quarks, electrons, DNA, and the like. Not that they are not real; but their reality is, to borrow a phrase from Searle, less brute than the reality of natural objects, for they are in part constituted by shared beliefs and intentions. We don't physically construct social institutions, as we do highways and skyscrapers; nevertheless, they are, in a different way, partially dependent on us.

The previous paragraph articulates what might be called a dual-level realism according to which natural objects and events are real in a stronger sense, social institutions only in a weaker. This should not be confused with the social constructivism of those who, explicitly or implicitly, deny the reality of theoretical entities in the social and the natural sciences alike. A scientific realist can acknowledge that theoretical concepts such as *gene*, *electron*, etc., are the creation of the scientists who devise them; that scientific theories come to be accepted or rejected by means of a complex social process within the relevant scientific sub-community; that sometimes what natural scientists describe are not so much natural as laboratory phenomena. A dual-level scientific realist can acknowledge that shared beliefs, etc., are partially constitutive of social institutions. No scientific realist, however, will grant, as some radical social constructivists appear to think, that the entities referred to in true scientific theories, natural and social, are brought into existence by the intellectual activity of scientists, let alone that they are created by 'social negotiation' within the scientific community.

As this last phrase hints, of late not a few sociologists and some philosophers of science have come to think of science as a cultural product like systems of religious belief or myth; and as a result another kind of non-realism, this time relativist or radically irrealist rather than instrumentalist in character, has begun to be newly influential.

REALISM VERSUS RELATIVISM, ETC.

Here things are complicated by the fact that 'relativism' itself refers, not to a single, simple thesis but to a whole family, each holding that something (truth, reality, moral values, and so forth) is relative, in some sense, to something else (language, theory, scientific paradigm, culture, and so on). For present purposes, the most relevant members of the relativist family are those that relativize truth and/or reality to language, theory, paradigm, or conceptual scheme.

...IS RELATIVE TO...

(1) meaning	(a) language
(2) reference	(b) conceptual scheme
(3) truth	(c) theory
(4) metaphysical commitment	(d) scientific paradigm
(5) ontology	(e) version, depiction, description
(6) reality	(f) culture
(7) epistemic values	(g) community
(8) moral values	(h) individual
(9) aesthetic values	•
•	•
•	•
•	

TABLE 2: Varieties of Relativism

Some forms of relativism are self-undermining; in consequence it is sometimes supposed that all must be, and that to show that a position is relativist is to refute it. But this over-simplifies. Tarski, for example, construes truth as language-relative; but there is nothing self-undermining about his position. According to Tarski, it must be sentences, or more strictly speaking well-formed formulae of formal languages, which are the bearers of truth and falsity; for only such linguistic items possess the syntactic structure which his truth-definition exploits. But sentences or wffs are sentences or wffs of some specific language; a string of symbols which is true in one language could be false or, more likely, meaningless, in another. For this reason – and also because his solution to the Liar Paradox requires a hierarchy of object-language, meta-language, etc. – Tarski defines, not 'true', but 'true-in-L'.

Whether or not it is true that 'true' is language-relative, this claim can be applied to itself without embarrassment. Granted, insofar as it takes the bearers of truth and falsity to be, not eternally existing propositions which sentences merely express, but linguistic items which presumably would not exist unless there were human beings and human languages, Tarski's approach is slightly to the left of the most stringent forms of realism with respect to truth (though, as Tarski points out, it is neutral with respect to other forms of realism). Only slightly to the left, however; as Tarski defines it, truth is independent of theory, belief, culture, etc. – indeed, part of his motivation was precisely to give a definition which makes it clear that truth is a concept quite distinct from provability-in-theory-T.

Unlike Tarski's thesis that it makes no sense to describe a sentence as true except relative to some language or other, the thesis that it makes no sense to describe a statement as true except relative to a theory really is self-undermining; at least if it is asserted as true-full-stop and not merely as true-in-theory-T. The form of realism with respect to truth that contrasts with this is, of course, the thesis that 'true' is not theory-relative – and that the question, 'is theory T true?' makes perfectly good sense.

Realists who deny the relativization of truth to theory will likely also deny cultural relativism, which relativizes truth to culture (a term sometimes used, of late, in a very extended sense which treats race or gender as cultures). What is accepted as true certainly varies from culture to culture – at least in the ordinary sense of that term; but the cultural relativist goes further, claiming that it makes no sense to describe a statement as true except relative to this or that culture. This too is self-undermining, at least if it is asserted as true-full-stop and not just as true-in-culture-C.

Goodman is even more radical, relativizing not just truth but reality, so that his position is perhaps best described as 'irrealism'. There is no one real world, Goodman holds, only many 'versions', the descriptions or depictions made by scientists, novelists, artists, etc. Versions of what? – he doesn't say.

This radical Goodmanian irrealism has some affinity with radical forms of social constructivism which think of theoretical entities as created by scientists' intellectual activity, and with radical styles of rhetoric of science which think of theoretical entities as a kind of illusion created by scientists' rhetorical activity. The affinity is sometimes disguised, however, by a twist of terminology: Gross, for instance, calls the thesis that scientists cannot but believe that the entities they describe are real, 'psychological realism', and then compounds the confusion by likening his approach to Goodmanian irrealism.[19]

The 'metaphysical realism' to which Putnam once subscribed is opposed to both cultural relativism and Goodmanian irrealism: there is one real world, consisting of a fixed totality of mind-independent objects, and one true description of this one real world, a description couched in a privileged, 'scientific' language. This is a very strong form of realism-as-opposed-to-relativism, stronger than required simply by the repudiation of cultural relativism and irrealism (and reminiscent of the old logical atomist metaphysics and philosophy of language); so perhaps it is not surprising to find Putnam retreating first to an 'internal realism' – arguably, perhaps, a misnomer – reminiscent of a moderately radical style of pragmatism, and then to the thesis of conceptual relativity.

According to this thesis there is one real world, but it does not consist of a fixed totality of mind-independent objects. The question, how many and what kind of objects there are, makes sense only relative to a conceptual scheme; there is no absolute, privileged, scientific vocabulary which describes the world as it is independent of our conceptual contributions. And truth is a matter neither of a description's copying or corresponding to the mind-independent objects in the world, nor of its being accepted in this or that community. It is a matter, rather, of the description's being such that, in epistemically ideal circumstances, we would be justified in accepting it. (This is further from Peirce's conception of truth than it sounds; for Putnam construes ideal justification as context-dependent, which precludes the possibility of that unique final representation envisaged by Peirce.)

Those of more realist inclination who are puzzled by what it could mean to say, 'relative to conceptual scheme C_1 there are rocks, but relative to conceptual scheme C_2 there aren't', may not be reassured when Putnam observes that to acknowledge conceptual relativity is only to admit that 'you can't describe the world without describing it' – which sounds suspiciously like a tautology; or that it is to admit that incompatible descriptions of the world can be both true – which sounds suspiciously like a contradiction.[20] Indeed, Putnam himself seems to have felt this realist pull, recently urging that, in the spirit of a 'second naïveté', we explore what habitable middle ground there may be between metaphysical realism and conceptual relativity.

The position I call 'innocent realism', I believe, identifies just such habitable middle ground between rigid realism and rakish relativism.

INNOCENT REALISM

The world – the one real world – is largely independent of us. Only 'largely', not 'completely' independent of us; human beings intervene in the world in various ways, and human beings, and their physical and mental activities, are themselves part of the world.

Natural things and events are entirely independent of us, but our shared beliefs and intentions are partially constitutive of social institutions such as law, the economy, government, religion, etc., etc. The world we humans inhabit is not brute nature, but nature modified and reconstituted by our physical activities and overlaid by the labyrinth of signs, the semiotic webs that constitute our social worlds, the theories of scientists, and the imaginative constructions of novelists and artists.

Though only fallibly and imperfectly, we humans are able to know something of how the world is. This is possible only because we have sense organs competent to detect information about things and events around us; and because the particular things and events in the world of which we can be perceptually aware are of kinds and are subject to laws, so that we can sometimes categorize things into real kinds, discern their inner constitution, and discover laws of nature. And that is how natural-scientific knowledge – deeper and more detailed, better unified, more accurate, yet still thoroughly fallible, imperfect and incomplete – could grow out of our common-sense knowledge of the world. (These thoughts, prompted by Peirce's arguments for the reality of generals, also bear some affinity to Bhaskar's 'transcendental realism'.)

Our sensory organs enable us to detect some of the information afforded by things around us. Our evidence with respect to any claim about the world – the evidence of our senses, plus relevant background beliefs (our reasons) – though always imperfect, is better or worse depending on how supportive it is, how much of the relevant evidence it includes, and how secure our reasons are. Our perceptual apparatus is imperfect and limited, and our judgments of what we perceive – usually influenced by background beliefs, expectation, set – are fallible. And our judgments of the relevance, supportiveness, comprehensiveness of evidence are inevitably perspectival, dependent on background beliefs themselves fallible. This dependence on perspective creates a kind of illusion of incommensurability, an illusion which may have encouraged epistemic relativism. Though our judgments of evidential quality are perspectival, however, evidential quality is not perspectival or relative, but objective.

We humans describe the world, sometimes truly, sometimes falsely. Whether a (synthetic) description of the world is true or is false depends on what it says, and on whether the world is as it says. What such a description says depends on our linguistic conventions; but, given what it says, whether it is true or it is false depends on how the world is. We learn language by learning to assert/to assent to sentences and sentence-fragments in the circumstances in which they are assertible; but our languages permit the construction of linguistically meaningful sentences – grammatically correct strings of meaningful words – of which we are unable, and perhaps would be unable however long inquiry continued, to determine the truth-value.

To say that a statement is true is to say that things are as it says. Some descriptions describe us, and some describe things in the world that depend on us; and whether such a description is true or is false depends on how we are, or how those things that depend on us are – for such descriptions, those are the relevant things, the relevant aspects of 'how the world is'. But whether even such a description is true or is false does not depend on how you or I or anybody *thinks* the world is.

We can describe how the world is, or would have been, if there were, or had been, no human beings. Before there were human beings or human languages, the English sentence 'There are rocks' did not exist; and so, if sentences are bearers of truth and falsity, it is not the case that 'There are rocks' was true before there were people, or that 'There are rocks' would have been true even if there had never been people. Nevertheless, there were rocks before there were people, and there would have been rocks even if there had never been people; and that is a (partial) description of how the world would be, or would have been, if there were, or had been, no human beings.

There are many different vocabularies, and many different true descriptions of the world. Though pieces of furniture are physical objects, descriptions of sofas, armchairs and tables cannot be rendered without loss into the vocabulary of physics; for the relevant descriptions will need to refer to the functions of sofas for sitting, tables for writing or eating. And, though a person's beliefs, hopes and fears are complex, federal, multiform dispositions to verbal and non-verbal behaviour which are somehow neurophysiologically realized, descriptions of what someone believes, etc., cannot be rendered without loss into the vocabulary of physics either; for the relevant descriptions will need to refer to the things in the world the beliefs are about, and to the linguistic communities to whose patterns of speech the person's verbal behavior belongs.

Two descriptions in different vocabularies may say the same thing about how (some part or aspect) of the world is, or different things. If they same the same thing, they are of course compatible with each other; if they say different things, they may be compatible or incompatible with each other. Compatible descriptions may be combined in a longer, conjunctive description, which will be true just in case its conjuncts are; incompatible descriptions, however, cannot be jointly true.

There are many different truths about the world. All these many different truths must, somehow, fit together; there can't be rival, incompatible truths or 'knowledges'. But this doesn't mean that all the truths about the world must fit together by being reducible to a privileged class of truths expressed in a privileged vocabulary (that they must be 'unified' in the logical positivists' strong sense of that term). A better analogy would be the way a road map can be superimposed upon a contour map of the same territory.

The tension between independence and accessibility, the innocent realist believes, is not so severe as is sometimes supposed – in fact, it may be altogether superable if our understanding of independence is modest enough and our understanding of accessibility fallibilist enough.

Susan Haack
Unversity of Miami

NOTES

[1] Peirce, *Collected Papers*, 8.144.
[2] See Chisholm's marvellous cartoon of the various positions in philosophy of mind in Taylor, *Metaphysics*, p.17.
[3] Peirce, *Collected Papers*, 6.25.
[4] Peirce, *Collected Papers*, 5.423.
[5] Peirce, *Collected Papers*, 4.1.
[6] James, *Pragmatism*, p.37.
[7] James, *Pragmatism*, p.106.
[8] Dewey, *Logic, The Theory of Inquiry*, p.345n.
[9] Rorty, *Consequences of Pragmatism*, p.xiii.
[10] Rorty, 'Introduction' to Murphy, *Pragmatism from Peirce to Davidson*, p.1.
[11] Rorty, *Philosophy and the Mirror of Nature*, pp.308-9.
[12] Rorty, 'Trotsky and the Wild Orchids', p.141.
[13] Rorty, *Consequences of Pragmatism*, p.xvii.
[14] Schlick, 'Positivism and Realism,' pp.105, 107.
[15] Carnap, 'The Rejection of Metaphysics', p.211.
[16] Feigl, 'Some Major Issues and Developments in the Philosophy of Science of Logical Empiricism', pp.16ff..
[17] Peirce, *Collected Papers*, 5.470.
[18] Fine, 'The Natural Ontological Attitude', p. 98.
[19] Gross, *The Rhetoric of Science*, pp.193-208.
[20] Putnam, *Renewing Philosophy*, pp.112-3.

REFERENCES

Alston, W. P.: 1979, 'Yes, Virginia, There is a Real World', *Proceedings and Addresses of the American Philosophical Association* **52**, 779-808.

Armstrong, D. M.: 1978, *Universals and Scientific Realism*, Cambridge University Press, Cambridge, two volumes.

Armstrong, D. M.: 1989, *Universals: An Opinionated Introduction*, Westview Press, Boulder, CO.

Armstrong, D. M.: 1995, 'Naturalism, Materialism, and First Philosophy', in Moser and Trout, *Contemporary Materialism: A Reader*, pp. 35-50.

Ayer, A. J.: 1936, *Language, Truth and Logic*, Victor Gollancz, London (2nd ed., 1946; and Dover, New York, 1952).

Ayer, A. J. (ed.): 1959, *Logical Positivism*, Free Press, New York.

Berkeley, G.: 1710, *A Treatise Concerning the Principles of Human Knowledge*, Part I, Turbayne, Colin (ed.), Bobbs-Merrill, Indianapolis, IN, 1957.

Bhaskar, R.: 1975, *A Realist Theory of Science*, Leeds Books, Leeds.

Bhaskar, R.: 1979, *The Possibility of Naturalism: A Philosophical Critique of Contemporary Human Sciences*, Harvester Press, Brighton.

Bhaskar, R.: 1986, *Scientific Realism and Human Emancipation*, Verso, New Left Books, London.

Blackburn, S.: 1993, *Essays in Quasi-Realism*, Oxford University Press, Oxford.

Boyd, R.: 1980, 'Scientific Realism and Naturalistic Epistemology', *PSA*, vol.2, P. Asquith and R. Giere (eds.).

Boyd, R.: 1984, 'The Current Status of Scientific Realism', in Leplin (ed.), *Scientific Realism*, pp. 41-82, and in Boyd *et al.*, *The Philosophy of Science*, pp. 195-222.

Boyd, R.: 1989, 'What Realism Implies and What It Does Not', *Dialectica* **43**, 5-29.

Boyd, R.: 1992, 'Constructivism, Realism, and Philosophical Method', in J. Earman (ed.), *Inference, Explanation, and Other Frustrations: Essays in the Philosophy of Science*, University of California Press, Berkeley, pp. 131-198.

Boyd, R., P. Gasper and J. D. Trout (eds.): 1991, *The Philosophy of Science*, MIT Press, Cambridge, MA.

Brown, J. R.: 1994, *Smoke and Mirrors: How Science Reflects Reality*, Routledge, London.

Bunge, M.: 1981, *Scientific Materialism*, Reidel, Dordrecht.

Campbell, K: 1976, *Metaphysics: An Introduction*, Dickenson, Encino, CA.

Carnap, R.: 1932, 'The Elimination of Metaphysics Through the Logical Analysis of Language', in Ayer (ed.), *Logical Positivism*, pp. 60-81.

Carnap, R.: 1935, 'The Rejection of Metaphysics', reprinted from *Philosophy and Logical Syntax*, Kegan Paul, Trench, Tubner & Co., London, Part I, in M. Weitz (ed.), *Twentieth-Century Philosophy: The Analytic Tradition*, Free Press, New York, 1966, pp. 207-219 (page references to the latter).

Carnap, R.: 1950, 'Empiricism, Semantics, and Ontology', *Revue Internationale de Philosophie* **4.11**, 20-40; reprinted in L. Linsky (ed.), *Semantics and the Philosophy of Language*, University of Illinois Press, Urbana, IL, 1952, pp. 208-28; in P. P. Weiner (ed.), *Readings in the Philosophy of Science: Introduction to the Foundations and Cultural Aspects of the Sciences*, Charles Scribner's Sons, New York, 1953, pp. 509-22 and pp. 633-4; in J. L. Jarrett and S. M. McMurrin (eds.), *Contemporary Philosophy: A Book of Readings*, Henry Holt, New York, 1954, pp. 377-90; in Carnap, *Meaning and Necessity: A Study in Semantics and Modal Logic*, University of Chicago Press, 2nd edition, 1956, pp. 205-22; and in H. Putnam and P. Benacerraf (eds.), *Philosophy of Mathematics: Selected Readings*, pp. 233-48.

Carnap, R.: 1928, *The Logical Structure of the World and Pseudoproblems in Philosophy*, trans. Rolf George, University of California Press, Berkeley, 1967.

Cartwright, N.: 1983, *How the Laws of Physics Lie*, Oxford University Press, Oxford.

Church, A.: 1952, 'A Formulation of the Logic of Sense and Denotation', in P. Henle (ed.), *Structure, Method and Meaning*, Liberal Arts, New York, pp. 3-24.

Church, A.: 1956, 'Propositions and Sentences', in *The Problem of Universals*, Notre Dame University Press, Notre Dame, IN.

Church, A.: 1958, 'Ontological Commitment', *Journal of Philosophy* **55**, 1008-14.

Churchland, P. M.: 1981, 'Eliminative Materialism and the Propositional Attitudes', *Journal of Philosophy* **LXXXVIII.2**, 67-89; reprinted in Churchland, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, MIT Press, Cambridge, Ma, 1989, pp. 1-22, and in Moser and Trout (eds.), *Contemporary Materialism: A Reader*, pp. 150-179.

Churchland, P. M., and C. Hooker (eds.): 1985, *Images of Science: Essays on Realism and Empiricism*, University of Chicago Press, Chicago.

Davidson, D.: 1965, 'Theories of Meaning and Learnable Languages', in *Proceedings of the International Congress for Logic, Methodology and Philosophy of Language*, Y. Bar-Hillel (ed.), North-Holland, Amsterdam, pp. 383-394; and in his *Inquiries into Truth and Interpretation*, pp. 3-16.

Davidson, D.: 1969, 'Truth and Meaning', *Synthese* **17**, 304-23; and in his *Inquiries into Truth and Interpretation*, pp. 17-36.

Davidson, D.: 1984, *Inquiries into Truth and Interpretation*, Clarendon Press, Oxford.

Davidson, D.: 1990, 'The Structure and Content of Truth', *Journal of Philosophy* **LXXXVII**, 279-328.

Davidson, D.: 1996, 'The Folly of Trying to Define Truth', *Journal of Philosophy* **XCIII.6**, 263-78.

Devitt, M.: 1984, *Realism and Truth*, Blackwell, Oxford (2nd edition, 1991).

Dewey, J.: 1938, *Logic, The Theory of Inquiry*, Henry Holt, New York.

Dummett, M. A. E.: 1959, 'Truth', *Proceedings of the Aristotelian Society* **59**, 141-62; reprinted in G. Pitcher (ed.), *Truth*, Prentice-Hall, Englewood Cliffs, NJ, 1964, pp. 93-112; in P. F. Strawson (ed.), *Philosophical Logic*, Oxford University Press, Oxford, 1967, pp. 49-68; and in Dummett, *Truth and Other Enigmas*, pp. 1-24.

Dummett, M. A. E.: 1975, 'The Philosophical Basis of Intuitionist Logic', in H. E. Rose and J. C. Shepherdson (eds.), *Logic Colloquium '73*, North-Holland, Amsterdam, pp. 5-40; reprinted in Dummett, *Truth and Other Enigmas*, pp. 215-247.

Dummett, M. A. E.: 1976, 'What Is a Theory of Meaning?', in G. Evans and J. McDowell (eds.), *Truth and Meaning: Essays in Semantics*, Clarendon Press, Oxford, pp. 67-137.

Dummett, M. A. E.: 1978, *Truth and Other Enigmas*, Duckworth, London.

Dummett, M. A. E.: 1982, 'Realism', *Synthese* **52**, 55-112.

Dummett, M. A. E.: 1991, *The Logical Basis of Metaphysics*, Harvard University Press, Cambridge, MA.

Dummett, M. A. E.: 1988, 'Is the Concept of Truth Needed for Semantics?', in C. Martinez, U. Rivas, and L. Villegas-Forero (eds.), *Truth in Perspective*, pp. 3-22.

Elgin, C. Z. (ed.): 1997, *Nominalism, Constructivism, and Relativism in the Work of Nelson Goodman*, Garland, New York.

Feigl, H.: 1956, 'Some Major Issues and Developments in the Philosophy of Science of Logical Empiricism', in H. Feigl and M. Scriven (eds.), *Minnesota Studies in the Philosophy of Science*, vol.I, University of Minnesota Press, Minneapolis, pp. 3-37.

Field, H.: 1980, *Science Without Numbers: A Defense of Nominalism*, Princeton University Press, Princeton.

Field, H.: 1982, 'Realism and Relativism', *Journal of Philosophy* **79**, 553-67.

Fine, A.: 1984a, 'The Natural Ontological Attitude', in Leplin (ed.), *Scientific Realism*, pp. 83-107.

Fine, A.: 1984b, 'And Not Anti-Realism Either', *Noûs* **18**, 51-65.

Fine, A.: 1986, 'Unnatural Attitudes: Realist and Instrumentalist Attachments to Science', *Mind* **95**, 149-79.

Fine, A.: 1991, 'Piecemeal Realism', *Philosophical Studies* **61**, 79-96.

van Fraassen, B. C.: 1980, *The Scientific Image*, Clarendon Press, Oxford.

Gibson, J. J.: 1966, *The Senses Considered as Perceptual Systems*, Houghton Mifflin, Boston.

Gibson, J. J.: 1967, 'New Reasons for Realism', *Synthese* **17**, 162-72.

Giere, R.: 1988, *Explaining Science: A Cognitive Approach*, University of Chicago Press, Chicago.

Glymour, C.: 1982, 'Conceptual Scheming or Confessions of a Metaphysical Realist', *Synthese* **51**, 169-80.

Goodman, N.: 1951, *The Structure of Appearance*, Harvard University Press, Cambridge, MA (2nd edition, Bobbs-Merrill, Indianapolis, IN, 1966; 3rd edition, Reidel, Dordrecht, 1977).

Goodman, N.: 1956, 'A World of Individuals', in *The Problem of Universals*, Notre Dame University Press, Notre Dame, IN; reprinted in H. Putnam and P. Benacerraf (eds.), *Philosophy of Mathematics: Selected Readings*, pp. 197-210.

Goodman, N.: 1975, 'Words, Works, Worlds', *Erkenntnis* **9**, and in *Ways of Worldmaking*, Harvester, Hassocks, Sussex, 1978, pp. 1-22.

Goodman, N. and W. V. O. Quine: 1947, 'Steps Towards a Constructive Nominalism', *Journal of Symbolic Logic* **12**, 105-122.

Gregory, R. L.: 1966, *Eye and Brain: The Psychology of Seeing*, Weidenfeld and Nicholson, London (second edition, 1972).

Gross, A.: 1990, *The Rhetoric of Science*, Harvard University Press, Cambridge, MA (2nd edition, 1996).

Haack, S.: 1977, 'Lewis's Ontological Slum', *Review of Metaphysics* **30.3**, 415-29.

Haack, S.: 1987, '"Realism"', *Synthese* **73**, 275-99.

Haack, S.: 1992, '"Extreme Scholastic Realism"; Its Relevance to Philosophy of Science Today', *Transactions of the Charles S. Peirce Society* **XXVIII.1**, 19-50.

Haack, S.: 1993, *Evidence and Inquiry: Towards Reconstruction in Epistemology*, Blackwell, Oxford.

Haack, S. 1994, 'How the Critical Common-sensist Sees Things', *Histoire, épistémologie, langage* **16.1**, 9-34.

Haack, S. 1995, 'Multiculturalism and Objectivity', *Partisan Review* **LXII.3**, 397-405; reprinted in *Manifesto of a Passionate Moderate: Unfashionable Essays*, pp. 137-48.

Haack, S.: 1996, 'Reflections on Relativism: From Momentous Tautology to Seductive Contradiction', in Tomberlin (ed.), *Philosophical Perspectives, 10: Metaphysics*, pp. 287-315, and in *Noûs*, Supplement, 287-315; reprinted in C. Martinez, U. Rivas and L. Villegas-Forero (eds.), *Truth in Perspective*, pp. 295-316; and in Haack, *Manifesto of a Passionate Moderate: Unfashionable Essays*, pp. 149-66..

Haack, S.: 1998, *Manifesto of a Passionate Moderate: Unfashionable Essays*, University of Chicago Press, Chicago.

Hacking, I. M.: 1983, *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*, Cambridge University Press, Cambridge.

Hacking, I. M.: 1999, *The Social Construction of What?*, Harvard University Press, Cambridge, MA.

Harre, R.: 1986, *Varieties of Realism: A Rationale for the Natural Sciences*, Blackwell, Oxford.

Harris, J. F.: 1992, *Against Relativism: A Philosophical Defense of Method*, Open Court, La Salle, IL.

Hempel, C. G.: 1958, 'The Theoretician's Dilemma', in H. Feigl, M. Scriven and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science*, vol. II, pp. 37-98; and in *Aspects of Scientific Explanation*, Free Press, Glencoe, 1965, pp. 172-226.

Hempel, C. G.: 1934/5, 'On the Logical Positivists' Theory of Truth', *Analysis* **2**, 49-59.

Horwich, P.: 1982, 'Three Forms of Realism', *Synthese* **51**, 181-21.

Hoy, R. C, and L. N. Oaklander (eds.): 1991, *Metaphysics: Classical and Contemporary Readings*, Wadsworth, Belmont, CA.

Hume, D.: 1739/40, *A Treatise of Human Nature*.

Hume, D.: 1748, *An Inquiry Concerning Human Understanding*, posthumous edition, with Hume's corrections, 1777.

James, W.: 1907, *Pragmatism*, F. Burkhardt and F. Bowers (eds.), Harvard University Press, Cambridge, MA, 1975.

James, W.: 1909a, *The Meaning of Truth*, F. Burkhardt and F. Bowers (eds.), Harvard University Press, Cambridge, MA, 1975.

James, W.: 1909b *A Pluralistic Universe*, Longmans, Green, New York.

James, W.: 1912, *Essays in Radical Empiricism*, R. B. Perry (ed.), Longman's, Green & Co., New York.

Kloesel, C. and N. Houser (eds.): 1992, *The Essential Peirce*, Indiana University Press, Bloomington and Indianapolis, IN.

Knorr-Cetina, K.: 1981, *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*, Pergamon Press, Oxford.

Koertge, N.: 1995, 'Wrestling With the Social Constructor', in P. R. Gross *et al.* (eds.), *The Flight From Science and Reason*, *Annals of the New York Academy of Sciences* 775, pp. 266-73; reprinted by Johns Hopkins University Press, Baltimore, MD, 1997.

Krausz, M. (ed.): 1989, *Relativism: Interpretation and Confrontation*, University of Notre Dame Press, Notre Dame, IN.

Kripke, S.: 1972, 'Naming and Necessity', in G. Harman and D. Davidson (eds.), *The Semantics of Natural Language*, Reidel, Dordrecht, pp. 253-355; published as a book by Harvard University Press, Cambridge, MA, 1980.

Kuhn, T. S.: 1962, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago (3rd edition, 1996).

Laudan, L.: 1984, 'A Confutation of Convergent Realism', in Leplin (ed.), *Scientific Realism*, pp. 218-49.

Laudan, L.: 1990, *Science and Relativism: Some Key Controversies in the Philosophy of Science*, University of Chicago Press, Chicago.

Leplin, J. (ed.): 1984, *Scientific Realism*, University of California Press, Berkeley.

Levin, M.: 1984, 'What Kind of Explanation is Truth?', in Leplin (ed.), *Scientific Realism*, pp. 124-139.

Lewis, D. K.: 1974, *Counterfactuals*, Blackwell, Oxford.

Locke, J.: 1690, *An Essay Concerning Human Understanding*, Alexander Campbell Fraser (ed.) (1891), Dover, New York, 1959.

Margolis, J.: 1986, *Pragmatism Without Foundations: Reconciling Realism and Relativism*, Blackwell, Oxford.

Margolis, J.: 1991, *The Truth About Relativism*, Blackwell, Oxford.

Martinez, C., U. Rivas and L. Villegas-Forero (eds.): 1998, *Truth In Perspective: Recent Issues in Logic, Representation and Ontology*, Ashgate, Aldershot, Hants.

Maxwell, G.: 1963, 'The Ontological Status of Theoretical Entities', in H. Feigel and G. Maxwell (eds.), *Scientific Explanation, Space and Time*, University of Minnesota Press, Minneapolis, pp. 3-27.

McCormick, P. (ed.): 1996, *Starmaking*, MIT Press, Cambridge, MA.

Migotti, M.: 1999, 'Peirce's Double-Aspect Theory of Truth', in C. J. Misak (ed.), *Pragmatism*, *Canadian Journal of Philosophy*, Supplementary Volume **24** (issue date 1998, published 1999).

Mill, John Stuart: 1865, *An Examination of Sir William Hamilton's Philosophy*, London (6th edition, 1889).

Moser, P. K., and J. D. Trout: 1995, *Contemporary Materialism: A Reader*, Routledge, London.

Niiniluoto, I.: 1987, 'Varieties of Realism', in P. Lahti and P. Mittelstaedt (eds.), *Symposium on Modern Physics, 1987*, World Scientific, Singapore, pp. 459-83.

Niiniluoto, I.: 1991a, 'Realism, Relativism and Constructivism', *Synthese* **89**, 135-62.

Niiniluoto, I.: 1991b, 'What's Wrong With Relativism', *Science Studies* **4**, 17-24.

Nola, R. (ed.): 1988, *Relativism and Realism in Science*, Kluwer Academic Publishers, Dordrecht, the Netherlands.

Peirce, C. S.: *Collected Papers*, P. Weiss, C. Hartshorne and A. Burks (eds.), Harvard University Press, Cambridge, MA, 1931-58. References are by volume and paragraph number.

Pihlström, S.: 1996, *Structuring the World: The Issue of Realism and the Nature of Ontological Problems in Classical and Contemporary Pragmatism*, Acta Philosophica Fennica, Helsinki.

Popper, K. R.: 1972, *Objective Knowledge,* Clarendon Press, Oxford.

Popper, K. R.: 1983, *Realism and the Aim of Science*, W. W. Bartley III (ed.), Rowman and Littlefield, Totowa, NJ.

Popper, K. R.: and J. C. Eccles: 1977, *The Self and Its Brain: an Argument for Interactionism*, Springer International, Berlin.

Prawitz, D.: 1998, 'Truth from a Constructivist Perspective', in C. Martinez, U. Rivas, and L. Villegas-Forero (eds.), *Truth in Perspective: Recent Issues in Logic*, pp. 23-36.

Putnam, H.: 1975, 'The Meaning of "Meaning"', in *Mind, Language and Reality*, Philosophical Papers, vol.2, Cambridge University Press, Cambridge, pp. 215-71.

Putnam, H.: 1978, *Meaning and the Moral Sciences*, Routledge and Kegan Paul, London.

Putnam, H.: 1982, 'Why There Isn't a Ready-Made World', *Synthese* **51**, 141-68; and in *Realism and Reason*, Cambridge University Press, Cambridge, 1983, pp. 205-228.

Putnam, H.: 1984, 'What is Realism?', in Leplin (ed.), *Scientific Realism*, pp. 140-53.

Putnam, H.: 1987, *The Many Faces of Realism*, Open Court, La Salle, IL.

Putnam, H.: 1990, *Realism With a Human Face*, J. Conant (ed.), Harvard University Press, Cambridge, MA.

Putnam, H.: 1992, *Renewing Philosophy*, Harvard University Press, Cambridge, MA.

Putnam, H.: 1994, 'Sense, Nonsense and the Senses: An Inquiry Into the Powers of the Human Mind', *Journal of Philosophy* **XCI.9**, 445-517.

Putnam, H. and P. Benacerraf (eds.): 1964, *Philosophy of Mathematics: Selected Readings*, Prentice-Hall, Englewood Cliffs, NJ.

Quine, W. V. O.: 1969, 'Natural Kinds', in *Ontological Relativity and Other Essays*, Columbia University Press, New York, pp. 114-38; reprinted in S. P. Schwartz (ed.)., *Naming, Necessity and Natural Kinds*, Cornell University Press, Ithaca, NY, pp. 155-75.

Reid, T.: 1785, *Essays on the Intellectual Powers*, in V. Hamilton (ed.), *Philosophical Works of Thomas Reid*, Edinburgh, 1846.

Rescher, N.: 1992, *A System of Pragmatic Idealism, vol. I: Human Knowledge in Idealistic Perspective*, Princeton University Press, Princeton, NJ.

Rescher, N.: 1993, *A System of Pragmatic Idealism, vol. II: The Validity of Values*, Princeton University Press, Princeton, NJ.

Rescher, N.: 1994, *A System of Pragmatic Idealism, vol. III: Metaphilosophical Inquiries*, Princeton University Press, Princeton, NJ.

Robinson, H. (ed.): 1993, *Objections to Physicalism*, Clarendon Press, Oxford.

Rorty, R.: 1979, *Philosophy and the Mirror of Nature*, Princeton University Press, Princeton, NJ.

Rorty, R.: 1982, *Consequences of Pragmatism*, Harvester Press, Hassocks, Sussex.

Rorty, R.: 1990, 'Introduction' to J. P. Murphy, *Pragmatism from Peirce to Davidson*, Westview Press, Boulder, CO, pp. 1-6.

Rorty, R.: 1991, *Objectivity, Relativism and Truth*, Cambridge University Press, Cambridge.

Rorty, R.: 1992, 'Trotsky and the Wild Orchids', *Common Knowledge* **1.3**, 140-53.

Rorty, R.: 1997, 'Realism, Anti-Realism and Pragmatism: Comments on Alston, Chisholm and Davidson', in C. Kulp (ed.), *Realism/Antirealism and Epistemology*, Rowman and Littlefield, Lanham, MD, pp. 147-72.

Russell, B.: 1910, 'Pragmatism', and 'William James's Conception of Truth', in *Philosophical Essays*, Longman's, Green & Co., New York.

Russell, B.: 1921, *The Analysis of Mind*, George Allen and Unwin, London, and Humanities Press, Inc., New York.

Schiller, F. C. S.: 1903, *Humanism: Philosophical Essays*, MacMillan & Co., London.

Schiller, F. C. S.: 1907, *Studies in Humanism*, MacMillan & Co., London.

Schiller, F. C. S.: 1939, *Our Human Truths*, Columbia University Press, New York.

Schlick, M.: 1932/3, 'Positivism and Realism', *Erkenntnis* **3**, 1-31, and in Ayer (ed.), *Logical Positivism*, pp. 82-107 (page references to the latter).

Searle, J.: 1995, *The Construction of Social Reality*, Free Press, New York.

Sellars, W.: 1963, *Science, Perception and Reality*, Routledge and Kegan Paul, London.

Smart, J. J. C.: 1963, *Philosophy and Scientific Realism*, Routledge and Kegan Paul, London.

Staniland, H.: 1972, *Universals*, Anchor Books, Doubleday, New York.

Tarski, A.: 1944, 'The Semantic Conception of Truth', *Philosophy and Phenomenological Research* **IV**, reprinted in H. Feigl and W. Sellars (eds.), *Readings on Philosophical Analysis*, Appleton-Century Crofts, New York, 1949, pp. 52-84.

Taylor, R.: 1963, *Metaphysics*, Prentice-Hall, Englewood Cliffs, NJ (4th edition, 1992).

Tomberlin, J. E. (ed.): 1987, *Philosophical Perspectives, 1: Metaphysics*, Ridgeview, Atascadero, CA.

Tomberlin, J. E. (ed.): 1996, *Philosophical Perspectives, 10: Metaphysics*, Blackwell, Oxford (also published as a Supplement to *Noûs*, 1996).

Trigg, R.: 1980, *Reality at Risk: A Defence of Realism in Philosophy and the Sciences*, Harvester, Brighton, Barnes and Noble Books, Totowa, NJ; Wheatsheaf, New York, 1989.

de Waal, C.: 1996, 'The Real Issue Between Nominalism and Realism: Peirce and Berkeley Reconsidered', *Transactions of the Charles S. Peirce Society*, **XXII.3**, 425-42.

de Waal, C.: 1998, 'Peirce's Nominalist/Realist Distinction, an Untenable Dualism', *Transactions of the Charles S. Peirce Society*, **XXIV.1**, 179-98.

Wolterstorff, N.: 1987, 'Are Concept-Users World-Makers?', in Tomberlin (ed.), *Philosophical Perspectives, 1: Metaphysics*, pp. 233-267.

GÖRAN SUNDHOLM

ANTIREALISM AND THE ROLES OF TRUTH

The position indicated by the term "antirealism" has come to the fore in philosophical discussion during the last thirty years, largely as a result of the writings of Michael Dummett.[1] *Prima facie* one would expect an antirealist critique of realism to comprise two parts: a negative part (as the name indicates) criticising the realist view, and a positive part setting out a constructive alternative to the realist position. The tradition starting with Dummett, it is fair to say, has mainly been concerned with the negative part of antirealism. Here, however, the antirealist appears to have a large, and definite, disadvantage in his perennial discussion with the realist, purely in virtue of the chosen label; his position is laid down *negatively* in opposition to a *prior* realist position. Accordingly, the antirealist's campaign will be fought on a field determined by his opponent, and in terms chosen by him. Therefore, in this Chapter, I do not take the views of Dummett as my point of departure, especially since the need for a survey of antirealism from this perspective has been excellently met by Hale (1997). Furthermore, in a Handbook of *Epistemology*, this manner of proceeding (that is, not taking Dummett's views as the point of departure) is quite appropriate, since one of Dummett's main tenets is the primacy of the theory of meaning over the theory of knowledge. Instead, I discuss a number of traditional epistemological notions and theories from an antirealist standpoint, and I also treat of the early work in modern mathematical intuitionism that provided the inspiration for much of the current work on semantical antirealism. Thus the present Chapter can be seen as providing an epistemological and philosophico-mathematical background to the modern semantical versions of antirealism.

1. THE ROLES OF TRUTH IN RELATION TO THE TRADITIONAL THEORIES OF TRUTH

Dummett famously recasts the debate between realism and idealism as a semantical dispute concerning the form that a proper theory of meaning has to take, and, in particular, what role has to be played by the notion of truth in such a theory. However, independently of these specific Dummettian concerns, it is clear that truth does have a major *semantical role* to play: after Frege and Wittgenstein many hold it to be a truism that

> the meaning of a declarative sentence is given by, or in terms of, its truth-condition.[2]

This semantical office, however, is not the only one in which truth has to serve; an equally important *epistemological role* for truth is given by a similar truism:

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 437-466.
© 2004 *Kluwer Academic Publishers.*

> what is known has to be true.

Finally, a third, *metaphysical role* of truth, so firmly stressed by Frege in the preface to his *Grundgesetze,* is that of making objectivity possible:

> really being true is conceptually different from appearing to be true,
> that is, the distinction between appearance and reality must be upheld.

In view of their truistic character, an account of truth, meaning and knowledge that respects these platitudes is, *ceteris paribus*, preferable to an account that does not, and one would certainly expect a correct account to throw light on why the maxims in question have been considered truistic.

The epistemological tradition knows various so called *theories of truth.* What these traditional theories offer are general conceptions of truth; in the modern jargon they are theories of truth, that is, of, or about, the *concept* of truth, but they are not (Tarskian) truth theories that tell us under what condition the sentences of a certain language are true. These general conceptions of truth turn out to be admirably geared towards various offices as given by the above truisms. Thus, for instance, the *evidence theory* of truth, according to which what is true is what can be made evident (that is, known), caters very well for the epistemological role.[3] Indeed, on this evidence-theoretical reading the maxim

> what is known has to be true

becomes

> what is known (what has been justified, warranted, made evident, etc.) has to be true (justifiable, warrantable, evidenceable, knowable, etc.),

and this is *a priori* obvious from the *ab esse ad posse* principle: what has already been done is certainly doable.[4]

Similarly, the traditional *correspondence theory* considers certain truth-bearers, be they judgements in the mind, or declarative sentences in the language, or propositions in the third realm of abstract entities, and relates these to appropriate truth-makers in the world:

> a truth-bearer is true if a corresponding truth-maker exists.

When the truth-bearers are sentences, this maxim provides just the sort of language-world link required for the semantical role of truth.

Finally, the metaphysical role of truth is catered for by the *pragmatic* and *coherence theories* of truth. The main task for the notion of truth when serving in this office (and perhaps even in general) is to hold open the possibility of making mistakes, that is, to rule out *epistemological nihilism*, by which I mean an epistemological counterpart to *moral nihilism.* This ethical position is characterised by the maxim

> 'If God is dead, everything is permitted'.[5]

The (contraposition of the) corresponding epistemological maxim is

> 'If mistakes are possible, then there is a norm of rightness'.

Another way of characterising epistemological nihilism is via the *Homo mensura* thesis of Protagoras:

> Man is the measure of things.

If that be so, there is no difference between how things-seem-for-me and how they are, and mistakes are accordingly ruled out. The normative notion of rightness allows for an absolute distinction between appearance and reality and makes room for mistakes: a mistake is an act of knowledge that is not right. It is at issue in examples such as

> I thought that he was a friend, but he turned out not to be a *true* friend.
> In the Netherlands there is butter (known elsewhere as *margarine*) and *real* (or true)"*cream*" butter.
> For many years Kummer's "proof" or the Four Colour Theorem was accepted, but in the end it was rejected as invalid. It was not a *true* proof.

This type of truth – "truth of things" – is known in scholastic philosophy as *rectitudo*.[6] Now, when a mistake is discovered, or suspected, at least one act of knowing (that is, of getting to know) cannot be right. Accordingly, one act, at least, will have to be annulled. The coherence and pragmatist theories of truth provide criteria for how to choose among the candidates for annulment: clearly an epistemological act that issues in a result that does not cohere with the body of knowledge is a strong candidate for annulment. Similarly, deeds with results that do not work, or are otherwise of no use, will be annulled. The above discussion can be summarised in a schema:

TRADITIONAL TRUTH THEORIES AND THE ROLES OF TRUTH

	Correspondence Theory		semantical	
Truth according to the	Evidence Theory	caters for the	epistemological	role of truth
	Coherence and Pragmatic Theories		metaphysical	

2. ACTS AND OBJECTS

In his characterisation of the realism/idealism ("antirealism") debate Fichte noted that basically there are only two epistemological options.[7] The positions may be formulated in terms of the act/object dichotomy:

$$\begin{array}{cc}\text{act} & \\ \cline{1-1} & \text{object}\end{array}$$

Either you determine the object of knowledge as the object of the act, and then you are an *idealist*, or you determine the act in terms of a prior object towards which the act is directed, and then you are a *realist* (or "dogmatist", as Fichte said, being an idealist himself).[8] If Fichte is right in this (and I suspect he is), the point is moot whether there is a neutral background position from which the issue between the realists and idealists can be adjudicated. For Fichte, it is clear that there is not. If this be so, it would serve to explain why the realism/idealism debate so often makes a futile impression; in place of a clear-cut decision, we find endless refinements of positions into sterile scholasticism, and conversions from one side to the other rarely take place. Brentano, Husserl, and Putnam are examples that spring to mind, all of whom came to reject their original realist stance.[9] Moore and Russell, on the other hand, it is well-known, converted to realism from a prior adherence to British idealism.

It should be noted that, traditionally, the act through which we gain knowledge can be either mediate or immediate. The object of an immediate act is intuitive, or non-discursive, knowledge, that is, an *axiom*, but not in the current Hilbertian hypothetico-deductive sense: the axiomatic objects of immediate acts are self-evident truths, which neither need nor are capable of further demonstration. The axiom is evident in itself: knowledge of its terms and composition suffices for knowing its truth. In the scholastic terminology the axiom is a *propositio per se nota,* whose evidence is *ex vi terminorum*, that is, in virtue of the terms (concepts) out of which it is composed.[10] 'Meaning is what essence becomes when it is divorced from the object of reference and wedded to the word', quips Quine. Hence, after the so called *linguistic turn*, one often says that axioms are analytic, in the sense that they are 'true by virtue of meaning and independently of fact'.[11] In fact, owing to a prevalent conventionalist view of meaning, axioms are even held to be *conventions*.

A mediate act of knowledge, on the other hand, is nothing but an act of inference in which one draws the conclusion J from the known premises $J_1, \ldots, J_k$. This act thus makes use of the mode of inference I:

$$\frac{J_1 \quad J_2 \quad \cdots \quad J_k}{J.}$$[12]

The appropriate notion of correctness for such (modes of) inference is that of *validity*. A mediate act of knowledge according to the mode of inference I takes the form (where the $J_1, \ldots, J_k$ are the objects of prior acts of knowledge):

$$\frac{\begin{matrix} | & | & & | \\ J_1 & J_2 & \cdots & J_k \end{matrix}}{J.}$$

This disambiguation between the two readings of 'inference', namely, mediate *act* of judgement versus mode of inference, raises the question of the relation between the corresponding notions of correctness: acts of inference have to be *right*, whereas inference-modes have to be *valid.* How, if at all, do these notions relate to each other? Clearly, in a right mediate act of knowing, every axiom used must be really (truly) evident, and every inference-mode that has been applied must be really valid. Thus, the task remains of elucidating the validity of inferences. Below I shall treat of both realist and anti-realist criteria for inferential validity.

In the present Chapter I intend to explore antirealism from this act/object perspective. I shall canvas a series of positions and pay attention to the above truisms on truth, as well as to the traditional theories of truth and their proper place, if any, within the particular anti-realist framework that I am concerned to develop here. Throughout I am indebted to the example offered by Per Martin-Löf in the philosophical explanations of his constructive theory of types.[13] Taken together they constitute the only substantial anti-realist theory of meaning that has been developed so far.

3. TRUTH-BEARERS AND THE FORM OF JUDGEMENT.

A knowledge claim is commonly made through an assertion, that is, an assertoric utterance of a declarative sentence. For instance, by uttering the declarative

(1) Snow is white

assertorically, I assert that snow is white. An assertoric utterance of (1) makes no explicit knowledge claim, but the assertion made, nevertheless, comprises an implicit such claim. This can be seen from the fact that one is entitled to counter an assertion with a demand for the grounds upon which the assertion rests: once the assertion has been made, the "asserter" is obliged, when challenged, to provide answers to counter-questions such as

(2) How do you know this?
How do you know that snow is white?

Thus, incorporating the implicit claim to knowledge, the explicit form of the assertion made through an (assertoric) utterance of the declarative (1) is

(3) I know that snow is white.[14]

The assertion by means of (1), when understood in the sense of (3), might be called a *performative* knowledge claim, as opposed to a propositional one.[15]
An utterance of the declarative (1) suffices to effect the assertion that snow is white, but an utterance, on the other hand, of the propositional nominalization

(4) that snow is white

does not so suffice. In order to obtain an expression with which the assertion in question can be effected through a single assertoric utterance, however, it is enough to augment (4) into

(5) that snow is white is true

or, equivalently,

(6) it is true that snow is white.[16]

If the knowledge claim (3) is spelled out one obtains

(7) I know that it is true that snow is white

as the fully explicit form of the assertion made through an assertoric utterance of the declarative (1).

In this assertion, truth is ascribed to the propositional content given by (4). The declarative sentence expresses a *statement* of the form "truth ascribed to propositional content".[17] In the scholastic tradition, assertion is the external form of the interior act of judgement, and the assertion is the outward sign of the mental judgement made. The traditional form of judgement/assertion was

S is P,

that is, a two-term judgement of subject/copula/predicate form. The above tale, from (1) to (7), provides a reason for abandoning the traditional form of judgement in favour of

A is true,

where A is a proposition, that is, equivalently, in favour of the form

that S is true,

where S is a declarative sentence.

It should be noted that my above route to the novel form is not the one that was actually taken by the logical pioneers, to wit Bolzano, who introduced the new form, and Frege, who reached the same conclusion: *Ein Urteil ist mir nicht das bloße Fassen eines Gedanken, sondern die Anerkennung seiner Wahrheit.*[18]

In the light of the above discussion the following picture emerges with respect to the act of knowledge:

THE REALIST (BOLZANO-FREGE) THEORY

{content of object} ← act of knowledge
↓
[{Proposition A} is true]
↑
[object of the act]
= [asserted statement], [judgement known]

There are three interrelated levels in this schema:

(i) The assertoric/judgemental act or deed;
(ii) the statement used in the assertion/judgement ;
(iii) the propositional content of the statement.

To each of the three notions there corresponds a suitable correctness notion, to wit:

(i') the rightness (validity) of the act;
(ii') the correctness of the object of the act, that is, the asserted statement;
(iii') the truth of the propositional content of the asserted statement .

Our task is now to determine the relative order of priority of these notions within the realist respectively antirealist positions.

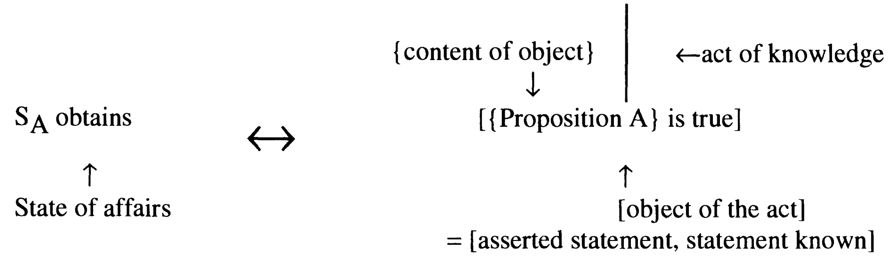
The notion (i') is the most fundamental; it is required in order to avoid epistemological nihilism. Without the notion of rightness applied to our deeds there is no way to differentiate between appearance and reality, between *Schein und Sein*: the distinction between right and right-for-me is abolished and anything goes. There are various options as to how to secure the norm of rightness in question. One that was followed by Bolzano (and also by Frege) is to take the classical – bivalent – truth of propositions as primitive: every proposition is true or false without further ado. One then readily explains the correctness of a statement in the following way: the statement that ascribes truth to a proposition is correct if the propositional content really is true, and the act of judgement is right if the (statement-)object is correct, that is, if the proposition that serves as content of the judgement made really is true. This radical and straightforward reduction of the rightness of acts, and of the correctness of judgements made, to the truth of propositions imposes a pleasing simplicity on the resulting realist epistemology.[19]

The Bolzano-Frege realist reduction makes all three notions of correctness subservient to propositional truth: the notion of truth that is applied to propositional contents then serves in all the offices of truth. The offices of truth, however, match the traditional theories of truth. Thus, under the realist reduction (whether tacit or explicit), with its conflation of the offices of truth, the traditional theories are turned into rival conceptions of truth for propositions: suddenly they are held to concern the same notion and impute different, and even contradictory, properties thereto.

For some realists the above epistemological reduction to propositional truth is not enough. Some, among whom Wittgenstein in the *Tractatus*, go even further and apply yet another reduction: the truth of a(n elementary) proposition A is reduced to the obtaining of a certain ontological state of affairs S_A, such that A is true if and only if S_A obtains. Through this reduction the desired epistemological notion of rightness is reduced to an ontological notion, namely that of the obtaining of states of affairs, which is then often thought of in a bivalent way: either the state in question does obtain or it does not. *Tertium non datur.*

The same pattern is obtained in modern versions of these views on knowledge and language, where a Tarskian model-theoretic semantics is applied to ordinary, or philosophical, discourse, and where the world is seen (often tacitly) as a (huge) relational structure in which every sequence of entities either does satisfy or does not satisfy a given "open sentence".[20] In such a way, then, via a realist semantics, be it Tarskian or not, an ontological norm of rightness is secured that suffices to avoid epistemological nihilism.

WITTGENSTEIN'S ONTOLOGICAL REDUCTION IN THE *TRACTATUS*



4. PROPOSITIONS AND TRUTH

The objects of the acts of assertion/judgement, that is the asserted statements/judgements made, ascribe truth to propositional contents, which can be rendered linguistically as that-clause nominalizations of declaratives. In order to complete the analysis, the notion of a proposition must be elucidated. Such an elucidation will contribute to both the negative and positive parts of the antirealist programme. The positive contribution, naturally enough, consists in a constructivist account of propositions, whereas the negative rests on the intuitionist criticism of non-constructive reasoning within mathematics.

This criticism was first voiced by Kronecker, who objected to the use of definitions by means of "undecidable", or perhaps better, as yet undecided cases. For certain number terms such definitions lead to computations that cannot be executed to a value in primitive, non-defined form. Consider the following example:

$$f(k) =_{\text{def}} \begin{cases} 1 \in \mathbf{N}, \text{ if Goldbach's Conjecture is true,} \\ 0 \in \mathbf{N}, \text{ if Goldbach's Conjecture is false.[21]} \end{cases}$$

This is meant to be a definition of a function f:**N**→**N**, but no values can be computed. For instance, according to the definition, f(23)∈**N**, but' we cannot indicate a natural number k such that f(23)=k.[22] Thus this "definition" by undecided cases introduces *defined* number terms, which cannot be eliminated in favour of terms in primitive notation. Such "definitions" allow for definitional equalities in which the definiendum cannot effectively be replaced by the definiens, thereby contravening the canons on definition that have been upheld ever since Pascal.[23]

The classical logical theory of Bolzano and Frege is *bivalent*: every proposition is true or false; in fact, being true-or-false, in the classical theory, is a characterising mark of propositions. But something is true-or-false only if it is true or if it is false, or so it is said. The need for every declarative to have a truth-value poses severe problems, for instance in connections with the quantifiers. Frege's explanation of the universal quantifier runs (where I have only made explicit the dependence on the domain of quantification) as follows:

∀-formation:

$$(\forall x \in D)A =_{\text{def}} \begin{cases} \text{The True, if } A[a/x] = \text{The True, for } a \in D. \\ \text{The False, otherwise,} \end{cases}$$

where A is a propositional function over the domain D, that is, A is a proposition, provided that x∈D.

When the domain D of quantification is infinite, or "unsurveyable", this separation of cases cannot be carried out effectively. It is exactly parallel to the above non-constructive way of attempting to define a function and it gives rise to similar difficulties: universal quantification, classically construed, introduces non-primitive means of notation that cannot be eliminated. This, however, is nothing but a version of Brouwer's (1908) criticism of the unrestrained use of classical logic within mathematics.[24] Note here that it is the universal quantifier *formation rule* that cannot be made evident on the classical conception of propositions: it simply is not clear that the universal quantifier takes a classical propositional function with respect to a domain into a classical proposition.[25]

Dummett has launched a controversial argument based on the presence of such "undecidable sentences" in the language (examples being quantification with respect to infinite domains, the remote past and future, sentience in others, and counterfactuals). Knowledge of a bivalent truth-condition for such an undecidable sentence cannot, in the end, Dummett holds, be "manifested", and so bivalent truth cannot serve as a key-concept in an adequate meaning theory for a sizeable language.[26]

To my mind the Kronecker-Brouwer rejection of classical bivalence more convincing, owing to its simplicity: if you want to avail yourself of classical logic across the whole board, irrespective of subject matter, you have to use defined expression that cannot be eliminated. Therefore, in a literal sense, the realist does not know what he is talking about. This, to me, is too high a price to pay. If one accepts this conclusion, the need for an alternative notion of proposition becomes obvious. Accordingly it is incumbent upon the constructivist to offer such an alternative.

5. Propositions: the constructive alternative

The 1920's constitute a period of transition in logical theory.[27] The universalist "Logic as language" paradigm that had been adhered to by the pioneers Frege, Russell, and Wittgenstein was gradually replaced by the metamathematical "Logic as calculus" approach that was emerging in the works of, among others, Skolem, Bernays, and Hilbert.[28] The novel metamathematical formal languages were, in the first instance, not meant for proving theorems *in*, but for proving (meta-)theorems *about*. The formal systems of the *Grundgesetze* and *Principia Mathematica*, on the other hand, were formulated as *interpreted* formal languages and the axioms and rules of inference had to be made evident under the given meaning explanations. Wittgenstein's *Tractatus*, in particular, can be seen as a grand attempt to provide a semantical foundation for *Principia Mathematica*. The works of Chwistek and Ramsey simplifying the theory of types also belong to this tradition.[29] The early systems of Church, of Curry, of Quine, and above all, of the mature Lesniewski, all fall under the Logic-as-language conception. It should be stressed that Heyting's seminal (1930) also belongs here: the formalization is understood as an interpreted one, but in that work the basic notions are left largely unexplained. Heyting's creation, nevertheless, intensified an already confused debate concerning "Brouwerian logic": was it not really a many-valued logic, using a third truth-value?[30]

Heyting (1930a) intervened in this debate, and, to all but few, put an end to the confusion. What he did was to give an explanation of the basic notions so that his formal systems became *interpreted* formal systems, more or less adequate for the expression of (part of) intuitionistic mathematics. In particular, Heyting articulated the relevant intuitionist notion of a proposition. From his intuitionist, or, as I prefer, *constructivist*, standpoint, a proposition is viewed as a *problem* (or *expectation*), which has to be solved by exhibiting a certain mathematical construction, namely its *proof*. His first example used the Euler-Mascheroni constant C:

The mathematical proposition

Euler's constant is rational

expresses the problem (or expectation) of finding a certain construction, namely a

pair of integers p and q such that C = p/q.

Heyting (1931, 1934) offered alternative formulations in terms of other basic concepts, and also Kolmogoroff (1932) gave an interesting alternative in terms of problems (*Aufgaben*) and their solutions. As Heyting (1958) came to realise, the various formulations were substantially equivalent. Finally, Howard (1969) introduced his formulae-as-types notion, which was refined by Martin-Löf (1984) into propositions-as-sets (of proof-objects).[31]

ALTERNATIVE CONSTRUCTIVIST NOTIONS OF A PROPOSITION

Heyting (1934)	Proposition	proof
Kolmogoroff (1932)	Problem (task)	solution
Heyting (1931) (1930)	Intention (expectation)	fulfilment (realization)
Howard (1969)	Type	term
Martin-Löf (1984)	Set	element

This notion of proof of a proposition, it must be stressed, is novel with intuitionism: in the tradition a proof is always of a theorem, that is, what is proved is always at the level of a judgement. Prior to Heyting, the notion of a proof of a judgemental content had found no use. Previously, proofs were either proof-acts (through which one gets to know a theorem) or (what I, following Martin-Löf, shall call) *traces* of such acts.[32] The novelty is reflected by the apt "proof-object" terminology that was introduced by Diller and Troelstra (1984). As is by now familiar, the meaning of a logical constant is explained in terms of how proof-objects may be formed for the propositions in which the constant in question serves as the main connective. The information may be presented in the form of "proof-tables" (which, from the point of view of meaning theory, play the same role as the truth-tables in classical semantics):

$(\bot)$ Nothing is a proof of $\bot$.
(&) When a is a proof-object for A and b is a proof-object for B,
<a,b> is a proof-object for A&B.
(v) When a is a proof-object for A, i(a) is a proof-object for AvB.
When b is a proof-object for B, j(b) is a proof-object for AvB.
$(\supset)$ When b is a proof-object for B, provided x is a proof-object for A,
$\lambda x.b$ is a proof-object for $A\supset B$.
$(\forall)$ When b is a proof object of B, provided that $x \in D$,
$\lambda x.b$ is a proof-object for $(\forall x \in D)P$.
$(\exists)$ When $a \in D$, and when b is a proof-object of B[a/x],
<a,b> is a proof-object for $(\exists x \in D)P$.[33]

One should here note the strong similarity between Gentzen's introduction rules in natural deduction style and these meaning explanations.[34] This suggests a necessary emendation in the reading of Heyting's clauses. What is given here is not a general formulation of how proof-objects for complex propositions may be formed out of proof-objects for their parts: these are explanations of how *canonical* (Brouwer), *direct* (Gentzen), *primitive* proofs may be formed out of parts, in analogy with how the primitive, or canonical, number terms

$$0, s(0), s(s(0)), s(s(s(0))), \ldots,$$

are given by the rules NI

$$0 \in N \qquad \text{and} \qquad \frac{a \in N}{s(a) \in N.}$$

However, as we all know, it is also possible to form non-primitive, defined, number terms, for instance,

$$14! + ([93/6] \cdot [3+8]).$$

The only constraint that is put on the use of such terms is that they admit of evaluation to a numerical value, that is, a number term in canonical form. Similarly, the only condition put on the means used for forming (non-canonical) proof-objects is that they admit of evaluation (or *normalization*, in the terminology of Prawitz) to canonical form.[35] In particular, in view of the reduction steps used in Prawitz's proof-theoretic normalization theorems, the standard elimination rules for the constructive logical constants are permitted in the formation of (non-canonical) proof-objects.

Consider, for instance, Prawitz's so called &-reduction according to which the derivation

$$\dfrac{\dfrac{\begin{matrix} D_1 \\ A_1 \end{matrix} \qquad \begin{matrix} D_2 \\ A_2 \end{matrix}}{A_1 \& A_2}\ (\&I)}{A_2}\ (\&E_2)$$

reduces to the derivation

$$\begin{matrix} D_2 \\ A_2. \end{matrix}$$

In linearized form, where the introduction- and elimination-rules are means for forming proof-objects, this becomes:

when D_1 is a derivation of A_1 and D_2 is a derivation of A_2, $\mathbf{\&E_2(\&I(}D_1, D_2\mathbf{))}$ = D_2 is a proof of A_2.

The analogy with:

when d_1 is a proof of A_1 and d_2 is a proof of A_2,

$\mathbf{p}_2(<d_1, d_2>) = d_2$ is a proof of A_2,

where $\mathbf{p}_2$ is the right-hand projection-operation associated with pair-formation (and which yields the second component upon application to a pair), should be obvious.

In summary, then, under the proof explanation of the constructive notion of proposition, to each proposition A there is associated a set (which might turn out to be empty in case the proposition is false) of proof-objects for A. Truth is then readily explained by means of a so called truth-maker analysis (with the proof-objects serving as truth-makers).[36]

The proposition A is true = there exists a proof of A.

This truth-condition for the proposition A is determined by (i) the proof-condition for A and (ii) the relevant notion of existence. The kind of existence that is here at issue is *not* that of the existential quantifier. The ∃-quantifier applies to propositional functions only, whereas the relation

$\Pi(a, A) =_{\mathrm{def}}$ a is a proof of the proposition A

is not propositional in nature: we do not explain $\Pi(a, A)$ by telling how a proof-object for this would be put together out of parts, owing to an infinite regress of ever-descending proof-explanations.[37] The relevant notion of constructive existence was made explicit by Hermann Weyl (1921):

I am entitled to claim that there exists an α only after having instantiated α.[38]

Here, then, we have a novel form of judgement: when α is a (general) concept

α exists

is a judgement, the assertion condition of which is given by the rule

$$\frac{a \text{ is an } \alpha}{\alpha \text{ exists.}}$$

The constructivists, though, are not the first to use existence as a form of judgement. In particular, they were anticipated by Brentano, who used only the two forms of judgement

α IS (or *exists*)

and

α IS NOT (or *does not exist*).[39]

The four traditional forms of categorical judgement were then reduced to these two forms. For instance,

All α are β

was reduced to

An α which is not β does not exist.[40]

When applied in the truth-condition of the proposition A, the constructive notion of existence yields a formulation of the assertion-condition for the statement that A is true:

One is entitled to assert that A is true,
only after having constructed a proof-object a for A.

This analysis (which is due to Per Martin-Löf (1994)) of the constructive form of judgement is, it should be noted, in a certain sense an epitome of the work of the previous century:

FORMS OF JUDGEMENT:

Traditional form:	S is P
Bolzano form:	proposition A is true.
Brentano form:	α exists
Truth-maker form:	there exists a truth-maker for A

Realist truth-maker analysis:	Constructive truth-maker analysis:
Bestehen of a *Sachverhalt*	there exists a proof of A which is reducible to a is a proof of A.

Thus the constructivist truth-maker analysis of

proposition A is true

is obtained by applying the Brentano form of judgement to the concept:

proof(-object) of A.

That form of judgement, that is,

proof(A) exists

is further reduced to an instance of the traditional form of judgement by means of an application of the constructivist analysis of existence:

a is (an element of) proof(A).

The constructive notion of propositional truth must not be understood in a modalized way: 'existence' does not mean 'possibility to find' in the formulation of simple truth. This can be seen by considering the different types of assumptions that result from the two notions in question. An assumption that A is true, that is, that a proof-object for A exists, is what we use in natural deduction when we aim to demonstrate the truth of a certain implication $A \supset B$. The use of such an assumption in no way presupposes that a proof-object for can *actually* be found; on the contrary, we all know of true implications with false antecedents. An assumption that A is true is compatible with the set proof(A) actually being empty.[41]

An assumption, on the other hand, that a proof-object *can be found* for A entails that the set A cannot turn out to be empty. It is instructive to carry out the discussion in terms of proof-objects, rather than in terms of truth. An assumption that A is true means considering an assumption of the form

$$x \in A.$$

Such an assumption can be used to infer, for instance, that $A \supset A$ is true, irrespective of the actual truth-value of A, by constructing the proof-object

$$\lambda x.x \in A \supset A.$$

For another example, consider a derivation involving the above &-elimination rule: assume that z is a proof of A&B. Under this assumption, $\mathbf{p}_2(z)$ is a proof of B. Therefore, discharging the assumption that z is a proof of A&B,

$$\lambda z.\mathbf{p}_2(z) \in A\&B \supset B,$$

which judgement holds irrespective of whether the propositions A and B are actually true.

The second assumption, that a proof-object can be found for A, assumes that A really is true, that is, that a proof-object $a \in A$ is obtainable. Under this assumption, it is incoherent that proof(A) turns out to be empty. An assumption that a proof-object of A can be found is, according to the explanations offered previously, the same as an assumption that the statement A is true is demonstrable (knowable), because in order to know that the proposition A is true, that is, in order to know that proof(A) exists, I must instantiate proof(A) by means of a proof-object. So, if it is demonstrable that A is true, a proof-object can be found. In order to grasp the difference we may consider an example. For every natural number k,

P(k) = that k is the number of window-panes in the City Hall of Leyden

is a proposition. Hence, for every $k \in N$,

P(k) is true

is a judgement (in the sense that its assertion-condition is determined). Truth can be demonstrated, however, only for *one* P(m) and the proof-object required for the truth of P(m) can be found as the result of a (tedious) counting-process.

An assumption, now, that P(k) is true may be used in the following way: "Assume that P(k) is true. The window-cleaning cost in Dutch guilders is 4 times the number of panes. Therefore, under our assumption, we should reserve 4 × k :- D. fl. in our budget. Therefore, the proposition

$(\forall k \in N)$ (that P(k)$\supset$that 4 × k :- D. fl. must be reserved in the budget)

is true. The truth of this proposition is compatible with *any* number of window panes.[42]

The other type of assumption leads to a different situation. "Assume that P(10.100) *really* is true. Then 40.400 :- D. fl. is the sum that we must reserve for the cleaning costs." Any other sum will, under the given circumstances, be off the mark and will make the budget incorrect.

Demonstrability ("provability") of statements, that is, truth for statements, is a modal notion, but truth for propositions is not. The matching two types of assumptions might be characterized as epistemic assumptions that statements are knowable versus alethic assumptions that propositions are true.

6. CORRECTNESS ("TRUTH") OF JUDGEMENTS.

For Brentano the judgement (statement), rather than the proposition (which notion he rejects), is the primary truth-bearer. His account of truth is a modal fusion of a correspondence theory and an evidence-theory of truth:

> a judgement is correct (*richtig*) if it agrees with (or corresponds to) the judgement that would be made by someone who judges with evidence.[43]

Above I tied the correspondence notion to the truth of propositions, and I therefore prefer, following Martin-Löf, to account for the modal and evidence-theoretical components in a slightly different fashion:

> a statement is correct (true) if it can be made with evidence.

The true statements are the evidenceable, knowable, warrantable, justifiable, ... ones. According to the discussion towards the end of the previous section, the statement that A is true is correct, that is, demonstrable, when a proof-object for the proposition A can be found.

Brentano, however, did not just construe judgemental truth according to an evidence theory. He also wished to locate the norm of rightness in the notion of evidence:

> Bei Evidenz ist Irrtum ausgeschlossen. Bei Evidenz ist auch Zweifel ausgeschlossen, aber weder Freiheit von Irrtum noch Freiheit von Zweifel macht das Urteil zum evidenten Urteil, sondern eine Eigentümlichkeit, die es als richtig charakterisiert.[44]

Here he goes to far in my opinion: the criterion of evidence is the Cartesian 'clear and distinct', but Brentano wishes this to be a criterion not only for evidence but also for freedom from error. According to him, when something is judged clearly and distinctly, error is ruled out. However, evidence is what makes us know, and thus, when evidence is taken to guarantee freedom from error, knowledge is infallible and error is ruled out. But error can never be ruled out. Hence evidence must not be conceptually equated with freedom from error. When error is diagnosed, the reaction will be: I thought it was evident, but in reality it was not.

The classical (Bolzano) view, as we saw above in section 3, gains great simplicity by reducing the correctness of the judgement

A is true

to the truth of the propositional content A. When the proposition A is true, Bolzano says, the judgement in question is correct (*richtig*), that is, it is a piece of knowledge (a cognition, an *Erkenntnis*).[45] On this reading, judgements which are, in the apt terminology of Brentano, *blind*, that is, unwarranted, are still held to be knowledge, simply in virtue of having a (classically) true proposition as content. An example would be a judgement made, completely without warrant, by hazarding a mere guess as to the number of window-panes in the Leyden City Hall, say, 8548, and hitting bull's eye by fluke. To my mind, blind *knowledge* is to high a price to pay for the Bolzano-reduction of judgemental correctness to propositional truth, since,

> opinions divorced from knowledge, are ugly things[.] The best of them are blind. Or do you think that those who hold some correct opinion without evidence differ appreciably from blind men who go the right way?[46]

Accordingly, I prefer the opposite route, explaining propositional truth as a particular form of judgement, and the judgemental correctness as evidenceability.

7. VALIDITY OF INFERENCES.

Corresponding blindness phenomena may occur also at the level of inference. Here, the classical notion of validity is applicable to an inference(-mode) of the form I':

$$\frac{A_1 \text{ is true} \qquad A_2 \text{ is true} \qquad \ldots \qquad A_k \text{ is true}}{C \text{ is true}}$$

Such an inference I is valid, or so Bolzano says, when a relation of logical consequence – *eine Ableitbarkeit* – obtains between the propositions that serve as premisses, respectively conclusion, of the inference in question, that is, when the consequence

$A_1, A_2, \ldots, A_k \Rightarrow C$

obtains logically. In this Bolzano was followed by virtually the entire modern tradition in classical logic. Similar accounts of validity can be found in Wittgenstein's *Tractatus*, as well as in Tarski's (1936) account of logical consequence, whose current model-theoretic (Tarski-Vaught (1957)) version can be found in any decent text-book.[47] However, also here the price paid for the ensuing simplicity is high. The key notion in the explanation of the (logical) holding of a consequence is that of truth under a variation (or truth in a model). A consequence holds logically if every variation that serves to make all the antecedent propositions true also makes the consequent proposition true. In fact, a consequence holds if the corresponding implicational proposition is true, and it holds logically if the implication is a logical truth. Accordingly, just as the reduction of judgemental correctness to propositional truth allowed for judgements that were blindly correct, so does the corresponding reduction of inferential validity to the (logical) holding of consequences allow for blindly correct inference. We get a similar epistemological slack between what is theoretically permissible and what is epistemically warranted: under the Bolzano reduction, the inference may be valid, even though no epistemological warrant has been offered in order to make the conclusion evident. Thus, under the classical Bolzano reduction of validity, we could find ourselves in the position that we knew the premisses of an inference, and, furthermore, that, unknowingly to us, logical consequence does obtain between the relevant propositional contents of premisses and conclusion. In such a position one would be allowed to carry out the inference – because under the Bolzano reduction the inference *is* valid – but still we would not know the conclusion. This situation would be an example of an unknown conclusion that is validly drawn from known premisses. We would have a (mediate) act of knowledge, in which all the premisses were known and the inference valid, according to the appropriate, Bolzano-reduced notion of validity, but which would not make its object evident.

It now remains to offer a constructivist account of validity that does not suffer from the shortcomings of the Bolzano reduction. The blindness-phenomena that impugn the Bolzano-reduced notions of validity and (judgemental) correctness have their origin in the circumstance that propositional truth is not primarily epistemic. Evidence is conferred upon what is known, namely a certain statement (judgement), by the act of knowing. In the case of an immediate act, the statement must be evidenceable from itself: the knowledge is intuitive rather than discursive. In a mediate act, discursive knowledge is inferred, that is, is drawn as a conclusion, from certain evident judgements. Thus, what is called for, in a constructivist elucidation of inferential validity, is not preservation of propositional truth, but transmission of judgemental evidence from statement(s) to statement. This leads straightforwardly to a resurrection of the old idea that the validity of an inference resides in the analytic containment of the conclusion in the premisses. Thus we say that the inference I

$$\frac{J_1 \quad J_2 \quad \cdots \quad J_k}{J}$$

is valid when a chain of evidence-preserving steps Σ_1, Σ_2, ...,Σ_m can be given, which links premisses and conclusion, and where each Σ_i is either an axiom, that is, a self-evident (immediate) judgement, or an immediately valid inference, that is, an inference the evidence-preservingness of which rests in the nature (essence) of the concepts that are used the inference in question, and which, accordingly, neither is capable, nor is in need, of further justification in terms of other inferences.[48] It is interesting to note that this notion of validity, in terms of chains of immediate evidences, crops up now and then, even in modern mathematical logic. Thus Gödel holds that 'the chain of definitions of the concepts occurring in the theorem together with certain axioms about the primitive terms forms by itself a proof, i. e., an unbroken chain of immediate evidences'.[49] Similarly, according to H. B. Curry's description of the intuitionist position,

> 'a proof is valid when it is a construction the individual steps of which are immediately evident; no matter what rules are given, a valid proof can be found which does not conform to them'.[50]

The notion of validity that was discussed above pertains to (modes of) inference. However, in the quote from Curry, another notion occurs. There, what is at issue is the notion of a valid *proof.* We have encountered three notions of proof that have to be carefully kept apart, namely,

(i) proof(-act)s of certain theorems;
(ii) proof(-trace)s of such acts, that is, demonstrations in mathematical texts;
(iii) proof(-object)s of propositions.

The second of these is the natural carrier of the above notion of validity. Proof(-trace)s are blueprints for, in general, discursive, mediate proof-acts. When a trace is valid, that is, when all the axioms that occur in the trace are (self-)evident, and all the inference(-modes) that occur are valid, then an act carried out according to the trace confers evidence upon its conclusion.

Rightness -*rectitudo* – is the relevant notion for the level of (proof- and other) acts as was already noted.[51] It is *sui generis* and is needed to account for the possibility of error. Without the notion of rightness, error would be an empty notion.

Of course, we can also speak of right proof-*objects*. This would yet again be an application of the notion of rightness, this time in the form of truth of things: concerning a judgement of the form

$$c \in \text{proof}(A)$$

the question may arise whether it really is evident and whether c really does belong to proof(A). Confronted with such a situation, we are perhaps able, after discussion and evaluation, checking each construction-step that has been used in synthesizing the putative proof c, to satisfy ourselves that we were not mistaken:

c is a *right* proof(-object) of A.

8. ANALYTICITY AND THE GÖDEL INCOMPLETENESS THEOREMS

Above, axioms were called analytic and the validity of inference was explained in terms of 'analytic containment' between conclusions and premisses.[52] From Aristotle's *Posterior Analytics*, with its treatment of *per se* predications, the notion of analyticity has had a central role in epistemology. Medieval epistemology, for instance, in Aquinas and Duns Scotus, treats of demonstrative knowledge in terms of "self-evident" judgements, that is, statements that are knowable in themselves. In fact, the explanation offered by St. Thomas Aquinas for the notion of a *propositio per se nota* is the same as that of Kant for the notion of an analytic judgement:[53] An S-is-P judgement is knowable *per se* when the predicate P is contained in the essence (or concept) of the subject S, so that knowledge of the definitions of S and P, that is, knowledge of their essences, suffices for knowledge of the judgement itself.[54] On the road to Kant, one encounters the trifling propositions of Locke, as well as the deviant variation that was adhered to by Leibniz, according to whom *all* truth is analytic. However, even Leibniz does not fall into the trap of making all of (analytic) truth knowable *per se*. Any S-is-P truth has an *a priori* proof that is obtained by resolving the terms S and P to their essential constituents. Owing to the analyticity of the truth in question the resolution has to stop in identities and the result is a proof when read in the opposite direction. However, only in the case of a truth of reason is the *priori* proof a *finite* one. In the case of a Leibnizian truth of fact, on the other hand, the process of resolution will, in general, not terminate after a finite number of steps, whence only the infinite mind of God is capable of taking in the *a priori* proof. Hence, according to Leibniz, other means, and not merely those present in the terms themselves, are required for us finite minds in order to know such truths of fact.

Martin-Löf (1994) notes that his type-theoretical judgements of the two forms

$a{:}\alpha$, that is, a is of object of type α,

and, where a and b are both object of type α,

$a{=}b{:}\alpha$, that is, a and b are equal objects of type α,

have the required analytic character. It is enough to have 'a' and 'α' in order to be able to decide whether $a{:}\alpha$ (and similarly for judgements of equality. For both kinds of judgement, the means of decision utilizes evaluation to, and inspection of, relevant canonical forms). On the other hand, judgements of the form

A is true,

that is, of the form

proof(A) exists,

are synthetic, since they cannot be known merely from their own formulation, but demand a construction, or *synthetization*, of a proof(-object). Thus, the truth of a mathematical proposition, indeed of '*ein jeder Existentialsatz*', is synthetic. On the other hand, that the proof(-object) is a proof(-object) of what it proves is something which can, in order to speak with Wittgenstein's *Tractatus*, be read off analytically *am Symbol allein.*[55]

A beautiful feature of Martin-Löf's view is that every synthetic truth is *grounded* in an analytic judgement: when a proposition C can be known to be true, that is, when we can know the judgement

C is true,

we can also know of a certain construction c, such that the judgement

c is a proof of the proposition C

is analytically correct, that is, can be known mechanically from the symbol alone.

In general, since it is a question of meaning, it is analytic (*ex vi terminorum*) that a certain proposition is made true by a certain kind of truth-maker. Whether such a truth-maker exists, on the other hand, is something that demands amplification (*Erweiterung*) of our knowledge, rather than mere elucidation (*Erläuterung*). Something of the this sort, of course, holds also for the classical truth-maker analysis in Wittgenstein's *Tractatus*. That a proposition (*sinnvolle Satz*) A is true cannot be known *a priori*, but demands comparison with the world; it must be checked that the presented state of affairs (*Sachverhalt*) S_A does indeed obtain. The relation between the proposition A and the state of affairs S_A that it presents, on the other hand, is internal. Thus, what a truth-maker is for A is internally determined from A, whereas the question of the existence of such a truth-maker is a material one, that is, one that cannot be answered merely from the symbol alone.

In Wittgenstein's *Tractatus*, as well as for Bolzano, the important notion is not that of an analytic judgement.[56] Instead the notion of a logically true *proposition*, which Wittgenstein calls *tautology*, and Bolzano *logically analytic proposition* (or 'analytic in the narrow sense'), that is, a proposition which is true, come what may, independently of what is the case, holds pride of place.[57] In the *Tractatus* Wittgenstein transfers the demand of *per se* recognizability from the notion of an analytic judgement to the notion of a logical truth: it must be possible, by mechanical calculation on the symbol alone, to determine whether the proposition is a tautology.[58] As noted by Wittgenstein, the decision method offered in the Tractatus, however, is applicable only in the case of quantifer-free propositions.[59] The undecidability of predicate logic, finally, that was established by Alonzo Church (1936), made it an illusionary hope that such a method could be found for the whole of language: in general, the notion of logical truth for propositions containing multiple generality, that is, occurrences of the quantifier combinations $\forall\exists$ and $\exists\forall$, is recursively undecidable. The logic of judgements of the two categorical forms a is an object of type α, respectively, a and b are equal objects of type α

is decidable, whereas the logic of judgements of the form

A is true, that is, proof (A) exists,

is undecidable, in virtue of Church's theorem.

How do matters stand with respect to the other great limitative theorem, namely Gödel's (1931) Incompleteness Theorem? Let us attempt to transpose the Gödel theory to the present framework. A system S of rules for generating proof(-objects) is consistent, if the judgement

$t \in \bot$

can be derived for no term t from the rules of S. Consider now a consistent system S that comprises a modicum of arithmetic, say, in the form of construction-rules for proof-objects corresponding to the natural deduction rules

&I, &E,□ ∨I, ∨E, ⊃I, ⊃E, ⊥E, ∀I, ∀E, ∃I, ∃E,□ IdI, IdE, NI, NE,

The rule IdI for identity (among the elements of the set A) takes the form

$$\frac{a \in A}{r(a) \in \mathrm{Id}(A, a, a),}$$

and IdE is the corresponding elimination rule. NI, on the other hand, is the introduction rule that generates the canonical forms of numbers, and NE allows for proofs by means of mathematical induction over the set N, by means of permitting the definition of functions by recursion. The work of Gödel, when transposed to the present framework, shows that, by inspection of the rules of S, we can explicitly indicate a *true* proposition G_S of L_S such that for no term t of the language L_S of the system S can the judgement

$t \in \mathrm{proof}(G_S)$

be derived from the rules of S, even though the proposition G_S is formulated using concepts in L_S only. However, the truth of G_S can be demonstrated in a suitable conservative extension S' of the system S, where a term t' can be found, together with a demonstration of the judgement

$t' \in \mathrm{proof}(G_S)$

from the rules of S'. The true proposition G_S is itself arithmetical, that is, formulated in terms of purely arithmetical concepts, but its proof can be obtained only using non-arithmetical concepts: the term t' cannot be formed in the language L_S of the system S, but only in the extended language $L_{S'}$ of the system S'. The Gödel Incompleteness strikes, not at the complete logic of analytic judgements, but against the incomplete logic of propositional truths, that is, judgements of the form

proposition A is true.

Let S_{PROP} be the system of propositional truths that is obtained by stripping off the proof-objects from the theorems of S, that is, by replacing the S theorem

$c \in \text{proof}(A)$

by the truncated S_{PROP} theorem

A is true.

The Gödel theorem then says that, for certain systems S and S', the matching system of propositional truths S'_{PROP} will not extend S_{PROP} conservatively, even though the system S' is a conservative extension of S.[60] Gödel shows that the logic of propositional truth is incomplete. For analytic judgements, on the other hand, where the proof(-object)s of propositions have not have not been suppressed in the theorems, completeness does hold. If an analytic judgement J of the form

$a \in \alpha$

is formulated in a language L_S, one will find a demonstration of J by means of applying the introduction and elimination rules of proof-construction from the construction-principles in S backwards.

Here, I think, lies a definite advantage of the constructivist position. In the *Tractatus* Wittgenstein rejects in scornful terms any use of 'das Einleuchten' – (self-)evidence – in logic.[61] He has, however, not taken proper notice of the fact that his demand that the propositions of logic be mechanically decidable am *Symbol allein* is nothing but a variant of the traditional demand that the primitive propositions of logic be knowable *per se*, that is, that they can be made evident from themselves and do not demand an external comparison with the world. Just as Kant's analytic judgements they offer no extension but only elucidation of our knowledge. Wittgenstein simulataneously both rejects and imposes this demand for analytical self-evidence, whence his position becomes impossible. The constructivist epistemological alternative that I have been concerned to outline in the present Chapter, on the other hand, suffers no ill fate at the hand of the Incompleteness and Undecidability theorems, and this, to my mind, constitutes a powerful argument in its favour.

Göran Sundholm
Leyden University

NOTES

[1]See for instance the essays collected in his (1978) and (1993), as well as the synoptic (1991).

[2] *Grundgesetze*, I, § 32, respectively *Tractatus* 4.024. David Wiggins (1980), (1992), (1997), in particular, has persistently explored the possibilities of *this* truth-conditional paradigm.

[3] *Evidence* is the quality that pertains to what is evident, and it is commonly expressed in terms of the Cartesian 'clear and distinct'. This evidence *of* what is evident must be distinguished from the evidence *for* an opinion. (The latter notion is *not* at issue in the present chapter.) The *locus classicus* for the evidence theory of truth is Brentano (1930, IV). *'Das Problem der Evidenz'*, that is, Stegmüller (1954, Ch. I.4), is an excellent introduction to the role of evidence in epistemology. Patzig (1971) is also illuminating, whereas Schlick (1910) offers a critical exposition.

[4] I am here indebted to Per Martin-Löf (1998).

[5] See Olson (1967).

[6] St. Augustine's *Soliloquies* and St. Anselm's *De veritate* are the prime sources concerning the notion of *rectitudo.*

[7] Fichte (1797).

[8] The act/object dichotomy raises the issue of the corresponding correctness notions: as a rule I shall reserve *right* for acts and *correct* for objects, for example, "the object of a right act of judgement is correct".

[9] See, respectively, Brentano (1930, Section IV), Patzig (1967), as well as Putnam (1981) and many later writings.

[10] Duns Scotus (1987, p. 106 and p. 126, respectively). The fascinating scholastic teaching on these matters is admirably treated by Vier (1951).

[11] Quine(1951, p. 21 and p. 22).

[12] In the German tradition, for instance, in Frege, one finds the distinction between *Schluss* (act) and *Schlussweise* (mode of inference).

[13] See Martin-Löf (1985) and other works listed in its Postscript.

[14] The notion of an assertoric utterance is here the prior one. The assertoric utterances of declaratives are delineated by means of the criterion involving the legitimacy of the counter-question (2): inquiry as to *how* the utterer knows is legitimate after an assertoric utterance and in other cases not.

[15] The third person propositional claim *Göran Sundholm knows that snow is white* is different from the sense of (3) that is here at issue. In order to understand this, the Moorean paradoxes which lurk around the corner might be reflected upon: since I might forget, or be otherwise confused about my identity, an assertion by me of

Snow is white, but Göran Sundholm does not believe it

is not paradoxical, whereas an assertion by me (or anyone else) effected by means of an assertoric utterance of

Snow is white, but I do not believe it

is. Under the analysis in the text, in both cases, we get the illocutionary knowledge claim

I know that snow is white,

which, together with,

I do not believe it,

does yield a Moorean paradox. My assertion of

Göran Sundholm does not believe it,

on the other hand, does not yield a paradox, unless the additional claim that
I know that I am Göran Sundholm
is also given.

[16] In order to avoid offensive iterations of *that* it is often convenient to use the form (6).

[17] *Warning*. Ever since Cook Wilson the term *statement* has been overburdened in Oxford philosophy. My 'statements' do not coincide with Dummett's: *his* statements come close to my propositions, but are also intended to take indexicality into account, a topic that I prefer to leave out of consideration. My use is also different from that of Frege. For Frege, the declarative expresses a proposition (*Gedanke*) and this proposition may or may not be asserted. Given that it is the nominalizations 'that S' of declaratives S that stand for propositions, I hold that Frege is wrong in this and that propositions are not *behauptungsfähig*. On the contrary, it is the statement expressed by the declarative S, namely the statement that it is true that S, that is capable of being asserted.

[18] Bolzano (1837, § 34), and Frege (1892, p.34, fn. 7).

[19] Independently of the chosen epistemological position, a mistake is an act of knowledge which is not *right*. Under the realist reductions the rightness of acts of knowledge is reduced to the correctness of their products and that in turn to the truth of the propositional contents. But from a realist point of view propositional truth is bivalent: the proposition A either is, or is not, true, *tertium non datur*. From the constructivist point of view (which opts for the opposite alternative in the Fichtean dichotomy and upholds the primacy of acts), on the other hand, the *rightness of acts* is primitive, *sui generis*.

[20] This theme is worked out in some detail in my (1994). Niniluoto (1997) argues against some of the conclusions I drew there.

[21] The example is taken from Rogers (1967, p. 9-10). It is manufactured for a purpose, but it is not farfetched. Compare, for instance, the analogous example of Dirichlet, concerning the characteristic function of the rational numbers within the reals, which is highly significant from a mathematical point of view.

[22] The use of the set-theoretic $\in$ in place of the type theoretic colon : is natural when the type in question is also a set.

[23] These canons are well set out in Dubislav (1931, §14).

[24] I have learned this way of presenting Brouwer's argument from Aarne Ranta (1994, Chapter 2.14, pp. 37-38). With the benefit of hindsight I can find it already in Martin-Löf (1985, p. 33).

[25] There is, of course, nothing special about the *universal* quantifier: the analogous way of construing the existential quantifier produces the same quandaries.

[26] See, in particular, 'The philosophical basis of intuitionistic logic' (1975) and 'What is a theory of meaning? II' (1976), reprinted respectively in (1978) and (1993). There is, however, no consensus even as to how, precisely Dummett's argument goes; a massive scholarly debate has arisen, to which my (1987) is a relatively early contribution.

[27] The transition is beautifully described by Warren Goldfarb (1979).

[28] The distinction between the two logical paradigms was introduced by Jean van Heijenoort (1967), (1976). Jaakko Hintikka (1996) has tirelessly explored its possibilities.

[29] Gödel's (1944) essay for the Schilpp volume on Russell deals with these issues in considerable depth.

[30] Thiel (1988) and Franchella (1994) survey the debate in question.

[31] Detailed arguments concerning the equivalence between the formulations in terms of propositions that express intentions towards constructions and in terms of problems that require solutions can be found in my (1983, pp. 158-9), and (1997, p. 196).

[32] The notion of trace is dealt with in considerable detail in my (1993), (1997), (1998) and (forthcoming).

[33] These formulations are not taken directly from Heyting, but are inspired by formulations used by Martin-Löf (1984). A proof of a mathematical theorem (that we can be found in a certain mathematical text) is such a proof-trace. It can be used as a "blue-print" for proof-acts by other mathematicians in order to get to know the theorem in question. Other examples are the scores of chess-games (which can be used by other players to imitate opening novelties etc.) and, of course, scores of music.

[34] Martin-Löf (1987) discusses the significance of this fact.

[35] See Prawitz (1971, §§ 3.3-3.5), Dummett (1977, Ch. 4), or Tennant (1978, § 4.10, §5.4).

[36] For further information concerning this truth-maker perspective on the correspondence theory of truth, see Mulligan, Simons and Smith (1984). I have spelled out some consequences of adopting this perspective to the constructive notion of truth in my (1994).

[37] For which regress, see my (1983, p. 162).

[38] I have dealt with this notion at some length in (1994a).

[39] Brentano (1956, §27).

[40] Brentano (1956, §30). These reductions were known also to Leibniz and to Bolzano. Bolzano, however, used his standard form of judgement A is true, and the reductions in terms of the *Gegenständlichkeit einer Vorstellung* (exemplification of a concept) were carried out in the *Sätze an sich* (propositions) that serve as contents of his judgements.

[41] All "empty" propositions, that is, propositions with no proof-objects are (materially) equivalent (while false), but they need not be logically equivalent, nor are they identical propositions. For instance, the (sets of proofs of the) propositions $\perp$ and $A\&\neg A$ are both empty. (In this, and some subsequent examples, I find it convenient to identify the proposition A and the set proof(A), so as not to overburden the notational patience of the reader.) The propositions are not identical, though. For the propositions A and B to be identical the inferences from the judgement $a \in A$ to the judgement $a \in B$ (where a is canonical), and conversely, must be immediate from the meaning explanations of the propositions in question. In the example just given, by stipulation $\perp$ has no canonical proofs, whereas a canonical proof of $A\&\neg A$ has to be an ordered pair <a,b> the first component a of which is a proof of A and the second component b is a proof of $\neg A$. Clearly, then, these are not identical propositions. Applying the proof object b to a one obtains $\mathrm{ap}(b,a) \in \perp$, which is impossible since, according to its meaning explanation, $\perp$ has no canonical proofs (and so no proofs at all). For more discussion, see my (1994b).

[42] For plausibility, and feasibility, the numbers considered should be taken below, say, 10 000.

[43] Brentano (1930, p. 139) and (1956, §42). Note how the standard regress arguments (*Dialelle*), e. g. Kant (1800, Ch. VII) and Frege (1918, p. 60), against the correspondence theory are obviated here by letting the judgement correspond to another judgement.

[44] Brentano (1956, §35, p. 143). (My) English translation:

Error is precluded with evidence. Also doubt is precluded with evidence, but the judgement is not made evident by freedom from error, or by freedom from doubt, but by a peculiarity that characterizes it as correct.

[45] *WL* § 34, 3, a: 'Jedes Urteil enthält einen Satz, der entweder der Wahrheit gemäss ist oder ihr nicht gemäss ist; und in dem ersten Falle heisset das Urteil ein richtiges, im zweiten ein unrichtiges.' *WL* § 36: (Bolzano) 'versteht unter dem Worte Erkenntnis ein jedes Urteil, das einem wahren Satz enthält.'

[46] Thus Socrates in Plato's *Republic*, 506c. I am indebted to Per Martin-Löf for drawing my attention to this splendid passage.

[47] Frege is the only prominent exception to this almost universal acceptance of the Bolzano-reduction of the validity of an inference between statements to the logical holding of

a matching consequence-relation between the propositional contents of the statements in question, cf. Currie (1987).

[48] I discuss this notion of validity, and its roots in medieval logic, in my (1998) and (1998a).

[49] (1972, p. 275, fn. h).

[50] (1963, p. 10). (On my reading Curry conflates proofs as acts and proofs as objects.) The final part of the quote attempts to find a place within constructivism for the effects of Gödel's incompleteness theorem. Martin-Löf (1994) gives an account of Gödel incompleteness for the constructivist framework, and some details can be found in section 8 below.

[51] See section 1 above.

[52] This section draws heavily on Martin-Löf (1994), and in some measure also on Sundholm (1990), (forthcoming).

[53] *Summa contra Gentiles*, Ch. X, and *Summa Theologica*, QII.1, respectively, *K.d.r.V.* B6.

[54] I am not unaware that several Quinean (1951) bullets are being bitten here.

[55] 6.113. Wittgenstein's formulation – *am Symbol allein* – recalls the scholastic *ex vi terminorum* (or *ex terminis*). (See footnote 10 above.) In my (1990) the analogy is noted between

(i)	Kant	The judgement S is P is analytic, that is, the predicate P is contained in the concept of the subject S
(ii)	Wittgenstein	P is a formal (or internal) property (feature) of a
(iii)	Martin-Löf	a:α, that is, a is an object of type α

[56] Indeed, the notion of judgement as such gets very short shift in the Tractatus; in 4.442 Frege's *Urteilsstrich* – in fact the combination of the *Urteils-* and *Inhalts-striche* – is dismissed as being entirely without logical significance.

[57] *Tractaus* 4.46, respectively, WL § 148(3).

[58] 6.11, 6.113.

[59] 6.1203.

[60] Formal theories in so called standard first order fomalization are, from the present meaning-theoretical perspective, obtained by the step from S to S_{PROP}. In the process of jettisoning the proof-objects much information is lost. For certain purposes, this is of no consequence. Sometimes, however, unwanted phenomena arise, such as in case of the Gödel theorem, where a conservative extension including proof-objects, is changed into a non-conservative extension by suppressing them.

[61] 5.1363, 5.4731, 6.1271.

REFERENCES

Bolzano, B.: 1837, *Wissenschaftslehre*, J. von Seidel, Sulzbach.

Brentano, F.: 1930 (1974^{II}), *Wahrheit und Evidenz*, O. Kraus (ed.), Felix Meiner Verlag, Hamburg.

Brentano, F.: 1956, Die *Lehre vom richtigen Urteil*, F. Mayer-Hildebrand (ed.), A. Francke Verlag, Bern.

Brouwer, L. E. J.: 1908, 'De onbetrouwbaarheid der logische principes', *Tijdschrift voor Wijsbegeerte* **2**, 152-158.

Church, A.: 'A note on the Entscheidungsproblem', *Journal of Symbolic Logic* **1**, 40-41, corrections 101-102.

Currie, G.: 1987, 'Remarks on Frege's Conception of Inference', *Notre Dame Journal of Formal Logic* **28**, 53-68.

Curry, H. B.: 1976 (1963^{I}), *Foundations of Mathematical Logic*, Dover Pub., New York.

Diller, J. and A. Troelstra: 1984, 'Realizability and intuitionistic logic', *Synthese* **60**, 253-282.

Dubislav, W.: 1931^{III} (1981^{IV}), *Die Definition*, Felix Meiner Verlag, Hamburg.

Dummett, M.: 1977, *Elements of Intuitionism*, Oxford U. P.

Dummett, M.: 1978, *Truth and Other Enigmas*, Duckworth, London.

Dummett, M.: 1991, *The Logical Basis of Metaphysics*, Duckworth, London.

Dummett, M.: 1993, *The Seas of Language*, Duckworth, London.

Duns Scotus, J.: 1987, *Philosophical Writings*, A. Wolter and O. F. M. (eds.), Hackett Pub., Indianapolis.

Fichte, G.: 1797, 'Erste Einleitung in die Wissenschaftslehre', *Philosophisches Journal* **5**, 1-47.

Franchella, M.: 1994, 'Heyting's Contribution to the Change in Research into the Foundations of Mathematics', *History and Philosophy of Logic* **15**, 149-172.

Frege, G.: 1892, 'Über Sinn und Bedeutung', *Zeitschrift für Philsophie und philosophische Kritik* **100**, 25-50.

Frege, G.: 1893, 1903, *Grundgesetze der Arithmetik*, Band I, Band II, H. Pohle, Jena.

Frege, G.: 1918, 'Der Gedanke', *Beiträge zur Philosophie des deutschen Idealismus* **2**, 58-77.

Gödel, K.: 1931, 'Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme I', *Monatshefte für Matematik und Physik* **38**, 173-198.

Gödel, K.: 1944, 'Russell's mathematical logic', in P. A. Schilpp (ed.), *The Philosophy of Bertrand Russell*, Library of Living Philosophers, Evanston, pp. 123-153 and in Gödel, 1990, pp. 119-143.

Gödel, K.: 1972, 'On an extension of finitary mathematics which have not yet been used', in Gödel, 1990, pp. 271-280.

Gödel, K.: 1990, *Collected Works, Vol. II*, Oxford U. P., New York.

Goldfarb, W.: 'Logic in the twenties: the nature of the quantifier', *Journal of Symbolic Logic* **44**, 1979, 351-368.

Hale, B.: 1997, 'Realism and its oppositions', in B. Hale and C. Wright, *A Companion to the Philosophy of Language*, Blackwell, Oxford, pp. 271-308.

Heyting, A.: 1930, 'Die formalen Regeln der intuitionistischen Logik', *Sitzungsberichte der preussischen Akademie von Wissenschaften*, Phys.-math. Klasse, pp. 42-56. English translation in P. Mancosu (ed.), *From Brouwer to Hilbert*, Oxford U. P., 1997, pp. 311-327.

Heyting, A.: 1930a, 'Sur la logique intuitionniste', *Acad. Roy. Belgique, Bull. Cl. Sci.*, V, **16**, 957-963; English translation in P. Mancosu (ed.), *From Brouwer to Hilbert*, Oxford U. P., 1997, pp. 306-310.

Heyting, A.: 1931, 'Die intuitionistische Grundlegung der Mathematik', *Erkenntnis* **2**, pp. 106-115. English translation in P. Benacerraf and H. Putnam, *Philosophy of Mathematics* (second edition), Cambridge U. P., Cambridge, 1983, pp. 52-61.

Heyting, A.:1934, *Mathematische Grundlagenforschung. Intuitionismus. Beweistheorie*, Julius Springer, Berlin.

Heyting, A.: 1958, 'Intuitionism in mathematics', in Raymond Klibanski (ed.), *Philosophy in the Mid-Century*, La nouva editrice, Florence, pp. 101-115.

Hintikka, J.: 1988, 'On the development of the model-theoretic viewpoint in logical theory', *Synthese* **77**, 1-36.

Hintikka, J.: 1996, *Lingua Universalis vs. Calculus Ratiocinator: An Ultimate Presupposition of Twentieth-Century Philosophy*, Kluwer, Dordrecht.

Howard, W.: 1969, 'The fomulae-as-types notion of construction', in Seldin, Jonathan, and Roger Hindley (eds.), *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus and Formalism*, Academic Press, London, 1980, pp. 479-490.

Kant, I.: 1800 (1904$^{\text{III}}$), *Logik*, B. Jäsche (ed.), Felix Meiner Verlag, Leipzig.

Kolmogoroff, A. N.: 1932, 'Zur Deutung der intuitionistichen Logik', *Mathematische Zeitschrift* **35** (1932), 58-65; English translation in P. Mancosu (ed.), *From Brouwer to Hilbert*, Oxford U. P., 1997, pp. 328-334.

Martin-Löf, P.: 1984, *Intuitionistic Type Theory*, Bibliopolis, Naples.

Martin-Löf, P.: 1985, 'On the meanings of the logical constants and the justifications of the logical laws', in *Nordic Journal of Philosophical Logic* **1** (1996), 11-60; originally delivered in 1983 and published in 1985; electronically available at http://www.hf.uio.no/filosofi/njpl/.

Martin-Löf, P.: 1987, 'Truth of a proposition, evidence of judgement, validity of a proof', *Synthese* **73**, 191-212.

Martin-Löf, P.: 1994, 'Analytic and synthetic judgements in type theory', in P. Parrini (ed.), *Kant and Contemporary Epistemology*, Kluwer, Dordrecht, pp. 87-99.

Martin-Löf, P.: 1995, 'Verificationism then and now', in Schimanovich, W. De Pauli, E. Köhler, and F. Stadler (eds.), *The Foundational Debate: Complexity and Constructivity in Mathematics and Physics*, Kluwer, Dordrecht, pp. 187-196.

Martin-Löf, P.: 1998, 'Truth and knowability: on the principles C and K of Michael Dummett', in H. G. Dales and G. Oliveri (eds.), *Truth in Mathematics*, Clarendon Press, Oxford.

Mulligan, K., P. Simons, and B. Smith: 1984, 'Truth-Makers', *Philosophy and Phenomenological Research* **44**, 287-321.

Niniluoto, I.: 1997, 'Tarskian truth as correspondence – replies to some objections', in J. Peregrin (ed.), *The Nature of Truth (If Any). Proceedings of the International Colloquium, Prague, September 17-20, 1996*, Filosofia, Prague, 1997, pp. 153-160.

Olson, R. G.: 1967, 'Nihilism', in P. Edwards (ed.), *The Encyclopaedia of Philosophy*, Macmillan, New, York.

Patzig, G.: 1971, 'Kritische Bemerkungen zu Husserls thesen über das Verhältnis von Wahrheit und Evidenz', *Neue Hefte für Philosophie*, Heft **1**, Vandenbroek und Rupprecht, Göttingen, 12-32; translated into English by J. N. Mohanty as 'Husserl on truth and evidence', in J. N. Mohanty (ed.), *Readings on Husserl's Logical Investigations*, Nijhoff, The Hague, 1977, pp. 179- 196.

Prawitz, D.: 1971, 'Ideas and results in proof theory', in J. E. Fenstad, *Proceedings of the Second Scandinavian Logic Symposium*, Amsterdam, North-Holland, pp. 235-307.

Putnam, H.: 1981, *Realism and Reason*, Cambridge U. P., Cambridge.

Quine, W. V. O.: 1951, 'Two dogmas of empiricism', reprinted in *From a Logical Point of View*, Harper & Row, N. Y., 1961$^{\text{II}}$, pp. 20-46.

Ramsey, F. P.: 1925, 'The foundations of mathematics', *Proceedings of the Aristotelian Society* **25**, 338-384.

Rogers, Jr., H.: 1967, *Theory of Recursive Functions and Effective Computability*, McGraw-Hill Book Company, New York.

Schlick, M.: 1910-11, 'Das Wesen der Wahrheit nach der modernen Logik', *Vierteljahrschrift für wissenschaftliche Philosophie und Soziologie* **34-35**, 386-477; reprinted in M. Schlick, *Philosophische Logik*, B. Philippi (ed.), Suhrkamp Taschenbuch, 1986.

Stegmüller, W.: 1954, *Metaphysik, Wissenschaft, Skepsis*, Humboldt-Verlag, Vienna.

Sundholm, G.: 1983, 'Constructions, proofs and the meaning of the logical constants', *Journal of Philosophical Logic* **12**, 151-172.

Sundholm, G.: 1987, 'Proof theory and meaning', in D. Gabbay and F. Guenthner, *Handbook of Philosophical Logic*, Vol. III, Reidel, Kluwer, pp. 471-506.

Sundholm, G.:, 1990, '*Sätze der Logik*: an Alternative Conception', in R. Haller and J. Brandl (eds.), *Wittgenstein – Towards a Re-Evaluation. Prroceedings of the 14th International Wittgenstein-Symposium Centenary Celebration*, Verlag Hölder-Pichler-Tempsky, Wien, pp. 59-61.

Sundholm, G.: 1993, 'Questions of Proof', *Manuscrito* (Campinas, S. P.), *16*, pp. 47-70.

Sundholm, G.: 1994, 'Ontologic versus epistemologic', in D. Prawitz and D. Westerståhl (eds.), *Logic and Philosophy of Science in Uppsala*, Kluwer, Dordrecht, pp. 373-384.

Sundholm, G.: 1994a, 'Existence, proof and truth-making: a perspective on the intuitionistic conception of truth', *TOPOI* **13**, 117-126.

Sundholm, G.: 1994b, 'Proof-theoretical semantics and Fregean identity criteria for propositions', *The Monist* **77**, 294-314.

Sundholm, G.: 1997, 'Implicit epistemic aspects of constructive logic', *Journal of Logic, Language, and Information* **6**, 191-212.

Sundholm, G.: 1998, 'Inference, Consequence, Implication', *Philsophia Mathematica* **6**, 178-194.

Sundholm, G.: 1998a, 'Inference versus Consequence', in *Logica Yearbook 1997* (T. Childers, ed.), Philsophia Publ., Czech Academy of Science, Prague, pp. 26-35.

Sundholm, G., forthcoming: 'Proofs as Acts versus Proofs as Objects: Some Questions for Dag Prawitz', to appear in a special issue of *Theoria* dedicated to the work of Dag Prawitz.

Tarski, A.: 1936, 'Über den Begriff der logischen Folgerung', *Act. Congr. Phil. Sci.* (Paris), **VII**, *Actualités Scientifiques et Industrielle* **394**, 1-11.

Tarski, A. and R. Vaught, 'Arithmetical Extensions of Relational Systems', *Compositio Matematica* **13**, 81-102.

Tennant, N.: 1987, *Anti-Realism and Logic*, Clarendon Press, Oxford.

Tennant, N.: 1997, *The Taming of the True*, Clarendon Press, Oxford.

Thiel, C., 1988, 'Die Kontroverse um die intuitionistische Logik vor ihre Axiomatisierung durch Heyting im Jahre 1930', *History and Philosophy of Logic* **9**, 67-75.

van Heijenoort, J.: 1967, 'Logic as calculus versus logic as language', *Synthese* **17**, 324-330.

van Heijenoort, J.: 1976, 'Set-theoretic semantics', in R. O. Gandy and M. Hyland (eds.), *Logic Colloqium '76*, North-Holland, Amsterdam, pp. 183-190.

Vier, P. C.: 1951, *Evidence and its Function According to John Duns Scotus*, The Franciscan Institute, St. Bonaventure, N.Y.

Weyl, H.: 1921, 'Über die neue Grundlagenkrise der Mathematik', *Mathematische Zeitschrift* **10**, 37-79; English translation in P. Mancosu, *From Brouwer to Hilbert*, Oxford U. P., 1997, pp. 86-118.

Wiggins, D.: 1980, 'What would be a substantial theory of truth?', in Z. van Straaten, *Philosophical Subjects: Essays Presented to P. F. Strawson*, Clarendon, Oxford, pp. 189-221.

Wiggins, D.: 1992, 'Meaning, truth-conditions, proposition: Frege's doctrine of sense retrieved, resumed and redeployed in the light of certain recent criticisms', *Dialectica* **46**, 61-90.

Wiggins, D.: 1997, 'Meaning and truth-conditions: from Frege's grand design to Davidson's', in B. Hale and C. Wright (eds.), *A Companion to the Philosophy of Language*, Blackwell, Oxford, pp. 271-308.

Wittgenstein, L.: 1922, *Tractatus Logico-Philosophicus*, Routledge and Kegan Paul, London.

MARKUS LAMMENRANTA

THEORIES OF JUSTIFICATION

During the past two or three decades, justification has become a central topic in epistemology. The interest in justification grew out of the attempts to give the correct analysis of knowledge in the face of the famous counterexamples given by Edmund Gettier to the so-called traditional conception of knowledge in his 1963 paper. The interest in knowledge may have decreased, but the disputes about the right account of justification go on more vigorously than ever. Indeed, there are at present so many distinct theories of epistemic justification advocated by different disputants that it makes one doubt whether they are actually talking about the same thing at all. Before considering these theories, we shall therefore make first an attempt to locate the common concept or property that they are all theories of.

We need some pretheoretic understanding of the target concept or property if we are to evaluate the different theories. It is often assumed that we already possess such an understanding because we have all learnt the language to which the term 'justification' belongs. That is why we have intuitions about the applicability of the concept that we can use to test the theories. However, it is far from clear that there is any ordinary concept of epistemic justification. The term 'epistemically justified' does not have such a customary use in ordinary language as the term 'to know', as William Alston (1989, p. 5) has pointed out. And even if there were such a concept, many epistemologists would not seem to be interested in it. 'Justification' is a term of art in epistemology.

This makes it even more important to try to identify the concept of justification we are interested in. Unfortunately, there is no unanimity among epistemologists about how to do this. There are at least three different ways of characterising our target concept. All these characterisations need not be incompatible. However, if they motivate competing theories about the substantive conditions of justification, they must be understood as descriptions of distinct concepts. Indeed, the most permanent and fundamental disagreements about the substantive issues become understandable when different conceptions of justification are seen to motivate them. We must therefore acknowledge that there are several distinct concepts of justification and that some competing theories of justification are not actually in conflict at all but are theories about different matters.[1]

I will divide this presentation into two parts that are often called meta-epistemology and substantive epistemology. Meta-epistemology is concerned with the central concepts of epistemology, such as knowledge and justification, trying to give, if not a complete analysis, at least some kind of description of their content. The task of substantive epistemology is to apply these concepts to different kinds of beliefs, trying to determine what knowledge and justified beliefs we have. With respect to justification, this is done by formulating epistemic principles that specify the conditions under which various beliefs qualify as justified. The distinction may also be put by

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 467-497.
© 2004 *Kluwer Academic Publishers.*

saying that while meta-epistemology is concerned with the nature of justification, substantive epistemology tries to specify the criteria of justification.

I will start by looking at those three conceptions of justification and discuss the most important substantive theories after that. Because I will conclude that there are different concepts of epistemic justification in circulation, I will end up by considering what kind of concept or concepts of justification we need in epistemology. This requires that we think over what is the point of having a concept of justification at all. What is the purpose for which we need it? So I will finally address the question concerning the nature of epistemology itself. What is its proper task and what are its most important questions?

I JUSTIFICATION AND KNOWLEDGE

In the *Theaetetus* dialogue, Plato raises the question 'What distinguishes knowledge from true belief?' According to a view that has been popular at least in the twentieth century, the answer is justification. Knowledge is true and justified belief in this view, which is often called the traditional conception of knowledge. It is sometimes even attributed to Plato who suggests in *Meno* (98a) that what turns true belief into knowledge (*episteme*) is the possession of an account (*aitias logismos*) — working out of an explanation. However, for Plato, this account seems to be rather an answer to the question 'What is *X*?' than to the question 'How do you know that *p*?' It is the latter question that is taken by contemporary philosophers to be the central question of epistemology and to which the proper answer is to give a justification for believing that *p*. So it is, at best, controversial to say that Plato himself accepted the traditional conception or analysis of knowledge.[2]

Be the historical truth what it may, it is at least true that several epistemologists in the last century have advocated the traditional analysis of knowledge. They have thought that knowledge is something more than a lucky guess. A belief that is true by accident is not knowledge. What is required is justification. If we accept this, we get a characterisation of justification: it is something that turns true belief into knowledge. The concept of justification can therefore be understood in terms of truth, belief and knowledge. This is appropriate, because justification seems to be the least understood of these four concepts. Theories of justification would thus be answers to the question Plato raises in *Theaetetus*: what distinguishes knowledge from true belief?

Things are not quite so simple. Edmund Gettier (1963) showed by two counterexamples that the traditional analysis of knowledge is not correct: true and justified belief is not sufficient for knowledge. In spite of this, most epistemologists go on taking justification to be a necessary condition of knowledge. What must be done is to add some fourth condition to the traditional analysis to rule out Gettier's original and other similar counterexamples. So most philosophers who have written about epistemic justification in the last forty years have taken it for granted that justification is something that is required for knowledge and that at least contributes to making true belief knowledge.

It is often argued that if justification is to distinguish knowledge from an accidentally true belief, it must be truth-conducive. A justified belief must be probably true. After Gettier, this argument has lost some of its force. If there must be a fourth condition of knowledge, justification need not in itself be truth-conducive. It is only required that justification together with the fourth condition is truth-conducive. So justification may be a purely internal matter and not necessarily conducive to the objective truth. We will discuss this controversy between internalism and externalism in more detail later on.

II THE NORMATIVITY OF JUSTIFICATION

There is another common way of thinking about justification. It emphasises the normative or evaluative character of the concept. To say that a person's belief is justified is to appraise or evaluate it positively. Her holding the belief is right, good, desirable, acceptable, or approvable. This normative character may be understood either in deontological or non-deontological terms. In the former case, justification is explained in terms of obligation, permission and duty. Justification is a matter of fulfilling or not violating an obligation. In the latter case, justification is thought to be just something good, desirable, favourable without being a matter of fulfilling or not violating obligations.[3]

The term 'justification' does seem to have a deontological flavour, and it does seem to suggest that justification is more a matter of permission than obligation. To say that a person is justified in believing that *p* is not to say that she is obligated to believe that *p* but that she is permitted to believe that *p*, that she is not violating any intellectual obligations in believing that *p*.

Permission and obligation are associated with responsibility, blame, praise and other normative consequences of one's situation with respect to fulfilling and violating obligations. To say that a person is justified in her belief is thus to say also that she is not to be blamed for believing so. Her believing is not culpable or blameworthy. It is responsible. So, we can say also that justification, according to this conception, is a matter of epistemic responsibility, of not being a subject to blame. The deontological conception is therefore also called the responsibilist conception of justification.

Alvin Plantinga (1993a, p. 11-14) cites Descartes and Locke as the originators of the deontological tradition in Western epistemology. According to Descartes, we have a duty or obligation not to affirm a proposition unless we perceive it clearly and distinctly. If we affirm something that we do not perceive clearly and distinctly, we are misusing our free will and are guilty and blameworthy for doing so. Locke thinks, on the other hand, that we have a doxastic duty not to affirm a proposition unless we have a good reason for it. So both think that justification is a matter of not violating one's duties and not being subject to blame.

What seems to be wrong with these early formulations of the deontological view is that they are committed to doxastic voluntarism according to which believing is under our direct voluntary control. Namely, if we have an obligation to refrain from believing a proposition whose truth we do not perceive clearly and distinctly or for which we do not have adequate evidence, we must be able to refrain from believing it. This follows from the famous Ought Implies Can Principle. However, it seems that we cannot believe or refrain from believing just by the act of will, by deciding. First of all, beliefs are not acts. They are mental states. They are not something we do. So it may not even make sense to talk about deciding to believe. But even if we omit this objection, we do not seem to have the required effective control of our beliefs. It is usually not within my power to believe or not to believe that a truck is approaching me, when I am about to cross a street and a truck seems to be approaching. And the same is true of our other beliefs whether they are perceptual, introspective, inferential or memory beliefs. The point is not, as it is often claimed, that even though we do not have direct control of our beliefs, we can control them indirectly. We do not usually have even any indirect control of our beliefs. Suppose that someone offers me one million pounds if I believe that I have no nose. Even if I were more interested in the money than in believing the truth, I would have no idea of what to do to get myself to believe that, except to cut my nose.[4]

It does not follow that we ought to stop thinking of epistemic justification as freedom from blameworthiness. We just have to avoid thinking that the relevant obligations are obligations to believe and to refrain from believing. Even though we cannot voluntarily decide to believe, we can influence our beliefs by voluntary actions. We can check whether we have considered all the relevant evidence, whether the observation conditions are normal, ask other people for their opinion. Or we can influence our belief forming propensities by training ourselves to be more careful in our inferences, more critical of authorities, to avoid wishful thinking and paying more attention to the evidence. So we can be responsible and blameworthy for what we believe because there are obligations that relate to actions that influence our beliefs. We can be blameworthy because we believe something we would not believe if we had done our duty.[5]

William Alston (1989, p. 95) argues that this is not yet what we expect of epistemic justification. What is missing, in his view, is an adequate truth-conducive ground, an objective connection between justification and truth. This is why I may have done everything expected of me in regulating my belief formation and still hold a belief on outrageously bad reasons. Suppose that I have grown in an isolated community in which everybody accepts the traditions of the tribe as authoritative. It has never entered my mind to put the traditions in question. There is nothing I could have been expected to do to change my belief-forming tendency. According to Alston, I cannot in these conditions be blamed for taking the traditions as authoritative. I am deontologially justified in believing what I believe even though I may have very poor reason for believing so.

Why is this an objection to the deontological conception of justification? It is clear that justification in this sense guarantees neither truth nor even probable truth. But why should there be such a connection between justification and truth? Why should justification be truth-conducive? The obvious answer is that justification is something that tends to make true belief knowledge. If justification is not truth-conducive, it does not help in distinguishing knowledge from accidentally true belief. As we saw, this answer is not adequate because after Gettier we are able to argue only that justification together with the fourth condition of knowledge is truth-conducive. Nevertheless, we will see that taking justification to be something required for knowledge does give support to truth-conduciveness accounts of justification.

However, justification may be interesting quite independently of knowledge. And if Plantinga (1993a, p. 14) is right about the origin of the concept, it has very little to do with our current interest in finding the necessary and sufficient conditions for knowledge. For Locke, in particular, knowledge and belief are quite distinct states, and the concept of justification applies only to the latter. This is because we have knowledge only of something about which we are certain. If some proposition is certain for me, there is, in Locke's view, no question of regulating my belief with respect to it and thus no question of my being justified or not justified in believing it. Plantinga (1993a, p. 4) thinks himself that justification in this deontological sense is not required for knowledge and uses the term 'warrant' of that quantity enough of which distinguishes knowledge from mere true belief.

If we, nonetheless, have doubts about the existence of epistemic duties or the philosophical significance of the deontological conception of justification, we may try to characterise a non-deontological normative concept of justification. Alston (1989, p. 97) gives the following suggestion as an alternative to the deontological conception of justification: to say that *S*'s believing that *p* is justified is to say that *S*'s believing that *p* is a good thing from an epistemic point of view. So we can evaluate our beliefs as being good, desirable or favourable from an epistemic point of view without relying on any deontological notions. To distinguish epistemic evaluation from moral, aesthetic and practical evaluation, we need to tell what is this particular epistemic point of view. According to Alston (1989, pp. 83-84), it is defined by the aim at maximising truth and minimising falsity in a large body of beliefs. The qualification 'in a large body of beliefs' is added because otherwise the aim would be achieved by restricting one's beliefs to those that are obviously true.

Alston suggests that this 'evaluative' conception of justification is pretty much the common ground that every epistemologist is ready to accept. Of course, then, we must understand it so broadly that it covers also the deontological conception. Still, Alston is not quite right in this suggestion because there are philosophers, such as Paul Moser (1989, p. 42), who takes justification to be purely descriptive concept that is in neither deontological nor non-deontological sense normative or evaluative. For Moser (1989, p. 36), epistemic justification is needed just to exclude coincidentally true belief and to provide the adequate relation between the belief and truth conditions for knowledge. He thinks thus that justification is something that contributes to making true belief knowledge, and that this something need not be anything evaluative.

Alston seems to think that it is a virtue of this evaluative conception of justification that it is neutral with respect to different substantive accounts of justification, that it does not rule out some such theories by definition. However, it is also so generic that it is not helpful in our trying to determine what kind of beliefs are justified, i.e. good from the epistemic point of view. For example, it would seem that true beliefs are good from the epistemic point of view and that false beliefs are bad. However, it seems to be wrong to say that all true beliefs are justified and all false beliefs unjustified. To distinguish justification from truth, Alston (1989, pp. 4-5) introduces an internalist constraint. What makes a belief justified must be internal to the subject, something to which she has direct cognitive access. This is, however, completely unmotivated simply on the basis of the generic evaluative conception of justification.

III The Argumentative Conception of Justification

There is a third way of thinking about justification, which may come closest to the way the term 'justified' is actually used in ordinary language. The grammatical form of the term suggests that to say that a belief is justified is to say that the person in question has justified it, that she has successfully argued for its truth.

One may think that this is, however, too restrictive use of the term. There are not many beliefs that we have actually justified. That is why it is usually thought that it is enough that the person is able to justify her belief. She need not already have done it. So what is crucial is that the person possesses an argument or reasons in favour of her belief whether or not she has actually presented this argument or these reasons.

It is sometimes suggested that this view of justification would involve a confusion between the state of being justified and the process of justifying.[6] The former is a state in which one's belief has a normative epistemic property; the latter is something one does to defend one's belief. However, no confusion need be involved. The issue is rather whether we should understand the state in terms of the process or the process in terms of the state. The argumentative conception of justification accepts the former path; the normative conception accepts the latter one.

All these three conceptions of epistemic justification are quite common in philosophical literature, in spite of being rarely explicitly expressed. A philosopher advocates usually some combinations of these. So it is typical that one who accepts either the normative conception or the argumentative conception thinks also that justification in this sense contributes to making true belief knowledge. But there are also philosophers who want to distinguish justification from anything that has this role. Anyway, it is useful and illuminating with respect to the controversies among substantive theories to keep these conceptions initially apart. It helps us to see the motivation behind different suggestions for the substantive conditions of justification and finally to evaluate what kind of concept or concepts of justification we need in epistemology.

IV The Epistemologist's Question

When we now ask the central question of substantive epistemology 'What makes *S* justified in believing that *p*?', it is important to notice that the question is ambiguous, which is to be expected remembering the different conceptions of justification. The same question is often expressed by using the term 'know' in the sentence 'How does *S* know that *p*?' and the same ambiguity remains. By both question, we may mean asking for (1) *S*'s reasons for believing that *p*, or (2) those conditions that are sufficient for *S*'s being justified in believing that *p*.[7]

The first interpretation accords very well with how these questions function in ordinary language. When *S* makes the claim that she knows that *p* or simply the claim that *p*, we may ask her 'How do you know?' or 'What makes you justified in believing that *p*?' These questions may be understood as challenges for *S* to defend her belief that *p*. We expect her to give reasons for her belief, to defend it by an argument. So we expect her to express some of her other beliefs that serve as her reasons for believing that *p*.

On the second interpretation, the question is not a challenge. It is a question concerning those properties of *S*'s belief that make it justified or those conditions under which it is justified. It is a question about the substantive conditions or the factual basis of justification – or, to put it shortly, the sources of justification. As an answer to it, epistemologists try to formulate epistemic rules or principles that specify these sufficient conditions of justification.

If one confuses these two meanings of the question, it is easy to think that it is *S*'s reasons that makes her justified in believing that *p*. And, indeed, some theories of justification are accused of being guilty of this confusion. However, if there is nothing incoherent in the argumentative conception of justification in itself, one can acknowledge the intimate connection between these two understandings of the question without any confusion. It may be *S*'s capacity to give reason for her belief that makes her justified in believing that *p*.

By making the question under the second interpretation and trying to answer it, epistemologists assume a doctrine that is currently widely accepted. According to it, justification, like other epistemic properties, supervene on natural properties. What this means is that for a belief to be justified it must have a natural property in virtue of which it is justified. To put it more accurately, for every justified belief, there must be a natural, non-epistemic, property *N* such that

(1) the belief has *N*,
(2) necessarily, whatever belief has *N* is a justified belief.[8]

Why do we think that epistemic properties supervene on natural properties? If epistemic properties did not supervene on natural ones, they would be autonomous, unanchored in any natural properties. That a belief is justified would be a brute fundamental fact unrelated to any of its natural properties. We feel strongly that this cannot be so. If a belief is justified, there must be an explanation for this. It must be justified in virtue of its nonepistemic properties.[9]

So, in epistemology, we are to understand the question 'What makes *S* justified in believing that *p*?' as a question about the sufficient conditions of justification that are to be specified in nonepistemic terms. Different theories of justification are to be understood as different answers to it.

V FOUNDATIONALISM AND THE REGRESS PROBLEM

Substantive theories of justification are traditionally divided into two types: foundationalism and coherentism. Both types of theories require that the justified beliefs of a given individual instantiate a certain kind of structure. According to foundationalism, some of the beliefs form the foundation that supports the other beliefs. The traditional metaphor for this structure is a building or a pyramid that rests on its own foundation. Coherentism denies that there is any foundation in our belief system. All justified beliefs have the same status. The justification of every belief depends on its relation to other beliefs. It depends on the coherence relations between beliefs. The traditional metaphor for this structure is a ship or a raft, the different parts of which support each other. We will see that this division is not exhaustive, but let us begin with these two traditional alternatives.

Foundationalism is the view that there are two kinds of justified beliefs. All beliefs are not justified in the same way. Some justified beliefs are foundational or basic, and all other justified beliefs owe the justification to these basic beliefs. Basic beliefs are understood as beliefs that have their justification independently of other beliefs. So foundationalism is characterised by two theses:

> A. There are basic beliefs.
> B. The justification of all other beliefs depends on their relation to basic beliefs.

The ultimate source of justification is thus the basic beliefs, which are immediately justified. All mediately justified beliefs derive their justification from immediately justified beliefs. All justified beliefs form a hierarchy that is organised in terms of epistemic dependence or priority. Beliefs in the upper layers depend in their justification on the beliefs in the lower layers, which do not depend in their justification on beliefs in the upper layers. And finally there is the foundation that supports the whole structure and that does not need any support from other beliefs. We may say that justification originates in the foundation and is then transmitted to the upper layers.

There are different varieties of foundationalism depending on how the justification of basic beliefs is explained and how justification is thought to be transmitted from basic beliefs to mediately justified beliefs. Historically, basic beliefs are understood as beliefs whose truth we can directly apprehend. Rationalists thought that such truths are comprised of self-evident propositions (Descartes' clear and distinct perceptions and Locke's perceptions of the agreement and disagreement of ideas). Empiricists added truths pertaining to what is directly given in experience. Rationalist thought that justification is transmitted by deductive reasoning. Empiricists allowed also induction.

Both traditional rationalists and empiricists can be described as radical foundationalists because they thought that basic beliefs have a very strong epistemic status. The degree of justification of basic beliefs amounts to certainty. This certainty derives from a variety of epistemic immunities. Basic beliefs are variably described as being immune from error, doubt and refutations. The terms customarily used about the foundation of basic beliefs are thus 'infallibility', 'indubitability' and 'incorrigibility'.

The standard criticism of radical foundationalism is that there are not many – if any – such certain beliefs, definitely not enough to support all those beliefs that we take to be mediately justified. It is pointed out that even the prime candidate of traditional foundationalists for basic beliefs, our beliefs about our own conscious mental states, is not in the required sense certain. This is so because my belief that I am in such and such a mental state involves a concept of such a state, and a concept can always be applied incorrectly. So there is always a possibility of error and with it the possibility of doubt and refutation. And even if it were admitted that our beliefs about conscious states are certain, we would still have the problem of explaining how these certain beliefs are able to support our justified beliefs about the external world – the infamous problem of our knowledge of the external world.

That is why contemporary foundationalists reject radical foundationalism. Basic beliefs need not be certain. It is enough that the degree of justification that they have independently of other beliefs is sufficient to satisfy the justification condition of knowledge – assuming that knowledge requires justification. We may call them moderate foundationalists. In their view, the justification of basic beliefs is defeasible. It may be lost when new justified beliefs are acquired. So even though basic beliefs do not need the support of other beliefs for their justification, this justification may be defeated by other beliefs.

There is even a weaker form of foundationalism that requires also positive support from other beliefs to basic beliefs. This minimal foundationalism attributes to basic beliefs just a very low degree of independent justification that is not enough to satisfy the justification condition of knowledge. Support from other beliefs is also needed. If we stick to our initial characterisation of foundationalism and coherentism, this theory is actually a combination of foundationalism and coherentism.[10]

Foundationalists defend typically their positions by the so-called regress argument. It is an indirect argument that purports to show that all alternatives to foundationalism are unviable. Let us suppose that there are no basic beliefs, the argument goes. Then the justification of every belief depends on its relation to other beliefs. This leads to an infinite regress, because these other beliefs must also be justified, and they must also owe their justification to still further beliefs. Take any belief that *p*. Suppose that it gets its justification from the belief that *q* and the belief that *r*. We may call these beliefs the reasons for the belief that *p*. To be able to justify the belief that *p*, these reasons must, of course, themselves be justified. And because these reasons get their justification in the same way, we get into an infinite regress. So, because the denial of basic beliefs leads to absurdity, there must be basic beliefs.

Actually an infinite regress is not the only alternative to the existence of basic beliefs. The chain of reasons may form a loop or terminate in an unjustified belief. It is clear for a foundationalist that these alternatives are equally unsatisfactory. This is so because she thinks that justification is transmitted in the chain of reasons from one belief to another belief. If there were no basic beliefs, there would be nothing to be transmitted and no actually justified beliefs. In an infinite chain of reasons, each belief is just potentially justified. Each belief is justified if the belief that serves as its immediate reason is justified. But no belief is actually justified. The same is true of circular chains. In such a chain, the belief that *p* is justified if the belief that *q* is justified, the belief that *q* is justified if the belief that *r* is justified, and the belief that *r* is justified if the belief that *p* is justified. So the belief that *p* is justified if it is justified. It is only potentially justified, not actually justified. In the chain in which the last member is an unjustified belief, no belief is justified because there is no justification to be transmitted. So it seems to be clear that if any belief is to be actually justified, there must be conditions in which justification is generated. Conditions in which justification is transmitted from belief to belief are not enough.[11]

What we have here is an elimination argument for foundationalism. All alternatives to foundationalism are eliminated on the ground that they are not able to explain how our beliefs get their justification. To be sure, the argument assumes that there are some mediately justified beliefs. If there were no mediately justified beliefs, there would not be any need for basic beliefs either. So it does nothing to show that scepticism is false. To defend herself against scepticism, a foundationalist must explain how her own position avoids elimination. She must explain how there can be any basic beliefs and nonbasic beliefs that owe their justification to basic beliefs. She must tell us what makes basic beliefs justified and how this justification is transmitted to nonbasic beliefs.

It is important to notice that the regress argument does not support radical foundationalism. All that is required to stop the regress is that there are basic beliefs that do not owe their justification to other beliefs. They need not be certain. Their degree of justification must just be sufficient to satisfy the justification condition of knowledge. Neither does it support minimal foundationalism. The weakly justified beliefs are not capable of stopping the regress because they need support from other beliefs and these other beliefs create a regress or circle. So the regress argument supports only moderate foundationalism.

To explain how basic beliefs are justified, a moderate foundationalist can choose one of three approaches. She may think that basic beliefs are (1) self-justified, (2) justified by non-doxastic, non-propositional, experience or (3) justified by a reliable non-doxastic source of the belief. I will discuss the latter two options under the titles of evidentialism and reliabilism. To explain how the justification of basic beliefs is transmitted to other justified beliefs, a moderate foundationalist may appeal to (1) deduction, (2) induction, (3) inference to the best explanation, and (4) inference permitted by epistemic principles.

The critics of foundationalism focus typically their attention on the efforts to explain how basic beliefs get their justification. They argue that these efforts fail. Beliefs are always justified by other beliefs. So there can be nothing outside of beliefs that is able to justify them. Neither can beliefs justify themselves. Merely having a belief is never enough to make it justified. I will discuss the two most influential antifoundationalist arguments in the next section.

VI THE DOXASTIC ASCENT ARGUMENT AND THE MYTH OF THE GIVEN

There are two influential arguments against moderate foundationalism. One is directed particularly against the empiricist form of foundationalism. It is argued that this form of foundationalism is committed to the myth of the given. The other one, called the doxastic ascent argument, is presented against all forms of foundationalism. Both are given in Wilfrid Sellars' famous but very difficult paper 'Empiricism and the Philosophy of Mind' (1963). They are further developed and more accessibly formulated by Laurence BonJour. I will therefore follow Bonjour's versions.

BonJour (1985, pp. 30-33) starts his criticism of foundationalism by the doxastic ascent argument. It is a *reductio ad absurdum* argument trying to show that the assumption that there are basic beliefs leads to a contradiction. (This is at least the way I will construe it.) So let's assume that *S*'s belief that *p* is a basic belief. To be basic, BonJour points out, *S*'s belief must have a feature in virtue of which it qualifies as basic and this feature must also constitute a good reason for thinking that the belief is true. In other words, there must be the following justificatory argument:

(i) *S*'s belief that p has feature *F*.
(ii) Beliefs having feature *F* are highly likely to be true.
Therefore, *S*'s belief that *p* is highly likely to be true.

But it is not enough that justification along the above lines exist in the abstract, BonJour argues. In order to be justified in believing that *p*, *S* must also be in cognitive possession of that justification. In other words, *S* must believe the premises of the justificatory argument and must be justified in believing them. What this means is that *S*'s belief that *p* is not basic after all because now *S*'s being justified in believing that *p* depends on other justified beliefs. So, because the assumption of there being basic beliefs leads to the conclusion that there are no basic beliefs, there are no basic beliefs. Every justified belief requires reasons constituted by other justified beliefs.[12]

BonJour mentions two possible foundationalist responses to this argument. One is to admit that, for a belief to qualify as basic, it must have a feature that makes it highly likely to be true but to deny that the person in question needs to know or believe justifiedly, or believe at all, that her belief has that feature and that beliefs having the feature are highly likely to be true. A foundationalist of this sort is an externalist because she thinks that what justifies a belief may be some facts that are external to the believer's conception of the situation.

BonJour's (1985, p. 8) opposition to externalism is based on the deontological conception of justification. To accept a belief without having any reason for it is to neglect the pursuit of truth. According to BonJour, such an acceptance is epistemically irresponsible. *S*'s belief may have a feature that makes it highly likely to be true, but if *S* has no conception of this fact, she is epistemically irresponsible in accepting the belief. From her own point of view, she has no reason what so ever for believing what she does, and her belief is thus not justified.

Some critics of foundationalism rely on the argumentative conception of justification. Thus Keith Lehrer (1974, pp. 187-188) writes:

> In whatever way a man might attempt to justify his beliefs, whether to himself or to another, he must always appeal to some belief. There is nothing other than one's belief to which one can appeal in the justification of belief. There is no exit from the circle of belief.

This seems to be obvious if one is talking about the process of justifying beliefs. To justify a belief is to present it as a conclusion of an argument the premises of which one believes. In the process of justification, there is nothing else than one's beliefs to which one can appeal. The situation is the same if one thinks that the state of being justified in one's belief requires that one is able to justify it. For being able to justify one's belief, one must have other beliefs to which one can appeal – or so one may argue. In the case of perceptual and introspective beliefs that foundationalists have typically taken to be candidates for basic beliefs, these other beliefs are metabeliefs concerning the reliability of the sorts of beliefs in question, as BonJour's doxastic ascent argument suggests.

The force of the doxastic ascent argument depends thus on the deontological or argumentative conception of justification. If one rejects both, the need for metabeliefs does not arise, and one can defend the existence of basic belief that are made justified by their reliability. I will discuss this sort of reliabilist views below.

The other foundationalist response to the doxastic ascent argument is more traditional. It concedes that in order for a belief to be basic there must be a justification of the sort sketched by BonJour and the person holding the belief must be in cognitive possession of the justification but this possession does not involve further beliefs. The believer must, indeed, have a reason for taking her basic belief to be true, but this reason is not constituted by beliefs. It is constituted by cognitive states of more rudimentary type, intuitions, immediate apprehensions, or direct awarenesses. The fact or state of affairs that makes the basic belief true is directly apprehended to obtain. It is directly given to the mind.[13]

BonJour (1985, p. 69) follows Sellars, however, and argues that this idea of givenness is a myth. It falls into the following dilemma: (1) if the intuitions, apprehensions or awarenesses are construed as being propositional in content, then they are capable of providing justification for basic beliefs but are in need of justification themselves; or (2) if they are construed as being non-propositional in content, then they do not themselves need justification, but neither are they capable of providing it. So, in neither case, can the appeal to the direct apprehension of the given explain how basic beliefs get their justification. This is why the doctrine of givenness is a myth.

What seems to be behind this argument is the argumentative conception of justification. Sellars (1963, p. 169) surely accepts it, and also BonJour uses sometimes words that discloses his commitment to it. (So he seems to accept both the deontological and the argumentative conceptions of justification.) Namely, if we accepted this conception, it would be clear why we would think that only mental states that have propositional content are capable of providing justification. This is because only this sort of states can serve as premises in an argument. And we would be forced to accept both horns of the dilemma.

But if we rejected the argumentative conception, then it would be possible for us to deny the second horn of the dilemma. We might even accept BonJour's official deontological conception of justification and argue that non-propositional, non-doxastic, experiences can make basic beliefs justified. Let us assume that it seems to me that there is a red book in front of me (a non-doxastic experience) and that I form a belief that there is a red book before me and that I have no reasons to doubt the truth of my belief. Why would I be epistemically irresponsible in forming this belief? The thought that epistemic responsibility always requires doxastic reasons for one's belief seems to be motivated by – implicit – acceptance of the argumentative conception of justification and not by the deontological conception as such.

So there are two ways out of the antifoundationalist's trap for a foundationalist. She may reject the deontological and argumentative conceptions of justification and be an externalist. I will discuss this sort reliabilism soon. What makes a person's belief justified, according to it, is its having a reliable source. It is not further required that she believes justifiedly or knows or believes at all that the source is reliable. The other way out is to reject just the argumentative conception and argue that non-doxastic experiences can make beliefs justified. This is an internalist view that does not require that the justifiers are within the believer's perspective or conception of the situation. It requires only that the believer has an epistemic access to the justifiers of her belief or is aware of them in some nondoxastic way. We may call them access internalism and awareness internalism respectively – in contrast to BonJour's perspectival internalism. I will discuss theories of this sort under the name of evidentialism.

Finally, it may be noted that BonJour's two arguments are overkill. It is not easy to see how even an antifoundationalist is able to escape from the trap. For a person's belief to count as justified, she must have reasons for taking the belief to be probably true and these reasons must be further justified beliefs of hers, according to BonJour. But then she must also have reasons for taking these beliefs to be probably true, and so on *ad infinitum*. Is there any way even for an antifoundationalist to avoid this regress?

VII COHERENTISM

According to coherentism, there are no basic beliefs. All beliefs get their justification in virtue of their relation to other beliefs. Beliefs are made justified by their 'cohering' with each other. How is this view able to avoid the regress problem? Even though it is sometimes suggested that there is nothing wrong with circular justification so far as the circle is large enough, current coherentists typically reject the linear view of justification that creates the regress problem. Justification is not transmitted along a linear chain of reasons. Justification is holistic. It depends on one's all beliefs taken together. All beliefs are made justified by their belonging to a coherent system of beliefs. The coherence of this system is a matter of complex reciprocal relations that obtain between the beliefs of the system. This is true both of ordinary first-order beliefs about non-doxastic matters and of second-order beliefs about the reliable sources of beliefs. A coherentist may thus think that she can avoid even the regress created by BonJour's doxastic ascent argument and take this argument to support her position.

There is no unanimity among theorists about how coherence is to be understood. The most simple-minded view identifies coherence with consistency. It is clear, however, that consistency is not sufficient for coherence and justification. That would make any consistent fairy tale justified for us. Even though coherentists take more typically consistency to be necessary for justification, it is arguable that even this is not true. It is not plausible to think that inconsistency in a small part of a belief system would make all one's beliefs unjustified. Furthermore, our systems of beliefs are huge networks of interrelated beliefs, only a tiny part of which is conscious at a time. It is probable that all such systems include some undetected – and even detected[14] – inconsistencies. The requirement of inconsistency leads therefore to scepticism.

More usually, coherence is thought to require relations of mutual positive support between beliefs. Some early coherentists thought that this mutual support is a matter of logical implication. So a coherent system of beliefs would be one in which every belief entails and is entailed by the rest (Blanshard, 1939). This is an extremely strong requirement. Many current advocates understand the required mutual support in terms of weaker explanatory relations (Sellars, 1963; Harman, 1973). According to this view, two beliefs support each other if one explains (the truth of) the other. According to a more subjective view, a belief coheres with others if the subject takes it to be more likely to be true than its competitors on the basis of these other beliefs (Lehrer, 1974, 1990).

John Pollock (1987, p. 72) has pointed out that, in addition to coherence theories that require positive support between beliefs, there are negative coherence theories according to which coherence relations have a purely negative role. These theories take all beliefs to be automatically *prima facie* justified. This *prima facie* justification can then be undermined or defeated by incoherence. Note that also a moderate foundationalist preserves this negative role for coherence. The justification of both basic and nonbasic beliefs can be defeated by their incoherence with other beliefs. However, it is difficult to find any actual advocates of pure negative coherentism.

Current coherentists think typically that there must be some metabeliefs or second-order beliefs in a coherent belief system. These are required for explaining how introspective, perceptual and memory beliefs get their justification. These are beliefs that arise spontaneously, without being inferred from other beliefs. Still, they may be made justified by their cohering with metabeliefs that attribute high reliability to beliefs of those kinds. And these metabeliefs are in turn made justified by their cohering with first-order beliefs and other metabeliefs.

In whatever way coherence is to be construed, there are classical objections to coherentism that must be faced. What makes it the target of these objections is the fact that it makes justification a function of the subject's beliefs and their relations to each other. Nothing outside of her beliefs affects justification. They are the objections from alternative systems of belief and the objection from isolation or detachment from reality.

It seems to be undeniable that there can be alternative incompatible systems of belief that are equally coherent. Coherence theory does not have any means to distinguish between them. It would make beliefs in all those systems equally justified, which seems to be wrong. This would be a serious objection against the coherence theory of truth because it would make both a proposition and its negation true and thus violate the law of contradiction. But it is not so clear that the coherence theory of justification is vulnerable to this objection. One person may very well be justified in believing a proposition that coheres with her system of beliefs while another person is justified in believing its negation that coheres with her system of beliefs. Our predecessors over 600 years ago may very well have been justified in believing that the earth is flat while we are justified in believing that the earth is not flat. There is no contradiction because justification does not guarantee truth.

The objection from isolation or detachment from reality draws out attention to the fact that coherence theory does not require any input from the external world for justification. Our beliefs would be justified even if we were totally isolated from the world. This second objection is a mirror image of the first one – as Sosa (1991, p. 184) points out. In the first one, we imagine that our system of beliefs is allowed to vary while the world is fixed. In this one, the world is allowed to vary while the system is fixed. Even though these objections against coherentism may not be conclusive, they show at least that the coherence theory implies that there is no necessary connection between justification and truth or even probable truth. There is at most a doxastic connection: the subject must believe that her beliefs are connected to truth and reality, but they need not actually be so connected. So a coherentist must reject the view that justification is by its nature truth-conducive.

A coherentist may try to avoid this result by two manoeuvres. She may try to define truth in terms of coherence.[15] Absolute idealists at the turn of the century and more recently antirealist philosophers, such as Hilary Putnam (1981) and Nicholas Rescher (1985), have suggested that truth is to be understood as some kind ideal coherence. If truth is ideal coherence, then increasing coherence of our view brings it closer to truth. Or she may try to argue, like Donald Davidson (1986), that belief is by its very nature veridical. By relying on his view of radical interpretation and the principle of charity, Davidson argues that it is constitutive for having beliefs at all that most of one's beliefs are true. If this is so, we can understand why increasing coherence of one's view increases also the amount of truths in it. Both of these suggestions seem to compromise our view that truth is objective, independent of what we think about it, which is why neither of these strategies have been popular among realistically oriented epistemologists. If truth about the world and our own mind is independent of our epistemic and interpretative stances, neither of these manoeuvres will work.

However, it is also possible for a coherentist to stick to the objectivity of truth and argue from the deontological conception of justification that the missing truth connection is not a problem. According to this conception, there need not, indeed, be any necessary connection to truth. It is enough for justification that a subject is not violating any epistemic duties in believing what she does believe, that she does her best in pursuing truth. How successful she is an external matter that does not affect her being justified in her beliefs. If she does her best, it is not her fault that her beliefs are isolated from the world. A coherentist may argue further that theories that require truth-conducivity of justification are false. What is external to the system of beliefs does not affect justification.[16]

There is, however, another version of the isolation objection to which even deontological coherentism is vulnerable. It is argued that coherentism makes justification isolated from empirical evidence when evidence is understood to cover not just supporting beliefs but also sensory experiences.[17] According to coherentism, justification is a matter of relations between beliefs. Anything outside of our beliefs has no affect on justification, not even our experiences. We can imagine situations in which our experiences are allowed to vary while our beliefs are fixed. For example, I believe now that I am sitting quite comfortably staring at the computer screen and that I have no pain anywhere. I believe also that I have just visited a doctor who told me that I have no medical problems what so ever. I believe that I am a healthy man. Assume that suddenly I get a splitting headache for staring at the screen too long. According to coherentism, I would still be justified in believing that I have no pain because my nondoxastic sensations do not affect the justification of my beliefs. But this is clearly wrong. Surely, I am not justified in believing that I have no pain while I am really having a splitting headache.

A coherentist will claim that a headache has no affect on justification unless I believe justifiedly that I have a headache. But does the deontological conception of justification motivate this claim? It is more plausible to think that the real motivating force behind coherentism is the argumentative conception of justification. For, given this conception, it becomes understandable why nondoxastic experiences are not able to produce justification. Only something that has a propositional content is able to serve as a premise in an argument.

Finally, does holistic coherentism really avoid the regress created by BonJour's doxastic ascent argument against foundationalism? If it does not, then it can be directed against coherentism as well. According to coherentism, a subject's belief that *p* is made justified by its coherence with her system of belief *S*. But this cannot be so, according to Bonjour's argument, unless the subject in question has also the following justified beliefs:

(1) Her belief that *p* coheres with *S*.
(2) Beliefs that cohere with *S* are likely to be true.

And to be justified in these beliefs, she must have the justified beliefs that (1) and (2) coheres with *S*. And these latter beliefs to be justified, she must believe that they cohere with *S*, and so on *ad infinitum*.[18]

Ernest Sosa (1991, p. 183) has suggested that this regress is vicious because it violates the principle of supervenience, which says that there are nonepistemic conditions that are sufficient for justification. The regress arises if it is required that the subject must also justifiedly believe that the conditions obtain and are truth-conducive. And the principle of supervenience would be violated. For then there would not be any nonepistemic sufficient conditions for justification. Of course, a coherentist could accept the principle and argue that coherence with a belief system without the metajustification is sufficient for justification, but then she loses her best argument against foundationalism.[19] A foundationalist argues also that sense experiences or reliable nondoxastic sources can make beliefs justified without the need for any justified metabeliefs.

VIII CONTEXTUALISM

There are a few philosophers[20] who think that there is a solution to the regress problem that avoids both the problems of foundationalism and coherentism. They think that the chain of reasons is linear but does not terminate in a basic belief. The chain terminates in a belief that is not justified. This view is inspired by Ludwig Wittgenstein who says in *On Certainty* (1969, § 253) that "at the foundation of well-founded belief lies belief that is not founded".

At first sight, this solution to the regress problem seems to be hopeless. How can a belief that is not itself justified justify other beliefs? However, the idea seems to be this. When a subject claims to know something, she is required to respond to challenges that are raised against her belief. In responding to these, she appeals to other beliefs of hers. But the status of these beliefs may also be challenged. And she must go on defending those beliefs by still other beliefs. Finally, she ends up with beliefs that are not challenged and that are therefore not in need of justification. These are the unjustified beliefs that lie at the foundation of all justified beliefs.

At the background of this view, there is obviously a version of the argumentative conception of justification. To be justified in her belief, a person must be able to defend it by arguments. She must be able to defend her belief in the face of objections that have been raised or could be raised against it. So "justified belief" is here understood as belief one can successfully defend against objections.

In contrast to coherentism, contextualists do not require that a subject be able to meet all possible objections to her belief. Only those objections need be met that reflect some real doubt. So she need not be able to defend those beliefs that are taken for granted by her community or social group. In this way, justification depends on the social context of the subject. It depends on social consensus.

We cannot deny that there can be this kind of contextual justification. The question is what is its epistemological significance. Our central epistemic ends are to believe what is true and to avoid believing what is false. But this kind of contextual justification is not directed at these ends. It is directed rather at the end of enlarging the consensus. So it seems that, as epistemologists, we need not pay attention to this kind of justification.

Indeed, Richard Rorty (1980) who is the most outspoken defender of contextualism preaches also the death of epistemology. He suggests that we should not be interested in objective truth or how things really are. We should instead be interested in solidarity, in having as much intersubjective agreement as possible.[21] This is, of course, the aim to which contextual justification is the most obvious means.

It is not, however, clear how this shift of interests could avoid all considerations of objective truth and how things really are. Is not the existence of a consensus itself a matter of how things really are? We can surely be mistaken about its existence. There is always the question whether our belief about consensus is objectively true or not. And so, we are back in the epistemological problems. Rorty could, of course, suggest that also this question of the truth of the belief about consensus should be replaced by the question of there existing a consensus about this consensus. But in this way, he is heading for a vicious regress.[22]

IX RELIABILISM

One response to the doxastic ascent argument is to admit that a basic belief must have a feature that makes it highly likely to be true and that there, indeed, must be a sound argument that shows it to be most likely true but to deny that the believer must in any way be aware of this feature or have the argument in her cognitive possession. It is enough that a basic belief is of a type most of the instances of which are true. The believer need not know or believe justifiedly or even believe at all that her belief is of this type. In a word, a justified belief is a belief that is of a reliable type.

Among current reliabilist theories of justification, there are three main varieties: In their simplest forms, (1) the process theory says that a belief is justified if and only if it is produced (and sustained) by cognitive processes that are generally reliable (Goldman 1979, 1986), (2) the virtue theory says that a belief is justified if and only if it results from the use of intellectual virtues, where virtues are reliable faculties or dispositions (Sosa 1991), and (3) the indicator theory says that a belief is justified if and only if it is based on reasons or grounds that are indicative of the truth of the belief (Swain, 1981; Alston 1989, 227-245).

What is common to all these forms of reliabilism is the view that there is a necessary connection between justification and truth. To believe something justifiedly is to believe it in a truth-conducive way. Another common feature is that they all make justification a function of the causal history or source of a belief where causal history is understood broadly to include both what causally originates and what causally sustains the belief. This feature contrasts reliabilism with traditional foundationalism and coherentism that are said to be current time-slice theories. According to them, justification is a function of what is true of the believer at the time of belief.[23] However, reliabilism can share some foundationalist or coherentist features. It can make all justification dependent on a belief's causal relation to other justified beliefs, or it can take some beliefs to be basic beliefs that are made justified by their non-doxastic source. It is to be admitted that the latter foundationalist alternative is much more plausible.

We may want to make a distinction between propositional and doxastic justification.[24] We may say that a proposition is justified for a person whether or not she actually believes the proposition. This is a case of propositional justification. Or we may say that a person's actual belief is justified. This is doxastic justification. Reliabilism is a theory of doxastic justification, and it is controversial whether it has resources to offer also a theory of propositional justification.[25] Because knowledge requires actual belief, it is doxastic justification that is relevant to knowledge. If we are thus interested in justification as a necessary condition of knowledge, the focus on doxastic justification comes quite naturally.

There are four main problems for reliabilism: two technical problems and two counterexamples. Let us illustrate the technical problems in terms of process reliabilism. The first of them, the generality problem, concerns how the relevant process types are to be identified. Every belief is produced by a particular process token that may instantiate different process types. When we talk about the reliability of a psychological process, we are actually talking about the reliability of the process type the instance of which the particular token process producing the belief is. We want to know what is the proportion of true beliefs among the beliefs produced by processes of that type. The result varies with different ways of identifying the type. If the type were chosen too broadly (e.g. a perceptual process) some unjustified (perceptual) beliefs would be deemed justified. If the type is chosen too narrowly, there may actually be just one process token of the type in which case the truth ratio of the process type would be either 1 or 0, depending on whether the process token produces a true belief or a false belief. The other technical problem, the range problem, concerns the range of the process tokens that are taken into account in measuring reliability. Should only process tokens in the actual world be taken into account, or should we also consider tokens that exist in other possible world? In the latter case, we may be able to solve the problem in which there is only a single token in the actual world.[26]

One of the counterexamples is given by BonJour (1980, pp. 62-65; 1985, pp. 41-45), and its purpose is to deny that reliability is sufficient for justification. He asks us to imagine a person, say Norman, who has a reliable clairvoyance power. One day Norman comes to believe that the President is in New York though he has no evidence for or against his belief. Neither does he have any evidence for or against his possessing the power of clairvoyance. Still, he has this power, and his belief is produced by this power. According to reliabilism, Norman is justified in his belief, which is against our intuitions.

The other counterexample is due to Keith Lehrer and Stewart Cohen (1987, pp. 325-326) who try to show that reliability is not necessary for justification: Imagine that unknown to us, our cognitive processes, those involved in perception, memory and inference, are made unreliable by a powerful demon. Under these conditions, our beliefs produced by those processes are not justified according to reliabilism. This is wrong because it follows from the demon hypothesis that our experiences and our reasoning are just as they would be if our cognitive processes were reliable. The fact that our experiences and reasoning are kept fixed while the external world changes so that our cognitive processes become unreliable has no effect on our justification for believing what we do believe. So reliability cannot be a necessary condition for justification.

What both of these counterexamples suggest is that justification depends essentially on our having reasons or evidence for our beliefs. Norman lacks any evidence for his belief that the President is in New York. The victim of the demon has all the evidence that she would also have in the world in which her processes are reliable. So let us look at a theory that makes justification dependent on evidence.

X EVIDENTIALISM

The view that justification depends essentially on one's having evidence for a belief has been recently defended by Richard Feldman and Earl Conee (1987) though the view has long been implicit in Roderick Chisholm's writings. According to it, a proposition is justified for a person if and only if the person's evidence supports that proposition.[27] This is a theory of propositional justification. So it does not presuppose that the person in question already believes the proposition. It is, however, easily enlarged to offer also a theory of doxastic justification. We just add the requirements that the person believes the proposition and that her believing it is (causally) based on her evidence.

What can be person's evidence is first of all her other beliefs. But evidentialists think also that her evidence includes non-doxastic mental states, such as perceptual experiences. This makes evidentialism a version of foundationalism. The inclusion of non-doxastic experiences as a part of evidence gives evidentialism resources to offer neat solutions to the problems that bother other positions: It explains how to stop the regress. Basic beliefs are made justified by non-doxastic experiences that are themselves neither justified nor unjustified. It resolves also the dilemma created by the doctrine of the given that – it turns out – is not a myth. It has a sort of answer to the isolation objection to coherentism. Justification is not isolated from our experiences about the world. And it seems to give the right solutions to the counterexamples to reliabilism. The clairvoyant person is not justified in his belief because he has no evidence for it. (Let us assume that there are no clairvoyant experiences.) And finally, the victim of an evil demon has all the evidence we have for our beliefs and is thus as justified in her beliefs as we are in ours.

However, to be a developed theory of justification, it needs to be clarified at least in two points: (1) What is it to possess something as evidence? (2) Under what conditions does the evidence support a proposition? A highly restrictive answer to the first question would limit the evidence to what one is currently aware of or thinking of. A highly liberal view includes everything that is stored in one's mind as the evidence. An intermediate position would restrict evidence to something that is easily accessible or retrievable from memory.[28] Different views about the matter represent different internalist conceptions of justification that will be discussed in the next section.

For the evidence to support a proposition, it may be suggested that it must entail the proposition or at least make the truth of it objectively probable. This would relate evidentialism to indication reliabilism that requires that the evidence is a reliable indication of the truth of the belief, but it would also make evidentialism vulnerable to some of the objections to it. Though Conee and Feldman does not say whether supporting evidence must be truth-conducive in this way, Feldman (1992) denies elsewhere its sufficiency. It is not enough that evidence entails or makes probable the proposition. The person in question must also 'grasp' the connection between the evidence and the proposition. As Feldman notes, this requirement raises two problems: (1) It over-intellectualises the situation, because people do not grasp such evidential relations routinely. And (2) if grasping amounts to having the justified belief that the evidence supports the proposition, there is a danger of a regress.

It is clear that Chisholm (1989) does not accept a truth-conduciveness requirement for evidential support. His approach is to form a complex system of epistemic principles that specify the conditions under which evidence supports a proposition, and these principles are such that he is able to formulate them just by sitting in his armchair and considering his own state of mind. These principles do not support any reliability constraint because one must be able apply them correctly just by reflecting one's own state of mind. Chisholm's theory is an internalist theory that contrasts with externalist theories that typically require truth-conduciveness of justification.

XI INTERNALISM AND EXTERNALISM

Theories of justification divide into internalist and externalist theories. What is common to internalist theories is that they make justification a function of factors that are internal to the believing subject. According to externalist theories, the conditions of justification are external to the subject. There are also mixed theories that take the conditions of justification to be composed of both internal and external factors.

There is, however, no unanimity in how the line between internal and external factors is to be drawn. Mental internalism, that may be the most modest form of internalism, makes justification simply the function of the subject's mental states and their relations to each other.[29] Traditional foundationalism and coherentism would be internalist theories according to this view, and so would be the form of evidentialism that does not require truth-conduciveness of the supporting evidence. Reliabilism that takes justification to be truth-conducive would be an externalist theory. Mental internalism is supported by the thought experiment in which we are invited to imagine a demon-world in which our mental states stays the same as they are now but the demon arranges things in the external world so that our sources of belief are unreliable. The intuition is that our beliefs are as justified in this world as they are in the actual world in which those sources of belief are reliable. So the external factors, such as reliability, do not seem to affect our being justified in our beliefs.

Mental internalism seems to be – apart from these intuitions – unmotivated. Current internalist defend therefore one or another of the following more specific forms of internalism. (1) Perspectival internalism makes the requirement that whatever contributes to the justification of a belief must be within the subject's epistemic perspective on the world, in the sense of being something the subject knows or justifiedly believes. (2) Access internalism requires only that the justifying conditions are directly accessible to the subject, in the sense that she can determine just by reflection whether they are satisfied. (3) Awareness internalism requires that the subject is actually aware of those conditions.[30]

BonJour argued against foundationalism that it is not sufficient for the justification of a basic belief that it has a feature that makes it most likely true. The subject must also believe justifiedly that it has the feature and that beliefs having the feature are most likely true. He is thus defending perspectival internalism. This formulation suggests that it is necessary for justification that the belief actually has the feature that makes it most likely true in which case Bonjour's perspectival internalism would be a version of reliabilism and not a pure form of internalism. This requirement is too strong, however. The way internalism is defended makes it clear that justified beliefs are taken to be those that the subject has good reasons for taking to be most likely true. They need not actually be most likely true. So what actually does the justifying are the perspectival beliefs themselves, not the features believed to be present and to be truth-conducive. It is enough that these beliefs are justified; they need not be true.[31]

Perspectival internalism rules out there being any basic beliefs. Only other justified beliefs can make a belief justified. This requirement creates the threat of an infinite regress that is thought to be avoided by accepting holistic coherentism. Perspectival beliefs as well as other beliefs are made justified by their belonging to a coherent system of beliefs. A problem remains, and BonJour is very well aware it. Now, it is the coherence of the whole system of beliefs that contributes to the justification of the beliefs that belong to it. This coherence and its truth-conduciveness must itself be within the epistemic perspective of the believer. The requirement of justification for the new perspectival beliefs creates a final regress that is vicious, as we saw. BonJour's attempt to stop this regress is widely thought to be unsuccessful.

Perspectival internalism sets thus too severe demands for justification. Access internalism is more liberal. It requires only that the justifying conditions are directly accessible to the believer. She need not have actual knowledge or justified beliefs about those conditions. It is enough that she is able to come to know or believe justifiedly that they obtain if she just reflected about the matter. So there is no threat of an infinite regress.

Even access internalism may be too demanding for some theories of justification that are usually taken to be internalist. It is plausibly argued by Hilary Kornblith (1989) that people do not have a cognitive access to the coherence of their whole system of beliefs. Neither do we seem to keep track of the chain of reasons required for justification by some versions of foundationalism, as Gilbert Harman (1986, p. 41) has argued. So both coherentism and foundationalism – at least in some of their forms – may turn out be externalist theories if we make the distinction between internalism and externalism in terms of accessibility.

Access internalism is also taken to be too weak by some epistemologists. What is needed is actual awareness of the conditions that makes a belief justified. Putting it in this way makes the view too strong, however. As it is pointed out by Alston (1989, p. 233-234), nobody would be able to complete the formulation of the sufficient conditions of justification. For suppose that we begin by taking *C* as the sufficient condition for the justification of *S*'s belief that *p*. Now we would need to add according to awareness internalism that *S* must also be aware of *C*. Call this enlarged conditions *C*'. But neither is *C*' sufficient for justification because *S* must also be aware of *C*', and so on *ad infinitum*. As a reply to this criticism, an awareness internalist may make a distinction between the evidence required for justification and the sufficient conditions of justification.[32] *S* need be aware only of the former, the evidence for her belief.

So some internalists would be ready to welcome some externalist elements into their accounts of justification. Most typically, one thinks that the internalist restriction concerns the evidence required for justification. It must be composed of something accessible to the subject or something she is actually aware of. But the adequacy of the evidence or the way belief is supported by the evidence need not be in the same way internal. So these would be mixed views that combine internalist and externalist elements.

One may also save perspectival internalism from the vicious regress by accepting an externalist element. One may not apply the internalist requirement to the perspectival beliefs themselves. Then it is just required that the subject has reasons for taking her belief to be most likely true, but these reasons need not themselves be justified for her or known by her. This manoeuvre makes epistemic justification subjective, and it accepts a conservative element into the account of justification.

So, there are several different internalist constraints on the sufficient conditions of justification. Some epistemologists accept some such constraint on just one part of the conditions and accept therefore a mixed position that combines internalist and externalist elements. And there are, of course, also pure externalist theories, like process reliabilism, that do not accept any internalist constraints. But what motivates the acceptance of internalism? Let's look at the conceptions of justification with which we started this survey to see whether anyone of them gives support to any such constraint.

XII DOES THE DEONTOLOGICAL CONCEPTION OF JUSTIFICATION SUPPORT INTERNALISM?

Internalists appeal most often to the deontological conception of justification to defend their internalist constraints. Indeed, all three forms of internalism discussed above have been defended by appealing to this conception of justification. The basic idea is that since justification is a matter of responsibility and freedom from blame, my belief's being justified depends on how things appear from my perspective or on how things are so far as I can tell. If my belief is well supported from my own point of view or so far as I can tell, I cannot be blamed for my belief.[33]

BonJour argues thus that it is epistemically irresponsible to accept a belief without having good reasons for taking the belief to be most likely true. Paul Moser's (1985) lesson from the regress problems created by this view is that these reasons need not be further beliefs by the agent but awarenesses of a nonconceptual sort and that instead of perspectival internalism we should accept awareness internalism. Carl Ginet (1975, p. 36) argues that it follows from the 'ought implies can' principle together with the deontological conceptions of justification that the conditions of justification must be directly accessible to the subject. If it is our obligation not to believe that *p* when we lack justification for *p*, then it is within our power not to believe that *p*. But it is within our power not to believe that *p* when we lack justification for *p* only if justification is directly recognisable or accessible to us.

One problem of all these three defences of internalism is that they seem to attribute epistemic obligations directly to beliefs and other doxastic attitudes. They assume therefore that belief is under our voluntary control. As we saw, there are strong reasons to believe that believing is never or very rarely under such voluntary control. The other version of deontologism that attributes obligations to actions that influence belief does not support internalism at all because it makes justification a function of what the agent did or did not do before the time of the belief. And these sort of historical facts are not internal to her. So internalism supported by deontological conception of justification does not seem to be a tenable position.

Furthermore, the deontological conception that is used to defend internalism seems to support pure internalism and rule out more plausible mixed views. The same argument seems to apply to every part of the condition that makes a belief justified. This creates a danger of a vicious regress, at least for perspectival internalism and awareness internalism. And it may make access internalism a too demanding view for epistemologically unsophisticated cognisers.

All this does not mean that the deontological conception of justification is itself problematic. It just does not support internalism. One may develop externalist theories of justification on the basis of it. For example, Hilary Kornblith (1983, pp. 33-34) suggests that a justified belief is a belief that is produced by epistemically responsible action, i.e. action that an epistemically responsible agent might have taken. According to Kornblith, an epistemically responsible agent desires to have true beliefs and to have her beliefs produced by reliable processes. This view differs from process reliabilism because it does not require that justified beliefs are in fact produced by reliable processes.

XIII DOES THE ARGUMENTATIVE CONCEPTION OF JUSTIFICATION SUPPORT INTERNALISM?

Coherentism is sometimes defended by relying, more or less explicitly, on the argumentative conception of justification. Thus, Keith Lehrer (1974, pp. 188-189) argues against foundationalism that to justify a belief one must always appeal to other beliefs. There is no exit from the circle of one's beliefs. This strategy has no chance of success unless one accepts the argumentative conception of justification. Only then does the state of the belief's being justified depend on one's having other beliefs on which one has appealed in justifying it or on which one can appeal if such justifying is called for.

The argumentative conception would create a vicious regress if it required that the belief has been actually justified by appealing to other beliefs that themselves have been actually justified by appealing to further beliefs and so on *ad infinitum*. The regress is avoided, however, if it is just required – more plausibly – that one need only be able to justify it. One need not already have done it, which is, indeed, impossible if the process involves an infinite number of steps. One may be able to justify any particular belief appealed to in the process of justification without being able to go though the whole infinite chain.[34]

Furthermore, there is a pragmatic element in justification. One need only go on justifying one's belief and the beliefs appealed to in the process as long as one is challenged to do so. There will be a point where one's audience is satisfied and beyond which one need not continue the process of justification.[35]

We may now ask what makes one able to justify one's belief. The answer that suggests itself is the coherence of one's over all system of beliefs. To be able to justify one's perceptual, introspective and memory beliefs, one's system must include metabeliefs concerning the reliability of these apparently basic beliefs. The coherence is therefore perspectival including a perspective into the reliability of the sources of beliefs.[36]

A coherence theory defended along these lines has an important advantage compared to the deontologically defended coherence theory advocated by BonJour. Even though it is a species of access internalism, the over all coherence of one's system of beliefs need not be accessible to one. It is enough that the reasons appealed to in the process of justification are so accessible. It is thus psychologically much more realistic than the deontologically defended version.

Coherentism assumes, however, that one already possesses all the beliefs needed in the process of justifying one's belief. By relying on the distinction between dispositional belief and disposition to believe, Robert Audi (1994) argues forcefully that this is not correct. One must often form new beliefs that one appeals to in the process. One did not already have them. One had only a disposition to form those beliefs. For example, if I am asked to justify my belief that there is a cow in the field, I may appeal to my belief that I see it. However, I did not already believe that I see the cow before the question of justification was raised. I formed it during the process of justification. What I did have was a sensory experience of the cow and a disposition to believe that I see it, and this disposition was grounded on the experience. Now, we may say that it is my having the experience that enables me to justify my belief. Thus, in the process of justification, I do not appeal to other belief with which the belief to be justified coheres but to sensory experience directly. So, the argumentative conception of justification may actually support foundationalism or evidentialism, rather than coherentism.[37]

If we accept the often-neglected distinction between dispositional belief and disposition to believe, the argumentative conception of justification is best seen to support evidentialism according to which non-doxastic experience contributes to the justification of belief.

XIV What Kind of Justification Does Knowledge Require?

We seem to share the intuition that small children and even animals can have knowledge. It is clear that they are unable to fulfil epistemic obligations and defend their beliefs by arguments. One may draw the conclusion that knowledge does not require justification. Or, alternatively, one may think that the relevant kind of justification needed for knowledge is neither deontological nor argumentative.

If we choose the latter alternative, we may defend a purely externalist and reliabilist theory of justification. What seems to convert children's and animals' true beliefs into knowledge is their beliefs' having a reliable source. And if we think that justification is something that contributes to making true beliefs knowledge, it is natural to think also that justification is a matter of the reliable source of belief. We may also think that even the person with the reliable clairvoyant power knows that the president is in New York, and so even he has justification for his belief. The contrary intuitions are based either on the deontological or the argumentative conception of justification.

One may, however, argue that knowledge does require that the knower is able to justify what she knows. Suppose that you claim to know something, but when we ask you how you know what you claim to know, you are unable to justify your claim or belief in any way. Surely, we would conclude that you do not know what you claim to know. The fact that we are inclined to withdraw the knowledge attribution when the putative knower is unable to defend her belief suggests that knowledge does require argumentative justification. Maybe this explains why we tend to hesitate in attributing knowledge to the person with the clairvoyant power.

We may try to accommodate these two cases on the basis of indication reliabilism and argue that in both cases there are grounds (e.g. sensory experiences) that indicate the truth of the belief. Children and animals are unable to appeal to their experience but more sophisticated subject are able to do that and that is why we expect them to do so when challenged. This solution would deprive the clairvoyant person of any knowledge because he has no truth-indicative experiences.

Alternatively, we may follow Ernest Sosa (1991, pp. 240, 253-255, 282) and argue that there are two kinds of knowledge: animal knowledge and reflective knowledge. Animal knowledge requires just that the belief is produced and sustained by reliable faculties or processes. Reflective knowledge requires in addition an ability to defend one's belief. According to this view, animals and small children as well as the clairvoyant person have just animal knowledge. More sophisticated cognisers may have also reflective knowledge.

XV THE SOCIAL POINT OF KNOWLEDGE ATTRIBUTION

We can confirm the above remarks about knowledge by considering the point of making knowledge attributions. This helps also to understand the significance of those concepts of justification that are relevant to knowledge.

Edward Craig (1990, p. 11-12, 18-19) makes the plausible suggestion that the purpose of knowledge attribution is 'to flag approved sources of information'. He uses this idea about the point of the concept of knowledge to clarify the conditions of knowledge. He asks what someone seeking information about some topic wants of her source of information or informant. First of all, she wants an informant who has a right answer to the question that interests her, i.e. an informant who has a true belief about the matter. Second, and this is important for our purposes, she wants that her informant has a detectable property by which she is able to pick him out and to distinguish him from others to whom she would be less well advised to listen. Furthermore, this property should be such that correlates well with his being right about the matter. This property would be the property that makes a true belief knowledge or at least the one that does so together with some fourth property.

If we think that justification is the property that tends to make true belief knowledge and by virtue of which we pick our informants out, we have here a strong motivation for accepting some form of the reliabilist account of justification. It seems to be clear that we pick our informants out in virtue of their reliability. This is also accepted by Ernest Sosa (1991, p. 257), according to whom we care about justification because it indicates a state of the subject that is important to her community. It is the state of being a dependable source of information about a certain topic and in certain circumstances. To be such a source of information, the state involves the subject's having a belief that is based on a reliable source. So, Sosa concludes that justification (aptness) is a matter of the intellectual virtue of the subject where an intellectual virtue is understood as a reliable faculty.

This approach supports also Sosa's view that there are two kinds of knowledge. Sometimes, we use children and even animals as our informants even though they are unable to justify their beliefs and to evaluate their own reliability. Especially when it is a matter of information acquired through sensory perception, we may very well trust in their reliability. To have a more sophisticated informant who is able to justify her belief by giving reasons may also be useful. Their capacity to give reasons may help us to evaluate their reliability even in topics that are more abstract and theoretical.

On the other hand, deontological justification does not seem to be relevant to knowledge. A person who fulfils her epistemic obligations need not make a good informant. Aristotle was surely a responsible inquirer. Nevertheless, we do not rely on him any more as a source of information about the nature. We do not take him to be reliable in such matters.

The idea that we need the concept of justification because we are interested in sharing information with each other gives thus strong support for some form of reliabilism. But it also underwrites the significance of the argumentative conception of justification. This conception supports in turn either coherentism or evidentialism. Accepting indication reliabilism that combines reliabilism and evidentialism would give us the more economical view that requires just one concept of justification. However, there are some intuitions that this kind of justification is not necessary for knowledge.[38] If we want to save these intuitions, we need to accept the less economical view that there are two kinds of knowledge and two kinds of justification respectively.

XVI WHAT AM I TO BELIEVE?

We have so far focused on our need to evaluate other people as potential informants, but is there not a more fundamental need to evaluate ourselves and especially what we ourselves are to believe in the first place? Even when we are choosing an informant, we must first make up our own mind about whom we are to trust and to take to be reliable.

Our whole epistemological tradition has focused on the situation where a doubt has been raised, and we need to make up our mind about what to believe or whether to withhold judgement altogether. Ancient sceptics emphasised the existence of disagreement and the need for finding a criterion of truth by which the disagreement could be resolved. Unfortunately, they found out that there is also a disagreement about the criterion and ended up with suspending judgement about all matters and living without beliefs altogether. Modern thinkers have not taken this to be a feasible alternative. So the need for finding a criterion of truth is more pressing.

To talk about the criterion of truth suggests that there is some easily accessible property the existence of which guarantees truth. Even though early modern thinkers, such as Descartes, may have thought that there is such a property, contemporary epistemologists are more pessimistic. They have been content with seeking a property that makes a proposition justified for one. It need not guarantee the truth of the proposition.

Which concept of justification is relevant to this situation? The attribution of justification is to help us to decide what to believe. We are still deliberating and try to determine which attitude is justified towards a proposition: (1) Are we to believe it? (2) Are we to believe its negation? (3) Or are we to suspend judgement? First of all, we are here interested in the concept of propositional justification the attribution of which does not require an existing belief. When we were interested in knowledge and sharing of information, there was always a belief that we were evaluating. It was doxastic justification that was relevant then. Also argumentative conception of justification concerns doxastic justification. So, the only generic conception left is the normative conception of justification.

If we apply the deontological conception to this situation, justification seems to be a matter of obligation, rather than permission. We are interested in what attitude we ought to accept, not what attitude we are permitted to accept. Now we are, of course, applying deontological term 'ought' directly to beliefs, and does not this suggest that we can voluntarily control our beliefs. Maybe we can avoid this suggestion and deny that the use of 'ought' implies voluntary control in this case. Then it could be the case that I ought to believe that *p* even though I cannot believe that *p* or cannot avoid believing that *p*.[39]

It is, of course, true that when I am deliberating what to believe in a certain situation, I assume that I am able to believe whatever it is that I judge that I ought to believe. Otherwise, I would take deliberation to be totally useless. However, I need not assume that belief is within my voluntary control. It is not so that I first make the judgement that I ought to believe that *p* and then make an effective decision to believe that *p*. When I reach the judgement that I ought to believe that *p*, the belief that *p* follows usually quite automatically. This is so because when I am deliberating whether to believe that *p*, I consider what evidence I have for and against the proposition that *p*. And when the evidence seems to support *p*, I judge that I ought to believe that *p* and come to believe that *p*. It is my awareness of the evidence that causes me to make the judgement and to form the relevant belief. We need not make any assumptions about my being able to control my beliefs intentionally. It is the evidence that controls my beliefs.

So the deontological conception of justification seems to be after all something to which we need to pay attention in epistemology. And it seems to support evidentialism and a form of internalism. Considerations about deliberation suggest that we should opt for awareness internalism because only evidence that we are actually aware of can affect our deliberation and our making a particular epistemic judgement.

According to this approach, the central epistemological question is 'What ought I to believe now?' and in answering it I can only take account of something that I am currently aware of. But does it not lead inevitably to the conclusion that it is impossible to have any unjustified beliefs? The same evidence that causes the judgement that I ought to believe that *p* also causes the belief that *p*. The ought judgement seems to be a useless intermediary between the evidence and the belief. So this sort of deontological concept of justification may not after all be so important in epistemology.[40]

What has gone wrong is that we have focused exclusively on the time of deliberation and on – what we may call – the current state justification of believing a proposition. Maybe, the role of epistemic judgements in guiding our own belief formation is more indirect. They may help us to notice our past mistakes and to learn to avoid them in the future. Perhaps, we have realised that the belief we formed is false, and we want to evaluate whether the mistake we made was excusable or whether we are to blame ourselves for it. If we want to avoid similar mistakes in the future, it is important to make a distinction between cases in which there is nothing we could have done to avoid the mistake and cases in which we could have done something and are therefore responsible for the mistake. When we do not excuse ourselves, we may blame ourselves for not paying attention to some evidence, for not listening to other people, for not being careful enough in our inferences and so on. All this suggests that justified belief is excusable belief and that excusable belief is belief that is the result of epistemically responsible action. Whether a belief is the result of such action need not be accessible to the agent at the time of belief formation. It is something that the agent assesses afterwards for the purpose of trying to learn from her past mistakes. Neither are we committed to doxastic voluntarism because epistemic obligations concern now actions that influence belief.

Because the term 'justified' is more appropriately used for excusable or permitted belief than for obligatory belief, the deontological conception does not seem to support internalism after all, and neither is it committed to doxastic voluntarism.

XVII CONCLUSION

Epistemic justification has turned out to be a messy issue. There is not just one question that epistemologists raise when they ask questions about the justification of our beliefs. There are several such questions that serve different interests and utilise different concepts of justification. We cannot even begin to understand the disagreement about the conditions of justification until we have some idea of these different interests and concepts. This is what I have tried to give in this survey.

I have focused especially on three conceptions of justification: the deontological (or more widely normative) conception, the argumentative conception and the conception of justification as something needed for knowledge. It has turned out that they all lead to somewhat different substantive theories of justification. It has also turned out that they may all give us concepts that we ought to look for in epistemology. We should not really use the term 'justification' for all of them to avoid confusion. I have done so here for the purpose of describing the current debate. The term 'justification' is after all used in such a variety of ways in contemporary epistemology. The next step will be to coin new words for the different concepts.

Markus Lammenranta
University of Helsinki

NOTES

[1] William Alston (1993) makes a similar point.

[2] See Burnyeat, 1980, p. 134; and Everson, 1990, p. 4.

[3] I follow here William Alston's excellent discussion of these conceptions of justification. See especially his 'Concepts of Epistemic Justification' and 'The Deontological Conception of Epistemic Justification', both reprinted in Alston, 1989.

[4] This example was suggested to me by Alvin Goldman. For different kinds of voluntary control, see Alston, 1989, pp. 122-136.

[5] See Alston, 1989, pp. 136-142.

[6] Indeed, William Alston is used to remind us of this confusion.

[7] See Van Cleve, 1985, p. 91.

[8] See Sosa, 1991, p. 156.

[9] See Kim, 1988, p. 399.

[10] See Chisholm, 1989, pp. 85-86, and Haack, 1993, pp. 73-94.

[11] For this way of understanding the regress problem, see Sosa, 1991, pp. 176-177, and Moser, 1985, pp. 107-116, 1989, pp. 56-60.

[12] It is to be noted that actually this argument does not refute minimal foundationalism according to which sufficient justification requires both a basic source and the support from other beliefs.

[13] BonJour, 1985, pp. 33, 59-60.

[14] See Foley, 1979.

[15] There seems to be a fatal objection to coherence theory of truth. It creates a conceptual infinite regress. See Fumerton, 1995, pp. 138-140.

[16] See Cohen's and Lehrer's counterexample to reliabilism below.

[17] See e.g. Moser, 1985, p. 85.

[18] Sosa, 1991, p. 183; Moser, 1985, p. 137.

[19] Indeed, Keith Lehrer (1997, pp. 61-64) has recently argued that, to preserve her dialectical advantage against foundationalism, a coherentist should reject the principle of supervenience.

[20] For example Annis (1978), Rorty (1980) and Williams (1980).

[21] See especially Rorty, 1991, p. 23, and, 1998, pp. 1-11.

[22] Rorty (1998, pp. 1-2) says also that we should replace the question of objective truth by the question of usefulness. But the question of what it is useful to say raises also the question of objective truth. Is it objectively true that it is useful to say so and so? If Rorty says that also this question should be replaced by the question of usefulness, he faces a regress. For this sort of criticism of Rorty, see Moser, 1999, pp. 82-86.

[23] Goldman, 1979, p. 14.

[24] Firth, 1978, pp. 217-219.

[25] Goldman, 1979, p. 21; Firth, 1978, pp. 219-220; Kvanvig and Menzel, 1990, pp. 247-258.

[26] Feldman, 1985; Goldman, 1986, pp. 49-51, 1992, pp. 434-435.

[27] This formulation is given by Feldman (1992, p. 119).

[28] Feldman, 1988a; 1992.

[29] See Schmitt, 1992, pp. 120-129.

[30] See Alston, 1989, p. 233.

[31] See Alston, 1989, pp. 189-191.

[32] Moser, 1989, p. 111.

[33] Alston, 1989, p. 200.

[34] Lehrer, 1974, p. 156.

[35] Lehrer, 1974, pp. 156-157; Rorty, 1980, p. 159.

[36] See Sosa, 1991, pp. 96-97, 206-207.

[37] Audi, 1998, pp. 202-203.

[38] See Alston, 1989, p. 178.

[39] See Feldman and Conee, 1987, pp. 335-336; Feldman, 1988b, pp. 237-243.

[40] I have elsewhere tried to develop an account of a non-deontological evaluative concept of justification that might better fit the role that epistemic judgements have in deliberation. See Lammenranta, 1998.

REFERENCES

Alston, W. P.: 1989, *Epistemic Justification: Essays in the Theory of Knowledge*, Cornell University Press, Ithaca.

Alston, W. P.: 1993, 'Epistemic Desiderata', *Philosophy and Phenomenological Research* **53**, 527-551.

Annis, D. B.: 1978, 'A Contextualist Theory of Epistemic Justification', *American Philosophical Quarterly* **15**, 213-219.

Audi, R.: 1994, 'Dispositional Beliefs and Dispositions to Believe', *Noûs* **28**, 419-434.

Audi, R.: 1998, *Epistemology*, Routledge, London.

Blanshard, B.: 1939, *The Nature of Thought*, Allen & Unwin, London.

BonJour, L.: 1980, 'The Externalist Theories of Empirical Knowledge', in P. A. French, T. E. Uehling, Jr., and H. K. Wettstein (eds.), *Midwest Studies in Philosophy 5*, University of Minnesota Press, Minneapolis, pp. 53-74.

BonJour, L.: 1985, *The Structure of Empirical Knowledge*, Harvard University Press, Cambridge.

Burnyeat, M.F.: 1980, 'Aristotle on Understanding Knowledge', in E. Berti (ed.), *Aristotle on Science: The Posterior Analytics*, Padua.

Chisholm, R.: 1989, *Theory of Knowledge*, Third Edition, Prentice-Hall, Englewood Cliffs.

Craig, E.: 1990, *Knowledge and the State of Nature*, Clarendon Press, Oxford.

Davidson, D.: 1986, 'A Coherence Theory of Truth and Knowledge', in E. LePore (ed.), -*Truth and Interpretation*, Basil Blackwell, Oxford, pp. 307-319.

Everson, S. (ed.): 1990, *Epistemology*, Cambridge University Press, Cambridge.

Feldman, R.: 1988a, 'Having Evidence', in D. Austin (ed.), *Philosophical Analysis*, Kluwer, Dordrecht, pp. 83-104.

Feldman, R.: 1988b, 'Epistemic Obligations', in J. E. Tomberlin (ed.), *Philosophical Perspectives 2, Epistemology*, Ridgeview Publishing Company, Atascadero.

Feldman, R.: 1989, 'Foley's Subjective Foundationalism', *Philosophy and the Phenomenological Research* **50**, 149-158.

Feldman, R.: 1992, 'Evidence', in J. Dancy and E. Sosa (eds.), *Companion to Epistemology*, Blackwell, Oxford, pp. 119-122.

Feldman, R. and Conee, E.: 1987, 'Evidentialism', in P. K. Moser and A. van der Nat (eds.), *Human Knowledge: Classical and Contemporary Approaches*, Oxford University Press, Oxford, pp. 334-345.

Firth, R.: 1978, 'Are Epistemic Concepts Reducible to Ethical Concepts?', in A. I. Goldman and J. Kim (eds.), *Values and Morals*, D. Reidel, Dordrecht, pp. 215-229.

Foley, R.: 1979, 'Justified Inconsistent Beliefs', *American Philosophical Quarterly* **16**, 247-258.

Foley, R.: 1987, *The Theory of Epistemic Rationality*, Harvard University Press, Cambridge.

Foley, R.: 1989, 'Reply to Alston, Feldman and Swain', *Philosophy and the Phenomenological Research* **50**, 169-188.

Foley, R.: 1993, *Working Without a Net*, Oxford University Press, Oxford.

Fumerton, R.: 1995, *Metaepistemology and Skepticism*, Rowman and Littlefield, Lanham.

Gettier, E.: 1963, 'Is Justified True Belief Knowledge?', *Analysis* **23**, 121-123.

Ginet, C.: 1975, *Knowledge, Perception, and Memory*, D. Reidel, Dordrecht.

Goldman, A. I.: 1979, 'What Is Justified belief?', in G.S. Pappas (ed.), *Justification and Knowledge*, D. Reidel, Dordrecht, pp. 1-24.

Goldman, A. I.: 1986, *Epistemology and Cognition*, Harvard University Press, Cambridge.

Goldman, A. I.: 1992, 'Reliabilism', in J. Dancy and E. Sosa (eds.), *Companion to Epistemology*, Blackwell, Oxford, pp. 433-439.

Haack, S.: 1993, *Evidence and Inquiry*, Blackwell, Oxford.

Harman, G.: 1973, *Thought*, Princeton University Press, Princeton.

Kim, J.: 1988, 'What is "Naturalized Epistemology"', in J. E. Tomberlin (ed.), *Philosophical Perspectives, 2, Epistemology*, Ridgeview Publishing Company, Atascadero.

Kornblith, H.: 1983, 'Justified Belief and Epistemically Responsible Action', *The Philosophical Review* **92**, 33-48.

Kornblith, H.: 1989, 'The Unatainability of Coherence', in J.W. Bender (ed.), *The Current State of the Coherence Theory*, Kluwer, Dordrecht, pp. 207-214.

Kvanvig, J. L., and C. Menzel.: 1990, 'The Basic Notion of Justification', *Philosophical Studies* **59**, 235-261.

Lammenranta, M.: 1988, 'The Normativity of Naturalistic Epistemology', *Philosophia: Philosophical Quarterly of Israel* **26**, 337-358.

Lehrer, K.: 1974, *Knowledge*, Clarendon Press, London.

Lehrer, K.: 1990, *Theory of Knowledge*, Westview Press, Boulder.

Lehrer, K. and S. Cohen: 1987, 'Justification, Truth, and Coherence', in P.K. Moser and A. van der Nat (eds.), *Human Knowledge*, Oxford University Press, Oxford, pp. 325-334.

Moser, P.K.: 1985, *Empirical Justification*, D. Reidel, Dordrecht.

Moser, P.K.: 1989, *Knowledge and Evidence*, Cambridge University Press, Cambridge.

Moser. P.K.: 1999, 'Realism, Objectivity, and Skepticism', in J. Greco and E. Sosa (eds.), *The Blackwell Guide to Epistemology*, Blackwell, Oxford.

Plantinga, A.: 1993a, *Warrant: The Current Debate*, Oxford University Press, Oxford.

Plantinga, A.: 1993b, *Warrant and Proper Function*, Oxford University Press Oxford.

Pollock, J.: 1987, *Contemporary Theories of Knowledge*, Hutchinson, London.

Putnam, H.: 1981, *Reason, Truth and History*, Cambridge University Press, Cambridge.

Rescher, N.: 1985, 'Truth as Ideal Coherence', *Review of Metaphysics* **38**, 795-806.

Rorty, R.: 1980, *Philosophy and the Mirror of Nature*, Basil Blackwell, Oxford.

Rorty, R.: 1991, *Objectivity, Relativism, and Truth: Philosophical Papers, Vol. 1*, Cambridge University Press, Cambridge

Rorty, R.: 1998, *Truth and Progress: Philosophical Papers, Vol. 3*, Cambridge University Press, Cambridge.

Schmitt, F.: 1992, *Knowledge and Belief*, Routledge, London.

Sellars, W.: 1963, *Science, Perception and Reality*, Routledge & Kegan Paul, London.

Sosa, E.: 1991, *Knowledge in Perspective: Selected Essays in Epistemology*, Cambridge University Press, Cambridge.

Swain, M.: 1981, *Reasons and Knowledge*, Cornell University Press, Ithaca.

Van Cleve, J.: 1985, 'Epistemic Supervenience and the Circle of belief', *The Monist* **68**, 90-104.

Williams, M.: 1980, 'Coherence, Justification, and Truth', *Review of Metaphysics* **34**, 243-272.

Wittgenstein, L.: 1969, *On Certainty*, trans. by D. Paul and G. E. M. Anscombe, Blackwell, Oxford.

PAUL WEIRICH

BELIEF AND ACCEPTANCE

Tradition takes knowledge to be true, justified belief. Accordingly, any theory of knowledge needs a supplementary theory of belief. It should say what belief is, and especially what features of a belief determine whether the belief is true and justified. Epistemology does not have its own account of belief, but draws on interested branches of philosophy. I present their views and the implications for epistemology. The first section provides general orientation. The next three sections explore belief as viewed by philosophy of language, logic, and philosophy of mind. The section following these considers whether epistemology gains by introducing an attitude of acceptance to supplement or replace ordinary belief. The final section draws conclusions for epistemology.

1. THE LIE OF THE LAND

People, animals, and perhaps some computers have beliefs. Belief may be invested in a person, idea, report, sign, proposition, or sentence (perhaps without even understanding it). A propositional belief's content is also called a belief. For example, the sentence, 'We share the belief that snow is white,' calls a common content of our mental states a belief. Among states of belief with ideational content, theorists distinguish occurrent and dispositional beliefs, conscious and unconscious beliefs, explicit and tacit beliefs, *de re* and *de dicto* beliefs, and conditional and nonconditional beliefs.

I investigate nonconditional, *de dicto*, propositional beliefs of humans, taken as mental states that are explicit and conscious, but not necessarily occurrent. Beliefs taken this way are persistent. They are held even when not entertained. They are roughly dispositions to assent to propositions when those propositions are entertained. Beliefs of this type are the ones epistemology studies most.

Taking the object of a belief as a proposition rather than a sentence allows people with different languages to believe the same thing. An English speaker and a French speaker may both believe the proposition that snow is white although the first expresses the belief saying "Snow is white" whereas the second expresses the belief saying "La neige est blanche." Taking the object of a belief as a proposition also allows the same sentence to express different objects of belief in different circumstances. I may believe that he is tired (pointing at Jones) and also believe that he is tired (pointing at Smith). Although the expression of both beliefs uses a single sentence, 'He is tired,' that sentence expresses different propositions in different segments of the report of those beliefs.

We take the propositions forming the objects of beliefs to be conceptually articulated. As such, a proposition is a structured entity, commonly with a concept

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 499-520.
© 2004 *Kluwer Academic Publishers.*

representing a subject and a concept attributing a property, rather than an unstructured entity, such as a set of possible worlds. This view allows for multiple necessary propositions, each of which is true in all possible worlds.

What makes a belief true? Consider a belief with the content proposition *p*. The belief is true if and only if *p* is true. An account of true belief thus rests on an account of true propositions. The standard account of truth is a correspondence theory according to which truth consists in correspondence with reality. Different versions of the theory explicate this idea in different ways. Some rival views take truth to be a matter of coherence, or what science eventually settles upon, or take truth to be an empty property whose attribution to a proposition adds nothing to the proposition's assertion.

What justifies a belief, in particular, a propositional belief? This is the central question of epistemology. It is central because epistemology aims primarily to characterize the true beliefs that form knowledge. It seeks the true beliefs with the backing necessary and sufficient for knowledge. Some epistemologists, such as Plantinga (1993a, 45), argue that justification is not necessary for knowledge. They take some substitute, such as warrant, to be necessary for knowledge. But we may regard the substitute as a type of justification broadly construed, so that justification remains the main epistemological issue.

Epistemologists generally hold that a special type of justification is required for knowledge, different from the type required for rational belief. Suppose that Smith justifiably believes that Jones is in Barcelona, and so justifiably believes that Jones is in Barcelona or ticket 100 wins the lottery. If Jones is not in Barcelona but ticket 100 wins, then Smith's belief in the disjunction is true and justified, but nonetheless is not knowledge. Such Gettier problems show that a special, nondefective type of justification is necessary for knowledge. Because believing just in case one has the type of justification required for knowledge is beyond human control, some theorists, such as Foley (1993, 85), want epistemology to put aside its concern with knowledge and concentrate instead on rational belief.

Even when focusing on rational belief, epistemologists are typically concerned with justification by certain types of reasons only. Loyalty may justify a person's belief that a loved one is innocent of a crime. Such pragmatic justification does not provide knowledge, however, nor even an epistemic warrant for belief. Epistemologists put aside pragmatic reasons for belief and study reasons stemming from the goal of belief in the truth.

Also, epistemologists are typically interested in standards of justification that accommodate the cognitive limits of humans. In some fields, it may be claimed that one is justified in believing all the logical consequences of one's beliefs. Then if Goldbach's conjecture is true, one is justified in believing it, since it follows from one's mathematical beliefs. If one believes it, one has knowledge of it according to the traditional definition of knowledge. This result is wrong since we humans lack logical perspicuity, and so lack adequate grounds for belief in some logical consequences of our beliefs. To resolve the problem, epistemologists explicate knowledge in terms of justification for humans. They take account of the distinctive features of human cognition.

According to a traditional view in epistemology, a belief is justified if and only if its content bears the appropriate relationship to the believer's evidence. This evidence may be the content of certain self-justifying or foundational beliefs, or experience without propositional content, or some combination of these two. Logical relationships between the contents of evidence and the contents of beliefs may play an important role in justification on this view, but so may other grounding relationships between evidence and beliefs.

Some rival accounts of justification appeal to criteria less attentive to a belief's content. They are procedural rather than substantive. The procedures may be external rather than internal, that is, they may involve causal relations to the external world, not just mental operations. One rival account focuses on the process that produced a belief. Reliabilism claims that a belief is justified if and only if it is the product of a reliable belief-forming process. Another rival account focuses on the condition of a believer's epistemic equipment. It says that a belief is justified if and only if it is formed by a properly functioning belief-forming system. Some theorists advance procedural justifications for some types of belief, such as perceptual beliefs, and substantive, evidential standards for other types of belief, such as inferential beliefs.

In the next three sections, I discuss belief as viewed by philosophy of language, logic, and philosophy of mind. I emphasize points about the content of a belief since a belief's content is crucial for the belief's truth and justification according to traditional views such as the correspondence theory of truth and the evidential account of justification.

2. BELIEF VIEWED BY PHILOSOPHY OF LANGUAGE

One way of learning about a phenomenon is to consider how we talk about it. We can learn about belief by studying talk about belief. Philosophy of language formulates truth conditions for belief reports and propositions believed. Its analyses elucidate the nature of belief.

Belief is a subject's attitude toward a proposition at a time. A common account of belief reports holds that a sentence of the form *S* believes that *p* uses the that-clause to name the proposition believed. The embedded sentence names its sense rather than the truth value that is its ordinary denotation.

Frege introduced this view to resolve a puzzle about belief. Given that Hesperus is Phosphorus, how can the report that Caius believes that Hesperus is Phosphorus differ in truth value from the report that Caius believes that Phosphorus is Phosphorus? If the sentences expressing the beliefs' contents each name a truth value, then since their truth values are the same, Caius has the first belief just in case he has the second. To square analysis of belief reports with facts about their veracity, Frege distinguished two kinds of meaning: sense and denotation. He held that in belief contexts the semantic value of an expression is its ordinary sense. Hence the embedded sentence of a belief report names a proposition, not a truth value. In our example the identities embedded in the belief reports name the propositions they ordinarily express. The two identities, even if equivalent, ordinarily express different propositions and thus name different propositions in the two belief reports. Frege's analysis also offers a more detailed explanation of the failure of substitution of identicals in belief contexts. It explains why 'Phosphorus' may not substitute for 'Hesperus' in the embedded sentences of belief reports. The two names have different senses and so name different things in belief contexts.

Russell's view of propositions challenges Frege's analysis of belief reports. Russell's view allows for singular propositions having as a constituent a physical object rather than an associated sense. Examples are propositions expressed by subject-predicate sentences in which the subject term is a proper name, indexical, or demonstrative. These terms contribute only an object to a proposition, not a sense. Their reference is direct, unmediated by a sense.

The phenomenon of direct reference resurrects Frege's puzzle. If the semantic contribution of the proper names 'Hesperus' and 'Phosphorus' is the same object, what is the distinction between Caius's believing that Hesperus is Phosphorus, and Caius's believing that Phosphorus is Phosphorus? The object of each belief is the same proposition. Frege made the objects of belief fine-grained by using words' senses to compose them. But the grain obtained from words is not fine enough given that some words lack senses.

Besides believer and proposition believed, we need a third factor to individuate belief more finely—a way of believing, a mode of presentation, or a mental representation. The third factor may have the role of a Fregean sense, but it may be a different sort of entity, perhaps not word-bound, shareable, and graspable as Fregean senses are. The third factor may be specified contextually and pragmatically. It may be tacitly characterized by a belief report, according to pragmatic rules of tacit reference that consider context, or it may even be the semantic value of an implicit expression.

Crimmins (1992) formulates such a fine-grained analysis of ideational belief, or belief conceptually articulated, having notions and ideas as constituents, and so having a structured proposition as content. He says that an instance of explicitly believing (as opposed to tacitly believing) holds in virtue of a relation between an agent, time, cognitive particular, and the proposition that is the content of the cognitive particular at the time (57). The cognitive particular that realizes an explicit belief is a mental representation. An explicit belief is a persisting structured representation, a cognitive particular with propositional content (53–8). Its structure is similar to the structure of the proposition that is its content.

A mental representation is individuated by its causal role or cognitive function, but is a concrete particular and is not sharable. It is not a type. A mental representation realizing a belief is commonly composed of a notion of the belief's subject and an idea of a property. The notion and idea are also cognitive particulars, not types (78–9).

Mental representations may be classified according to type. A mental representation has a content at a time. Mental representations with the same content at a time fall into the same content type. Two people may have mental representations with the same classification. We assert that they do when we say that the two believe the same thing; we attribute explicit beliefs with the same content.

Besides classifying a belief according to its content, we may classify it according to its structure, or the thought map it realizes. A thought map is an abstract structured entity listing notions and ideas, and roles they occupy in a potential belief (61; Chp. 4). One believes a proposition according to a thought map.

Thought maps specify ways of believing and add grain to an account of belief reports, but insufficient grain according to Crimmins. He argues (44–9) that beliefs are insufficiently individuated by shareable belief states (modes of presentation, ways of believing, or thought maps). Consider an elaboration of one of Perry's examples. Hume has a belief he expresses by saying "I wrote the *Treatise*." The madman Heimson has a belief he expresses by saying "I wrote the *Treatise*." Hume and Heimson are in the same belief state, although the contents of their beliefs are different. Suppose Heimson accepts "I wrote the *Treatise*" and does not accept "I wrote *A Treatise on Human Understanding*." It seems he both believes and does not believe that he wrote the *Treatise*. There is just one agent, time, and proposition relevant. Heimson seems to be both in and not in the belief state associated with the first sentence.

To add grain sufficient for resolving the inconsistency, Crimmins appeals to the concrete particulars, mental representations, forming instances of believing. He individuates beliefs according to the mental representations that realize them (139). These cognitive particulars finely individuate Heimson's beliefs. Context yields the appropriate cognitive particular to settle the accuracy of a report that Heimson believes that he wrote the *Treatise*. The belief report is underarticulated, according to Crimmins, and makes tacit reference to an internal, mental representation. Context determines the representation to which the report tacitly refers. A belief report is true if the content proposition is believed by means of the particular mental representation to which the report tacitly refers.

A belief report's tacit reference to a mental representation works by means of tacit reference to the thought map the representation follows. For instance, take the report, 'Susan (explicitly) believes that Dean Smith fired Tom' (155). The report says that Susan (explicitly) believes a proposition in a tacitly characterized way. Her belief, a concrete particular, is of a type specified by a thought map giving constituent notions and ideas, and roles they occupy. Along with explicit reference to the proposition believed, the belief report tacitly refers to Susan's notions of Dean Smith and of Tom, which are notions associated with the names used in the report. The report asserts an (explicit) belief tacitly claimed to be of a certain type, a type in which Susan's notions of Dean Smith and Tom are responsible for roles in a thought map characterized by the content sentence. According to pragmatic rules, the context and words of the report tacitly indicate a thought map and notions or ideas responsible for filling roles in the map. They indicate more or less fully the (concrete particular) belief's type, its structure and constituents.

Like Frege, Crimmins makes the truth of belief reports depend on modes of presentation. Unlike Frege, the modes are the subject of tacit rather than explicit reference, and they are concrete particulars rather than abstract universals shared and understood. Like Russell, Crimmins allows for singular propositions with objects as constituents. Unlike Russell, Crimmins does not hold that belief in such propositions requires direct acquaintance with those objects (204–5).

Crimmins's account of belief reports is similar to Richard's (1990). Richard states the following truth condition for a report that a subject believes that *p*. The report is true in a context if and only if in the context *p* represents one of the sentences that constitute the subject's thoughts (4, 140). In a true belief report the semantic, external, and public represents the psychological, internal, and private. The grain of representation is given by a *Russellian* annotated matrix or RAM specifying the semantic values of constituents of *p* and also the sentence constituting the subject's corresponding thought (137–8). Crimmins's account differs in appealing to concrete mental representations to individuate beliefs, and in appealing to unarticulated constituents of belief reports to resolve certain puzzles about belief (29–32).

3. BELIEF VIEWED BY LOGIC

Logic, taken broadly as the study of reasoning, treats belief. Principles of inference explain why some beliefs warrant others. The principles of inference attend to the content of beliefs and rely on relations between propositions such as entailment. Some formulations of the principles do not explicitly mention belief, but nonetheless logic taken broadly aims to provide ultimately an account of justified inference. Logic has implications concerning belief because it formulates rules of inference that govern beliefs although the way in which they govern beliefs is controversial.

The connection between familiar logical principles and justified belief is not straightforward. Take the rules of deductive inference. How do they govern belief? A common view is that rational belief obeys those rules. That is, if it is rational to believe each premiss of a valid deductive argument, then it is rational to believe the conclusion, at least when the validity of the argument is recognized. Does this mean that a rational agent must draw all the obvious deductive consequences of his beliefs? That demand is too strong for humans, given our mental limits. Do human limits also excuse some violations of the rules of deductive inference, beliefs at variance with the obvious deductive consequences of other beliefs? Suppose a sincere but modest author believes each claim in her book but concedes she is bound to be mistaken somewhere. She believes each claim but also the denial of their conjunction. Are her inconsistent beliefs irrational? Or suppose that a gambler believes of each ticket in a lottery that it will lose but nonetheless believes that the lottery will yield a winning ticket. Are his inconsistent beliefs irrational? The charge of irrationality seems too harsh in these cases.

Kyburg (1997) takes the rules of deduction as applying to reasonable belief in the following way. When the conjunction of the premisses of a valid deductive argument is reasonably believed, the warrant for the conjunction is transmitted to the conclusion and makes belief in the conclusion permissible, but not obligatory. In cases where it is given only that the premisses are reasonably believed individually, he rejects the rule of adjunction (conjunction introduction) because of the lottery paradox and similar cases where the probability of each premiss is high but the probability of their conjunction is low (110–114).

Foley (1993) takes the rules of deduction as applying to attitudes besides belief, such as assumption and taking as evidence (166, 168, 196–7). He rejects universal application of the rules of deduction to belief. On his view, rational beliefs that *p* and that *q* do not mandate a belief that *p* & *q* (166). Also, although a person may not rationally believe a proposition she knows is contradictory, she may rationally believe a set of propositions she knows is inconsistent (164–5). The lottery and preface paradoxes move Foley to this position. He acknowledges that the inconsistency of a set of beliefs is relevant to their rationality since it establishes their inaccuracy, but holds that it is not decisive against their rationality since rational belief serves the goal of comprehensiveness as well as the goal of accuracy (94–6). Rational beliefs may tolerate an inaccuracy to achieve scope.

Harman divorces deductive logic and reasoning involving beliefs (1986; Chp. 2, App. A). He thinks that reasoning is governed by psychological principles of immediate implication and immediate inconsistency, whereas logic investigates a nonpsychological relationship of logical consequence. Perhaps rules of deduction can be reformulated in terms of immediate implication and inconsistency. But the details raise many unresolved issues.

In contrast, some theorists think that logic does regulate reasoning, but not by imposing inviolable standards. Kaplan takes conformity with the rules of deductive inference, in particular, deductive closure and consistency, as a regulative ideal for beliefs (1996, 36–8). Rationality requires conformity only from ideal agents in ideal circumstances. Other agents, if rational, aspire to conform although cognitive limits provide excuses for failures. Excused nonconformity is open to a type of criticism, however, since it deviates from the ideal.

Even granting Kaplan's view that deductive logic establishes a regulative ideal for belief, an account of justified belief needs supplementary principles that explicitly accommodate the human situation. It needs principles that tell how to pursue the regulative ideal in the face of obstacles to its attainment. One barrier to using deductive logic to govern beliefs directly is that many beliefs are not certainties. Although deductive logic may regulate certainties, it makes no provision for uncertainty. Inductive logic undertakes this task. In looking for principles of logic that govern beliefs generally, we turn to it.

Some approaches to inductive inference directly target belief. They develop rules for belief revision in response to new evidence. Gärdenfors's (1988) inductive system and some systems of nonmonotonic reasoning have this orientation. However, the classical approach to inductive inference directly addresses probability rather than belief and assumes that conclusions about probability regulate belief. Classical statistics adopts this approach. So does Bayesianism, but unlike classical statistics it addresses epistemic probabilities, which are relative to evidence. Let us consider its main tenets.

Epistemic probabilities are rational degrees of belief formed under certain idealizations, which concern cognitive power and the like. Rational degrees of belief conform to the standard laws of probability given the idealizations. The idealizations yield partial explanations of human rationality, explanations that detail the workings of some factors bearing on rationality. Bayesians do not assess a person's rationality by measuring his distance from the standards the idealizations generate. The idealizations serve explanation rather than measurement of rationality.

According to Bayesianism, the immediate outcome of inductive reasoning is a degree of belief, the degree of belief in the conclusion warranted by the degrees of belief in the premisses. Strictly speaking, it is the degree of belief warranted by all the evidence, whose salient points are assumed to be summarized by the premisses and their epistemic probabilities. The standard probability laws stand behind the warrant in typical cases. Other warrant-generating principles are controversial.

A degree of belief is an attitude manifest in betting behavior. As a person's degree of belief that *p* varies from 0 to 1, the lowest price she is willing to pay for a bet that yields $1.0 if *p* is true and otherwise nothing typically varies from $0.0 to $1.0. The term 'degree of belief' is standard, although it is somewhat misleading. A low degree of belief that *p* does not entail a belief that *p*. In fact, if a person's degree of belief that *p* is 0, then typically she is certain that ~*p*.

Since Bayesian reasoning yields a degree of belief rather than a belief, it needs a supplementary rule connecting degree of belief with belief to yield a genuine inductive inference culminating in a belief. What is the relation between degree of belief and belief? A common view is that belief is degree of belief above a certain threshold. Various pragmatic factors set the threshold, and may set it imprecisely.

The lottery paradox challenges this view. If a lottery has enough tickets, the degree of belief that any given ticket will lose is above the threshold, but the degree of belief that some ticket or other will win is also above the threshold. The threshold view entails that a rational person may believe of each ticket that it will lose and yet believe that some ticket or other will win. As mentioned earlier, some theorists tolerate inconsistent beliefs in such cases, but many refuse to condone them. Is there an alternative to the threshold view of belief?

According to Kaplan, the Bayesian approach to inductive reasoning creates the following challenge to belief (1996, xi; Chp. 3, esp. 98–101; 102–3). Since the immediate topic of inductive reasoning is degree of belief, belief may be dispensable. Indeed, what, if anything, is belief? It is not certainty, or we have few beliefs, contrary to the common conception of belief. It is not degree of belief (or confidence) above a threshold, or we face the lottery paradox. Since the certainty and threshold views of belief seem to be the only candidates, Kaplan concludes that the ordinary concept of belief is incoherent, and that a substitute should take over its role in inquiry.

In fact, the ordinary concept of belief seems eliminable in favor of degree of belief not only in inductive reasoning but also in practical reasoning. Traditional principles of practical reasoning attempt to explain how beliefs, along with desires, warrant decisions and actions. Aristotle's principle of practical reasoning uses a belief and a desire as premisses and issues an action as conclusion. To illustrate, suppose a rational person desires a cup of coffee and believes she can get one at Starbuck's. According to Aristotle's principle, she goes to Starbuck's. That action is allegedly warranted by the premisses. However, Aristotle's principle makes no provision for competing desires, such as the desire to stay in the office. It makes no provision for alternative means of achieving the end, say, getting coffee in the department lounge. And it makes no provision for obstacles, say, a parade blocking the route to Starbuck's.

Bayesian decision theory furnishes better principles of practical reasoning. It says to begin by computing each option's expected utility. For each of an option's possible outcomes, multiply its epistemic probability and its subjective utility. Then add the products to obtain the option's expected utility. In standard cases, adopt an option whose expected utility is at least as great as the expected utility of any other option. In other words, maximize expected utility.

Such Bayesian principles of practical reasoning replace belief with degree of belief. They elbow belief out of its role in practical reasoning as well as in inductive reasoning. As a result, common explications of belief in terms of its role in theoretical and practical reasoning are in jeopardy. Belief does not seem to have such a role. Belief is not input for fundamental laws of logic. It may not be output either. Laws of logic may completely dispense with the ordinary concept of belief. Stich (1983) goes further. He argues that belief has roles in practical reasoning and in inference that it cannot fill simultaneously. As Kaplan does, he concludes that the concept of belief is incoherent (230–7). In general, he claims that cognitive science provides good reason to doubt the viability of the concepts of folk psychology.

The Bayesian challenge to belief argues forcefully that belief's role in logic is peripheral, but, I think, does not warrant the conclusion that belief in the ordinary sense has no role in logic. In particular, the challenge does not establish that belief in the ordinary sense is incoherent. The argument that belief is incoherent misstates belief's roles in inference and practical reasoning. The concept of belief is not rent apart by incompatible roles. Belief's quantitative analogue, degree of belief, governs inference and action. Belief may be explicated, not by its traditional roles in inference and practical reasoning, but rather by association with degree of belief, which now occupies those roles.

Foley (1993) advances one response to the Bayesian challenge. He argues that degree of belief does not make belief dispensable. Belief is a way of taking an intellectual stand, something we find worth doing despite the oversimplification involved (171–3, 201). We want an accurate black-and-white picture of the world. He takes degree of belief above a vague threshold, lower than 1 but higher than .5, as sufficient for belief (140–2, 170, 198). Degree of belief above the threshold is not necessary for belief, however, since degree of belief may be unformed in some cases where belief is warranted. He escapes the lottery and preface paradoxes by denying that deductive logic governs rational belief.

The main lines of Foley's view are plausible. Belief seems to be related to degree of belief as hotness is related to temperature. Progress in physics requires reliance on temperature. Today the water was hot and it boiled. Yesterday the water was hot but it did not boil. Why does being hot sometimes cause boiling and sometimes not? Physics cannot state a regularity using the property of being hot. That property is not precise and depends on factors outside the taxonomy of physics such as sensation. Physical laws replace the property of being hot with temperature. Instead of saying that water boils when hot, they say that it boils at 100 degrees Centigrade. Although physical theory does not need the property of being hot, we may explain hotness by its relation to temperature. Being hot is having a temperature above a context-sensitive vague threshold.

Similarly, logic cannot progress if it relies on belief. Logic needs something more precise, and something not dependent on factors outside the taxonomy of logic. Because of its connection with assertion, belief is dependent on nonlogical factors such as rules of conversation. Degree of belief has the precision and independence necessary for principles of reasoning. Although basic laws of logic do not need belief, we may explain belief by its relation to degree of belief. Belief is degree of belief above a vague threshold.

In the face of the lottery and preface paradoxes, we need not adopt Foley's tolerance for violations of deductive principles. We may appeal to the context-sensitivity of a person's threshold for belief. The threshold is context-sensitive because belief is connected with assertion, which is context-sensitive, being a pragmatic matter. In the lottery paradox, the threshold may be set so high for an individual lottery ticket that belief that it will lose does not ensue. In the preface paradox, the author believes each claim in her book but not their conjunction. Rather than accept a failure of deductive closure, we may invoke the threshold's sensitivity to inferential context. Although the threshold for each claim taken in isolation is exceeded, the threshold for each claim taken as a premiss in the argument for their conjunction is not exceeded. The context of the argument boosts the threshold for belief in the premisses.

In view of the paradoxes, we may agree with Kaplan that belief does not reduce to degree of belief above a fixed threshold, but nonetheless hold that it does reduce to degree of belief above a vague context-sensitive threshold (or in nonquantitative cases, adoption of a constraint placing degree of belief, if formed, above the threshold). We may take belief in a context, roughly, as sufficient degree of belief in the context to motivate assertion (but not necessarily cause it) given only cognitive goals such as comprehensiveness and accuracy. The variable threshold is sensitive to factors influencing assertion in the context, including conventions of consistency.

Although belief has a peripheral role in logic, logic regulates belief by regulating degree of belief, to which belief is related. The details of the relation between belief and degree of belief, and the standards logic imposes on belief, remain an open issue.

4. BELIEF VIEWED BY PHILOSOPHY OF MIND

Taking belief as a mental entity with content, numerous questions arise. What sort of mental entity is belief? How does a mental entity have content? What sort of content does it have? Let us start with belief's characterization. I will assume that some form of physicalism is correct.

Once physicalists held that beliefs are brain states. This identity theory later ceded to the view that beliefs are functionally characterized states realized by brain states. Many physicalists now hold that believing a proposition *p* is a functional property of a brain. They identify belief by its causal role. According as belief is taken narrowly or widely, its causal role is narrow or wide. Taken narrowly, its causal role is internal. Taken widely, its causal role is not limited to the believer's internal psychology but may encompass his physical environment. A partial characterization may say that belief is a property caused by events such as being told that *p*, and resulting in events such as assertion that *p*. Wide functionalism is the heir of Ryle's (1949, 133–5) dispositional characterization of belief, according to which a belief is a tendency, proneness, or propensity to certain types of behavior, thoughts, and feelings.

If physical, how can a belief have propositional content? How can a physical state represent a proposition? Of course humans invest some physical states with meaning. We invest patterns of ink marks on paper with meaning. But we do not invest our beliefs with meaning. They have their content naturally. What natural process invests such physical states with meaning? Why do states realized by our brains have content naturally?

Dretske (1988) offers a physicalistic account of intentionality, or mental representation. Other naturalistic accounts are provided by Millikan (1984) and Fodor (1990, Chp. 4), but I will describe only Dretske's account. He says that brain states *indicate* features of the external world. They do this by being causally dependent on and thus correlated with those features. Many natural phenomena are indicators. The rings in a tree's stump indicate the tree's age when it was cut. If in a system an indicator has the function of indicating features of the world, then it *represents* those features. The tree rings do not have the function of indicating the tree's age, and so do not represent it. But in a thermometer the column of mercury has the function of indicating temperature and so represents temperature. Some components of natural systems, as opposed to artificial systems, have the function of indicating facts because of evolution rather than by design. Just as the heart has the function of circulating blood, so the parts of the brain connected to the eyes have the function of indicating the shape and movement of nearby middle-sized objects. Since human brains have the function of indicating facts about the worlds, they represent those facts.

Some brain states represent facts not because of evolution but because of learning. Suppose a brain state that indicates a fact comes to control movement, because of conditioning, say. The state acquires the job of (partially) controlling movement because it indicates the fact. It thereby also acquires the function of indicating the fact, and so represents the fact. Brain states with the dual function of indication and control are beliefs. They trigger movement, but do so because they indicate facts. Their being indicators explains their having become controllers. A state's representational content explains behavior taken as a causal process starting with the brain state and (typically) culminating in movement (given appropriate desires).

Dretske says, "Beliefs are precisely those internal structures that have acquired control over output, and hence become relevant to the explanation of system behavior, in virtue of what they, when performing satisfactorily, indicate about external conditions" (84). Although beliefs are drafted as indicators, they are not inerrant since a state's function may be indication despite cases where it does not indicate. In those cases it misrepresents the external world.

As an illustration of Dretske's account of belief, take the brain state arising when raindrops hit the face. The presence of the state indicates rain. As a result, it comes to trigger seeking shelter. As it acquires this role, it also acquires the function of indicating rain, and so becomes a representation of rain and thereby the belief that it is raining. The content of the belief explains behavior, taken as a process that starts with the brain state and culminates in seeking shelter. Dretske takes behavior as a brain-state-to-movement causal process to accommodate the distinctive causal role of the content of a belief. As a triggering cause of movement, a belief's content does not matter; its realization as a brain state triggers movement. Its content matters in a different way. A belief's content explains why the belief is a trigger. It is a structuring cause, a cause of the brain state's being a trigger of movement. That the belief's content helps explain behavior makes the belief a reason for the behavior not just a cause of the behavior's motive product.

According to Dretske, a belief has propositional content. Its content is articulated. It has a topic and makes a comment about the topic (70–2). Also, the brain state that realizes the belief (functionally characterized) is a repeatable state, one that may take on a function because it is a reliable indicator. It is not a concrete particular. However, each instance of the brain state is a concrete particular and is a belief in the sense of being a realization of a belief state. Dretske's account of belief therefore is reconcilable with Crimmins's account of belief reports.

Dretske's account of belief is a simplified sketch. It addresses beliefs acquired in sensory discrimination by an organism with standing background desires. Animals with unsophisticated cognitive apparatus may have beliefs in this sense. Perhaps their beliefs should be called "protobeliefs" (107). Perhaps genuine beliefs are part of a complex network of beliefs, desires, and intentions controlling behavior, and arise only in cognitively sophisticated organisms. Dretske explains how, given this perspective, his account of belief may be expanded to provide an account of genuine belief by making more versatile a belief's interaction with desires and other beliefs (Chps. 5, 6).

For simplicity, Dretske also takes beliefs as maps by which we steer, whereas in fact probability is the guide of life. His account of belief should be revised so that degrees of belief have the role of directing action and inference. Then it may define belief in terms of degree of belief as Section 3 explained.

Dretske's account of belief makes the content of a belief depend on external factors rather than internal factors exclusively. Some find this consequence objectionable. They argue that a belief's content is internally accessible. Introspection reveals what the belief is about. This is possible, they say, only if the content of the belief is an internal matter.

A powerful defense of externalism can be marshalled, however. Beliefs have linguistic content, and language is social. As a result, a belief's content may be individuated externally. It may be broad rather than narrow. Numerous examples argue that this is so. Take an example of Putnam's, which he uses to make a point about word meaning (1975, 223-4), but which we can use to make a similar point about belief content. Suppose that Earth has a twin in which water is replaced with another chemical substance XYZ indistinguishable by the human senses from water. You believe that water fills lakes here. Your twin on Twin Earth believes that XYZ fills lakes there. Yet you and your twin are in the same psychological state. The difference in the content of your beliefs is therefore external. Beliefs have broad content.

An example of Burge's (1979) makes the same point. A patient believes that he has arthritis in his thigh. He has the false belief because he incompletely understands the notion of arthritis. He does not know that arthritis occurs only in joints. Imagine a counterfactual case where his physical and non-intentionally characterized past is the same but his linguistic community uses 'arthritis' for any rheumatoid aliment, including those occurring in the thigh. Then the belief he expresses saying "I have arthritis in my thigh" is correct, but it is not a belief involving our notion of arthritis. According to Burge, "The patient's mental contents differ while his entire physical and non-intentional mental histories, considered in isolation from their social context, remain the same" (79). Burge concludes that the contents of beliefs are not individuated internally. They are in part socially individuated. The contents of a person's beliefs depend in part on his social environment.

If the content of a belief is externally individuated, can a belief be a functionally defined entity? Burge argues that it cannot if a functional role is specified using physical and non-intentional terms; his thought experiments obtain variation in the contents of a person's beliefs without variation in the person's physical and non-intentionally characterized features (105–8). Burge's point is persuasive for narrow functionalism. That view defines belief in terms of an internal causal role. Internally defined, belief cannot have broad content. However, wide functionalism, which defines belief in terms of a larger causal role, is not as vulnerable to the objection. It permits a belief's causal role and its content to be in part externally individuated. In any case, a physicalistic account of belief can survive the abandonment of functionalism. A physicalistic account may explain a belief's physical realization rather than define it physically, and may freely appeal to all aspects of the believer's physical environment.

Baker (1995; Chp. 2) uses belief's external individuation to argue that beliefs are not (identical with, supervenient upon, or constituted by) brain states. In particular, she argues against Dretske's account of belief as mental representation. She says that since belief involves a relation to environment, it cannot be realized by a person's internal states. In the following passage she states her case in terms of supervenience: "Psychological properties may fail to supervene on neurological properties because individuation of psychological states is more sensitive to the subject's environment than is individuation of neurological states" (65).

One reply to the objection is Dretske's (1988, 36). On his view, external factors determine indication and hence content. But, Dretske says, just as the content of a book need not be in the book, the content of a belief in the head need not be in the head. The internal mental entity is just a vehicle for external nonmental content. External individuation of content does not make a component of belief external. Belief in the sense of a mental entity is different from belief in the sense of the content of the mental entity. Although belief involves a relation to external objects, taken as a mental entity it is wholly internal.

This reply accords with the way we speak about belief. We distinguish a belief's being individuated externally and its local causal powers' being in the head, just as we distinguish a photograph of Mt. Everest's being individuated externally, by its relation to Mt. Everest, and its local causal powers' being in the paper constituting it. Thus in common parlance we distinguish a person's *believing* that Mt. Everest is the highest mountain from the person's *belief* that Mt. Everest is the highest mountain. The former is a relation to an external entity forming part of the content of the belief. The latter is the internal state generated by the relation. The belief relation is not like the relation of being exactly a billion miles from the Voyager 1 spacecraft; it creates an internal state. We mean the internal state when we say that the belief is internal. We are interested in the internal state because it is the seat of the belief's causal powers described internally. The internal state is the cause of the internal events and states the belief causes, for example, brain events' and states' internal realization of inferences and intentions to act. A belief taken as a mental entity is internal and may be realized by a mental representation.

Given that content is externally individuated and broad, Fodor (1994) claims that we confront another issue. The first step in the issue's presentation states that the success of human inference is evidence that inference is computational. In other words, inference is syntactic, the product of structures mentally realized. The structures are defined by narrow causal role (not neurologically characterized). Fodor calls the syntactic vehicle of inference "the language of thought" or "mentalese." The next step concerns the nature of belief. According to Fodor (47), it is a relation between a believer, proposition (broad content), and mode of presentation (sentence of mentalese). Moreover, since inference is computational, a belief is realized as a computational state (a sentence of mentalese). Its internal realization is syntactic, or structural. The final step is the statement of the main issue. If a belief is realized as an internal computational state and has broad content, there must be some mechanism for keeping the computational state and the belief's content in alignment. What is it?

The aligning mechanism is not metaphysical, some type of identity, since two external states may correlate with one computational state, as in Twin Earth cases. Also, two computational states may correlate with one external state, as in Frege's puzzles about identities. Fodor holds that contingent laws of nature maintain a general alignment of broad contents and computational states. That alignment is the foundation for *ceterus paribus* psychological laws appealing to beliefs and other intentional states. "Computational-syntactic processes can implement broad-intentional ones because the world, and all other worlds that are nomologically nearby, arranges things so that *the syntactic structure of a mode of presentation reliably carries information about its causal history*" (54).

5. ACCEPTANCE

Some epistemologists introduce a propositional attitude to supplement or replace belief. They generally call it acceptance, but a few call it assent or belief in a technical sense. Their motivations are diverse, and their accounts of acceptance differ according to their motivations. Is acceptance in some form a welcome addition to epistemology? Should it replace belief in the definition of knowledge? This section addresses these questions.

One motivation for acceptance is the introduction of an attitude to which deductive logic applies. Since certainty of the premisses of a valid deductive argument warrants certainty of its conclusion, Levi takes acceptance as certainty (1997; Chp. 3). Instead of using the term 'acceptance,' however, he uses the term 'full belief.' Full belief contrasts with partial belief. Full belief entails assigning a probability of 1, whereas partial belief admits of lesser probability assignments. Levi further explicates full belief in terms of what he calls serious possibility, a variety of epistemic possibility. Where X is an agent, *t* is a time, and *h* is a hypothesis, he says, "For every X and *t*, X is committed at *t* to fully believe that *h* if and only if X is committed at *t* to judge ~*h* impossible" (46). This principle is stated in terms of commitment to full belief rather than full belief since humans may fail to fully believe as they should because of cognitive limits, mistakes, insufficient motivation, and the like. Full belief generates commitment, but commitments extend farther than full belief. By a commitment to full belief, Levi means an obligation to fully believe insofar as able when demand arises, and an obligation to use reasonable opportunities to improve capacities for meeting the demand (41). Deductive logic applies more strictly to commitments to full belief than to full beliefs. The commitments are subject to principles of consistency and deductive closure (41). Thus, although a rational person who fully believes *p* and fully believes *q* may fail to fully believe *p* & *q*, she is committed to fully believing the conjunction.

One problem with taking acceptance as certainty is that little is justifiably accepted in this sense. Not much is justifiably taken as certain or assigned probability 1. It is a rare proposition on which one justifiably bets the farm.

Another motivation for acceptance is the introduction of an attitude attuned solely to epistemic considerations. Lehrer has this motivation. He takes acceptance as belief arising from concern for the truth (1990, 10–11, 20–1, 113–14). "There is a special kind of acceptance requisite to knowledge. It is accepting something for the purpose of attaining truth and avoiding error with respect to the very thing one accepts. More precisely, the purpose is to accept that p if and only if *p*" (11). Acceptance so construed is a type of belief, but not belief out of habit, or instinct, or for pragmatic reasons such as loyalty, piety, or felicity. Acceptance is a special type of belief, epistemically prompted belief.

Lehrer adds, "To accept the information that *p* implies a readiness in the appropriate circumstances to think, infer, and act on the assumption that the information is correct." This gloss on acceptance invites criticisms analogous to Section 3's criticisms of traditional views of belief's role in inference and practical reasoning. In those areas, as belief should, acceptance in Lehrer's sense should cede to degree of belief.

Lehrer's introduction of acceptance is inspired by the observation that a belief is knowledge only if the belief is supported by epistemic reasons. The traditional definition of knowledge may accommodate this observation without replacing belief with acceptance, however. Rather than reformulate the definition in terms of acceptance, one may make more precise the type of justification required before a true belief counts as knowledge. The requisite justification may be limited to epistemic reasons. The case for replacing belief with the new attitude of acceptance is therefore not strong. We put aside this conception of acceptance.

In statistics, hypotheses are accepted or rejected according to the results of statistical tests. Some statisticians resist defining the technical sense of acceptance involved. In fact, some statisticians prefer to speak of not rejecting a hypothesis, rather than accepting it. Their worries concern the number and variety of tests necessary to warrant acceptance. However, accepting a hypothesis is generally taken to be acting as if it is true. Rejecting it is acting as if it is false. These forms of action may occur despite lack of certainty in either the hypothesis or its negation. The argument for them is that the statistical policies that prompt them have a high relative frequency of success in the long run. See, for example Neyman (1950, 259–60).

Jeffrey criticizes the view that it is the job of science to counsel accepting hypotheses as a basis of action (1992; Chp. 2). The warrant for acting as if a hypothesis is true depends on circumstances, including the possible gains and losses of so acting. As a basis of action, rational acceptance or rejection depends on the risks of action, and so goes outside the province of statistical testing. Howson and Urbach formulate similar criticisms (1989, 162–5). These criticisms make the statistical conception of acceptance unattractive for epistemology. Even if the statistical conception has epistemological relevance, it is unattractive for epistemic theorizing since it does not carve at the joints.

A distinction between belief and acceptance figures in debates about scientific realism, or realism with respect to the entities posited by scientific theories. Van Fraassen doubts the existence of those theoretical entities, but acknowledges the usefulness of our best scientific theories. He says the appropriate attitude toward those theories is acceptance rather than belief (1980, 4, 12–13). Belief in a theory entails belief in its entities, whereas acceptance of a theory does not. Acceptance carries belief that the theory is empirically adequate, roughly, belief that its observational consequences are correct. It also involves a commitment to the research program to which the theory belongs. This commitment is the pragmatic side of theory acceptance. The reasons for it may concern simplicity, informativeness, and explanatory power rather than truth. Van Fraassen designs acceptance of scientific theories as a way of acknowledging the theories' pragmatic value without commitment to their theoretical entities; it delivers us from metaphysics (69).

We put aside acceptance in van Fraassen's sense since we are concerned primarily with believing and accepting propositions that do not express scientific theories. For these propositions, both belief and acceptance in van Fraassen's sense involve belief that the proposition is true. Acceptance, however, may be supported by pragmatic reasons not supporting belief. Interestingly, reasons for acceptance in van Fraassen's sense include pragmatic reasons that Lehrer's conception of acceptance eschews. Some authors conceive of acceptance in contrary ways.

Another motivation for distinguishing belief and acceptance is that belief, being involuntary, is not an appropriate object of our highest standards of normative evaluation. Acceptance, in contrast, is voluntary, and so may be appraised by standards akin to moral standards for voluntary actions. Cohen (1992) has this motivation for distinguishing belief and acceptance. He takes a belief that *p* as a disposition to feel it true that *p* and false that not-*p* (4). The feeling is cognitive and in the same family as doubt and certainty, suspicion and surprise. It is typically activated by thinking of the proposition that p. In contrast, to accept that *p* is to have or adopt a policy of deeming, positing, or postulating that *p*, that is, to use *p* as a premiss in reasoning and deliberation (4). Acceptance is linguistic and available to adult humans but not infants or animals.

Cohen uses his distinction between belief and acceptance to resolve some puzzles. First, belief, being involuntary, is less structured than acceptance, even when both attitudes are rational. For instance, belief is not bound by subjective deductive closure, but acceptance is (Sec. 5). Since acceptance is more constrained than belief, the preface paradox is resolvable if restated in terms of acceptance. Rules of rational acceptance prohibit accepting each claim of one's book while accepting the denial of their conjunction (Sec. 6). Second, conflicting roles for belief in inquiry and deliberation lead some theorists to claim that the ordinary concept of belief is incoherent. The apparent incoherence can be resolved, however, by dividing the conflicting roles between belief and acceptance (Sec. 15). Third, explaining a group's actions raises problems that acceptance handles better than belief does, since acceptance is voluntary whereas belief is a matter of feeling (Sec. 10). And fourth, some psychological states seem to involve contrary attitudes. Distinguishing between belief and acceptance explains their structure. Self-deceit involves belief that *p* but acceptance that not-*p*. Similarly, weakness of will involves belief that one should not do *A* but acceptance of *A*'s permissibility (Chp. 5).

Cohen distinguishes two kinds of knowledge, one involving acceptance and the other belief (Chp. 4). Knowledge of scientific theories involves acceptance, he says, whereas observational knowledge involves belief (92–3). Does epistemology gain from the bifurcation? Reliance on the distinction between acceptance and belief thwarts the theoretical objective of a unified account of knowledge. To overcome the obstacle, epistemology needs an explanation of acceptance in terms of belief and desire. Acceptance, being voluntary, might be explained by the degrees of belief and degrees of desire that lead to the policy of using *p* as a premiss in decisions about what to do or think. This explanation of acceptance systematizes epistemology's treatment of belief, acceptance, and knowledge, but also makes the attitude of acceptance derivative rather than fundamental, and so less important to epistemology. The basic epistemic attitude is passive, being attuned to evidence and not also the desires that direct active attitudes.

Furthermore, as Cohen is well aware (26), evaluating attitudes that are nonvoluntary is possible. We commonly evaluate emotions, for example. Grief may be appropriate or excessive. Jealousy may be reasonable or not. Belief is under a person's indirect control, and his beliefs are a sign of his cognitive character. Evaluation of beliefs directs the formation of cognitive habits and character even if belief is not voluntary. Plantinga (1993b), for example, assesses nonvoluntarily beliefs according to whether the epistemic equipment producing them functions properly. Epistemology need not supplement belief with acceptance to be normative. It is true that the standards of evaluation for the voluntary differ from the standards for the nonvoluntary. But normative objectives do not motivate shifting epistemology from belief to acceptance, or enlarging it to cover acceptance along with belief except as a derivative concept.

None of the concepts of acceptance reviewed so far is promising for the traditional work of epistemology. Keeping epistemology in mind, how is acceptance best defined? Acceptance, unlike belief, is not part of the taxonomy or proprietary language of psychology. It is a technical term of revisionist epistemology. It needs to be defined in a way that gives acceptance a role in significant epistemic principles. Since acceptance, however defined, cannot steal degree of belief's central role in inference and practical reasoning, I think the most interesting view of acceptance ties it to assertion. Acceptance defined in terms of assertion may, perhaps, figure in important epistemic principles since assertion is the public side of belief.

Kaplan takes acceptance as assertability in the context of inquiry. Instead of the term 'acceptance' he uses 'belief' in a technical sense to stand for acceptance. His definition states, "You count as believing *P* just if, were your sole aim to assert the truth (as it pertains to *P*), and your only options were to assert that *P*, assert that ~*P* or make neither assertion, you would prefer to assert that *P*" (1996, 109). Kaplan claims that deductive closure and consistency are regulative ideals for acceptance (112–21). He resolves the preface paradox by recommending that an author accept the conjunction of her book's assertions despite confidence that the book contains an error. This resolution of the paradox condones acceptance of the improbable, which

Kaplan contends may be justified, despite the risk of error, by inquiry's goal of comprehensiveness (142–8). For example, he argues, comprehensiveness justifies acceptance of scientific theories although they are improbable if detailed (146–8, 181–2).

A crucial feature of this view is provision for motives for acceptance beside truth, which by itself might result in acceptance of little. Inquiry generates two goals (at least): accepting truths and avoiding errors, that is, comprehensiveness and accuracy. Maher advances a definition of acceptance similar to Kaplan's, but adds a resolution of the various conflicting goals directing acceptance (1993; Chps. 6–9). On his view, acceptance of a hypothesis *H* is the mental state expressed by sincere intentional assertion that *H* (130). Acceptance is rational when it maximizes expected cognitive utility. The cognitive utility of accepting a hypothesis depends in part on its closeness to the truth and its informativeness. The total depends on how well acceptance serves, on balance, the various theoretical goals of inquiry.

This account of rational acceptance raises a question about cognitive utility. Does cognitive utility behave as the account requires? Let us assume the absence of impediments to sincere intentional assertion. Then, if acceptance is the attitude expressed by such assertion, rational acceptance should observe the same principles as rational assertion does. Hence if acceptance is rational for propositions whose acceptance maximizes expected cognitive utility, then assertion should be rational for those propositions too. For this to be so, those propositions must comply with all the canons of rational assertion. These canons include the rules of deductive logic. But the preface and lottery paradoxes suggest that maximizing expected cognitive utility fails to conform to deductive logic. It seems possible that for each premiss but not the conclusion of a valid deductive argument acceptance may maximize expected cognitive utility. Also, it seems possible that for each proposition in an inconsistent set acceptance may maximize expected cognitive utility. Furthermore, the canons of rational assertion prohibit asserting the improbable. But, according to Maher, acceptance of the improbable sometimes maximizes expected cognitive utility (146). It therefore seems that rational assertion does not follow expected cognitive utility. The apparent divergence raises doubts about cognitive utility's ability to fill its role in Maher's account of acceptance, and thus raises doubts about that account.

Let us table issues concerning cognitive utility, issues Kaplan avoids by dispensing with cognitive utility, and move to a general evaluation of acceptance defined in terms of assertion. Is the concept of acceptance important in epistemology? One positive argument claims that acceptance is needed as a replacement for the ordinary concept of belief since that concept is incoherent. Section 3 considers this charge against belief and dismisses it. Should epistemology replace belief with acceptance despite belief's coherence? Belief is an attitude with wider range than acceptance. Animals believe but do not accept propositions. Hence acceptance does not connect epistemology with research on the cognitive states of animals. Moreover, acceptance defined in terms of assertion lacks the accessibility of belief. To determine whether one accepts a proposition, one must entertain the proposition's assertion. According to Kaplan's definition, one must determine one's preferences concerning the proposition's assertion in a hypothetical situation. According to Maher's definition, one must determine whether one has toward the proposition the complex attitude, combining degree of belief and degree of desire, that would be expressed by the proposition's sincere intentional assertion. Being less accessible than belief makes acceptance less suitable than belief as an epistemic foundation.

Maher's account of acceptance, our most detailed account, does not motivate replacing belief with acceptance. Does it motivate treating acceptance in addition to belief? If acceptance does not replace belief, it occupies only a secondary role in cognition. Degree of belief holds the primary role. Degrees of belief regulate acceptance once cognitive utilities are assigned. Having degrees of belief, epistemology may do without acceptance.

Maher argues that a faithful account of science must attend to the attitude of acceptance (Chp. 7). First, scientists provide an account of hypotheses accepted, not probabilities assigned to them. A theory of acceptance is needed to explain this practice of science. Second, a theory of acceptance explains why hypotheses continue to be accepted until rivals are formulated. Accepting the received view maximizes cognitive utility until a rival presents a new option with higher cognitive utility. Third, a theory of acceptance explains why gathering data has scientific value. Comparing the expected cognitive utility of the set of hypotheses rationally accepted before data acquisition with that quantity's expected value after data acquisition, we see that the latter is at least as great as the former. Gathering data is cognitively promising from the standpoint of acceptance.

These three reasons for a theory of acceptance are not decisive. The three explanatory objectives are met by a theory of assertion of hypotheses. Probabilities and cognitive utilities govern assertion. Our cognitive goal of making comprehensive, accurate assertions directly explains scientific theorizing without appeal to the intermediary attitude of acceptance. Acceptance is not needed for the move from probabilities and cognitive utilities to assertion.

Epistemology is not obliged to treat acceptance to address traditional topics. It need not replace belief with acceptance because belief is incoherent, or because acceptance works better than belief in the traditional definition of knowledge. Should epistemology nonetheless treat acceptance? That depends on whether acceptance can earn its keep despite the work accomplished by probability and cognitive utility, and belief and assertion. No concept of acceptance we have reviewed shoulders a burden they do not already carry.

6. CONSEQUENCES FOR EPISTEMOLOGY

What are the implications for epistemology of our investigation of belief and acceptance? A conclusion from philosophy of language is that a belief's justification concerns not just the proposition believed but also the believer's understanding of that proposition. The belief that Hesperus is Hesperus may be justified while the belief that Hesperus is Phosphorus is not justified. Both beliefs have the same content, but the latter belief involves the believer's notion of Phosphorus as well as his notion of Hesperus. Justification of belief must go beyond the proposition believed to the notions involved in belief of that proposition.

A conclusion from philosophy of mind is that the content, and hence truth value, of a belief is a matter of concepts individuated externally, not internally. Suppose a person thinks that arthritis is deep, persistent pain wherever located, and he has such pain in his thigh. Hence he believes that he has arthritis in his thigh. His belief is not true. Its content does not involve his conception of arthritis. It involves the concept expressed by the word 'arthritis' in his linguistic community. Since his belief is false, it is not knowledge even if justified. Going further, Davidson (1991, 199) uses the external individuation of perceptual concepts, and some of the epistemic consequences Burge (1986) draws, to argue against global skepticism. He says, "If anything is systematically causing certain experiences (or verbal responses), that is what the thoughts and utterances are about. This rules out systematic error."

The section on acceptance concludes that epistemology need not introduce acceptance. Belief is sufficiently robust for its role in epistemology. Along with degree of belief, it generates acceptance's product. It yields a satisfactory account of assertion in the context of inquiry. Although acceptance may find roots in the conventions of inquiry, its fruit already grows on branches of the theory of belief.

The section on logic draws the most significant conclusion of our study. Belief is not the basic epistemic attitude. Rather, degree of belief is. Belief has a role in inquiry, but it is regulated by degree of belief. Hence probabilistic epistemology should supplement traditional epistemology.

Some epistemologists are probabilists. Their accounts of degree of belief vary. Jeffrey (1992) follows the tradition of Ramsey, taking degree of belief as a basis of action (30–4). It is revealed by a person's betting behavior. A person's degree of belief that *p* is roughly the largest percentage of a dollar that the person would pay for a bet that yields $1.0 if *p* is true and otherwise $0.0. This characterization of degree of belief can be made more precise by using the desirabilities of monetary outcomes rather than the outcomes themselves. Both desirabilities and degrees of belief may be elicited from betting behavior over a sufficiently rich assortment of bets given the bettor's rationality. Jeffrey's favored methods of elicitation, however, do not completely specify degrees of belief, except in special cases (33).

The elicitation of degrees of belief presumes idealizations about agents so that their rationality makes their degrees of belief conform with the standard laws of probability, and thereby in a common terminology makes them *coherent*. Under the idealizations, does rationality impose any structure beyond those standard laws? Probabilists disagree. Some advance a principle of indifference that allots equal degrees of belief to possible cases divided certain ways. Others advance principles making degrees of belief match relative frequencies in certain cases. Jeffrey (59–64) holds a principle of this sort first formulated by de Finetti.

One point of general agreement is that rational degree of belief depends on evidence. A common principle for updating degrees of belief when new evidence arrives is called "conditionalization." It says that degrees of belief after acquiring new evidence should equal prior degrees of belief conditional on that evidence. Jeffrey generalizes this principle for cases in which no proposition satisfactorily expresses all the new evidence acquired. An observation by candlelight, for instance, may alter degrees of belief about colors of objects without providing certainty of any proposition specifying the relevant content of the observation. The observation's relevant content may be nonpropositional. Since Jeffrey's probabilism allows for nonpropositional evidence, it allows for probabilities that do not rest on propositions that are certain. This makes it *radical* probabilism. Since probabilities may rest on mere probabilities, epistemology may be entirely probabilistic (1–13).

Probabilism as outlined leaves many questions open. In particular, how does the mind assign a degree of belief to a proposition? Pollock (1986, 100–2, 108–9) doubts probabilism because he doubts the existence of degrees of belief. A quantitative mental state seems unrealistic. Making the assignment an interval of degrees of belief does not help since the interval's endpoints also suggest quantitative mental states. Harman adds that there is not enough room in the head to store all the degrees of belief needed for conditionalization as new evidence arrives (1986, 25–7). Jeffrey's line of reply is that degrees of belief merely represent mental states and need not be in the head themselves (29). The quantitative representation of mental states serves a useful epistemic function even if the mental states are not themselves quantitative. Jeffrey takes a person's epistemic state to be represented by a set of probability functions, or, alternatively, conditions on probability functions (68–73). Coherence for an incomplete probability function is coherence for at least one of its completions (85). Probabilities provide a framework for an epistemological account of the bearing of evidence on an inference's warrant and an action's choiceworthiness, even if mental states have a nonquantitative psychological structure.

Acknowledgements. I thank Prasanta Bandyopadhyay, Peter Markie, Andrew Melnyk, and Wlodek Rabinowicz for illuminating discussions of belief and acceptance. I am also grateful to the editors for helpful suggestions.

Paul Weirich
University of Missouri-Colombia

REFERENCES

Baker, L. R.: 1995, *Explaining Attitudes,* Cambridge University Press, New York.

Burge, T.: 1979, 'Individualism and the Mental,' in P. French, T. Uehling, and Howard Wettstein (eds.), *Studies in Metaphysics, Midwest Studies in Philosophy,* Vol. 4, pp. 73–121, University of Minnesota Press, Minneapolis.

Burge, T.: 1986, 'Cartesian Error and the Objectivity of Perception,' in P. Pettit and J. McDowell (eds.), *Subject, Thought, and Context,* pp. 117–36, Clarendon Press, Oxford.

Cohen, L. J.: 1992, *Belief and Acceptance,* Clarendon Press, Oxford.

Crimmins, M.: 1992, *Talk about Beliefs,* The MIT Press, Cambridge, MA.

Davidson, D.: 1991, 'Epistemology Externalized,' *Dialectica* **45**, 191–202.

Dretske, F.: 1988, *Explaining Behavior,* The MIT Press, Cambridge, MA.

Fodor, J.: 1990, *A Theory of Content and Other Essays,* The MIT Press, Cambridge, MA.

Fodor, J.: 1994, *The Elm and the Expert,* The MIT Press, Cambridge, MA.

Foley, R.: 1993, *Working without a Net,* Oxford University Press, New York.

Gärdenfors, P.: 1988, *Knowledge in Flux,* The MIT Press, Cambridge, MA.

Harman, G.: 1986, *Change in View,* The MIT Press, Cambridge, MA.

Howson, C. and P. Urbach: 1989, *Scientific Reasoning: The Bayesian Approach,* Open Court, LaSalle, IL.

Jeffrey, R.: 1992, *Probability and the Art of Judgment,* Cambridge University Press, New York.

Kaplan, M.: 1996, *Decision Theory as Philosophy,* Cambridge University Press, New York.

Kyburg, H.: 1997, 'The Rule of Adjunction and Reasonable Inference,' *Journal of Philosophy* **94**, 109–25.

Lehrer, K.: 1990, *Theory of Knowledge,* Westview Press, Boulder, CO.

Levi, I.: 1997, *The Covenant of Reason,* Cambridge University Press, New York.

Maher, P.: 1993, *Betting on Theories,* Cambridge University Press, New York.

Millikan, R.: 1984, *Language, Thought, and Other Biological Categories,* The MIT Press, Cambridge, MA.

Neyman, J.: 1950, *First Course in Probability and Statistics,* Holt, New York.

Plantinga, A.: 1993a, *Warrant: The Current Debate,* Oxford University Press, New York.

Plantinga, A.: 1993b, *Warrant and Proper Function,* Oxford University Press, New York.

Pollock, J.: 1986, *Contemporary Theories of Knowledge,* Rowman & Littlefield, Totowa, NJ.

Putnam, H.: 1975, 'The Meaning of "Meaning,"', in *Mind, Language and Reality, Philosophical Papers,* Vol. 2, pp. 215–71, Cambridge University Press, New York.

Richard, M.: 1990, *Propositional Attitudes,* Cambridge University Press, New York.

Ryle, G.: 1949, *The Concept of Mind,* Barnes and Noble, New York.

Stich, S.: 1983, *From Folk Psychology to Cognitive Science: The Case against Belief,* The MIT Press, Cambridge, MA.

van Fraassen, B.: 1980, *The Scientific Image,* Clarendon Press, Oxford.

ILKKA NIINILUOTO

INDUCTION

Induction is a mode of inference which has important links with many epistemological problems. It is a common feature of the different varieties of induction that they are not necessarily truth-preserving. Thus, induction is weaker than logical deduction or entailment. However, unlike deduction, inductive inference is ampliative in the sense that at least part of the content of its conclusion is not explicitly or implicitly present in the premises. Hence, if there is a rational answer to Hume's problem concerning the justification of induction, inductive inferences can be claimed to be knowledge-increasing, i.e., they allow us to expand the domain of our rationally warranted beliefs. As responses to these challenges, philosophers have given probabilistic reconstructions of different types of induction and analysed their role in the methodology of the empirical sciences.

1. THE VARIETIES OF INDUCTION

1.1. Deduction and Induction

Deductive inference is characterized by the condition that the conclusion is a *logical consequence* of the premises: whenever the premises are true, the conclusion must be true as well. This idea of necessarily truth-preserving arguments can be explicated in systems of formal logic. They also link the concepts of deduction and logical truth: Q is a logical consequence of P if and only if the statement P → Q is logically true, where → is the connective of material implication.

According to Ludwig Wittgenstein's *Tractatus Logico-Philosophicus* (1922), logical truths are "tautologies" (e.g., 'Now it is raining or it is not raining'). In Leibniz's words, they are true in all possible worlds. Using the terms of the theory of semantic information by Rudolf Carnap and Yehoshua Bar-Hillel, the information content of logical truths is empty, since they allow all alternative states of the world (see Bar-Hillel 1964). The corresponding characterization of logically valid deduction or entailment within first-order logic was given by Alfred Tarski in 1935: all models of the premises are also models of the conclusion (see Tarski 1956). Hence, conversely, the semantic information content of the conclusion is included in the content of the premises.

Many philosophers have concluded that the tautologous character of deduction means that logical inferences are "uninformative": the feeling that chains of deductive inferences bring about new information about the world is only a psychological illusion. This view has been challenged by Jaakko Hintikka (1973).

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 521-545.
© 2004 *Kluwer Academic Publishers.*

He points out that, due to the undecidability of full first-order logic with relations, it is not always possible check effectively which alternatives allowed by a statement are logically consistent and which are not. Elimination of such inconsistent pseudo-alternatives is an objective or non-psychological feature of "non-trivial" deductive inferences, and allows us to speak of new information gained by deduction. To distinguish this idea from Carnap's and Bar-Hillel's "depth information", which is never increased by deduction, Hintikka calls his new concept "surface information" (cf. Hintikka and Suppes 1970).

Inferences which are non-deductive allow cases where the premises are true but the conclusion is false. Many examples of such inferences have been studied as "logical fallacies". For example, while the inference from P → Q and ~Q to ~P is logically valid (modus tollens), the attempt to derive P from the premises P → Q and Q commits "the fallacy of affirming the consequent".

Induction is often characterized negatively as inference which is non-deductive. Such inferences are fallacious from the viewpoint of deductive logic, but they are *ampliative* or *content-increasing* in the sense that the conclusion contains some depth-information not present in the premises. The challenge for a theory of induction is to show that such inferences may be reasonable in some sense. For example, in spite of not being necessarily truth-preserving, induction may be "probable" reasoning (see Section 3).

As the falsity of the conclusion Q is compatible with an inductive inference from the premise P, inductive reasoning is typically *non-monotonic* or *defeasible* in the sense that adding a premise S to P may preclude the inference to Q (cf. Gabbay and Smets 1998). In contrast, deduction is monotonic, since the deducibility of Q from premise P guarantees the deducibility of Q from P&S for any statement S.

1.2. Main Types of Induction

Suppose that Q is known to be a logical consequence of premises $P_1, ..., P_n$ which are known to be true. Then the logical derivation of Q from $P_1, ..., P_n$ constitutes a *proof* or *demonstration* of Q. This idea is prominent in the "metamathematical" study of formal systems, where the acceptable premises are axioms of logic or some mathematical theory.

The same axiomatic ideal for all science was formulated already by Aristotle who presented the first formal system of deductive logic in his theory of syllogistic. Aristotle required that the theorems of special sciences have to be demonstrated by deductive syllogisms which start from axioms or "the first premises". Aristotle thought that such axioms have to be necessarily true, but he realised that – due to the problem of circularity – they cannot be established in the same demonstrative way. The process of reaching these general axioms was called *epagoge* by Aristotle. This term was translated as *inductio* by the Latin commentators.

The standard interpretation has assumed that Aristotle had two conceptions of induction (see Ross 1949). First, in *intuitive induction* a universal generalization is grasped by the perception of some particular instances of the generalization. This is a psychological process which cannot be formulated as an argument involving propositions. Secondly, in *complete induction* (or "perfect induction") a generalization is obtained by enumerating all of its instances. For example, to prove that all animals are mortal, give an exhaustive enumeration of all species of animals and show that each species is mortal.

The latter idea is preserved in the term *mathematical induction*, which refers to a demonstrative method of proving arithmetical generalizations: to prove that all natural numbers n have the property $\phi(n)$, show that (i) the number 0 has the property $\phi(0)$, and (ii) if $\phi(k)$ then $\phi(k+1)$ for an arbitrary number k. The two conditions (i) and (ii) guarantee that the proof goes through all the members of the series 0, 1, 2, ..., n, n+1,... so that $\phi(n)$ holds for all natural numbers n. Within systems of formal arithmetic, where the Principle of Induction is assumed as an axiom, inferences of this kind are deductively valid.

It has been argued that these interpretations misrepresent Aristotle's view (cf. Niiniluoto 1995). First, his account of induction is inseparable from concept formation. In this sense, his view resembles William Whewell's (1840) doctrine where induction always involves the discovery of a new conception that the investigator "superinduces" over the known particular cases (cf. Butts 1968). Secondly, as Whewell (1860) himself argued, Aristotle cannot assume an exhaustive enumeration of all cases as the basis of induction, but only that the known particular instances constitute a representative class of the relevant cases.

In contrast to complete induction, *incomplete* induction is a genuinely ampliative form of inference from a proper part to a whole. Such inferences are called *enumerative induction.* Thus, *inductive generalization* is taken to proceed from a finite sample to a population, where the population may be indefinitely large or infinite. For example, all of the ravens observed until now have been black, hence all ravens are black. *Statistical generalization* goes from a sample to a statistical statement about a population. For example, according to Hans Reichenbach's (1949) Straight Rule of Induction, if 10 per cent in a random sample of the citizens of Finland are left-handed, then 10 per cent of the Finnish citizens are concluded to be left-handed. *Singular* inductive inference – or *eduction*, as John Stuart Mill (1843) called it – proceeds from a sample to a new individual from the population. For example, all of the swans observed so far have been white, hence also the next swan to be examined will be white.

The traditional examples of Aristotelian syllogisms proceed from a population to its parts. For example, the following argument is logically valid: all humans are mortal, Socrates is mortal, hence, Socrates is mortal. For this reason, some philosophers have called "deductions" all inferences from a population to its parts, or from a generalization to particular cases. An inference from a statistical statement about a population to an individual or a sample was called *statistical deduction* by Charles S. Peirce, even though Peirce knew well that such inferences are not logically valid. For example, 90 per cent of the Parisians are catholic, hence probably this randomly selected Parisian is a catholic. In the context of statistical prediction and explanation, this mode of argument is also called *direct inference* (see Carnap 1962; Levi 1967).

Another non-deductive argument, which often is treated as a species of induction, is *analogy*: from observed similarities between two objects it is inferred that they share some further property (see Hesse 1974). As Mill suggested, such inferences may be understood as enumerative inductions over the properties of objects. For example, argument by analogy is used when the results of medical experiment with animals are entended to cover human beings.

Francis Bacon in the early 17th century thought that induction by simple enumeration is "childish" (see Bacon 1960). He argued that induction should involve a systematic tabular method of excluding putative but false connections between the examined variables, so that the finally remaining only alternative is established with certainty. This idea has been called *eliminative induction* or *demonstrative induction*. It was further developed in John Stuart Mill's (1843) Rules of Experimentation (see von Wright 1951; Blake, Ducasse, and Madden 1960).

Finally, it is often said that true observable consequences of a hypothetical theory give *inductive support* to the theory (at least if no counter-examples refuting the theory have been found). In this case, induction proceeds in a direction which is converse to deduction. Inferences of this kind are typical when a proposed theory is indirectly tested by its consequences. When the theory is discovered to give an explanation of some known facts, this kind of inference was called *hypothesis* or *abduction* by Peirce. Initially Peirce applied this term to inferences that are converse to explanations in the sense that they proceed from effects to causes (see Niiniluoto 1999b).

2. THE JUSTIFICATION OF INDUCTION

2.1. Hume's problem

The rules of deduction are not empirical psychological laws concerning the "laws of thought", but rather serve as criteria of logically valid argumentation. Even though there are systems of "alternative logic", the debates about the proper rules of deduction arise only in some special contexts, such as reasoning within constructive mathematics. Thus, the basic justification of deduction can be simply expressed by its necessarily truth-preserving nature.

A similar justification of ampliative inductive inferences is not possible. For example, enumerative induction is fallible, since it is possible that the conclusion is false even when the premises are true. The classical example illustrating this was, after a long series of observations of white swans, the discovery of black swans in Australia.

David Hume, in his *An Enquiry Concerning Human Understanding* (1748), gave a powerful expression of doubts concerning the justification of induction (see Hume 1902). Hume distinguished two kinds of propositions. Those expressing relations of ideas, including the truths of arithmetic and geometry, can be known with demonstrative certainty by mere operation of thought. But if a proposition expresses a matter of fact, its negation is possible, and it can be known only by sense experience. As an empiricist, Hume assumed that a factual statement can be certain only if its truth is immediately based upon the present testimony of our senses or memory. All the other factual beliefs, including all of our expectations about the future, have to inferred from the evidence concerning present facts. Such inferences have to rely on the relation of cause and effect. But as this relation is a factual one, not logical or a priori, it has to be discovered not by reason but by experience. However, experience at best shows that there is a regular succession of some events A and B, and thereby we learn to associate the ideas of A and B in our mind. But, according to Hume, we have no sense impression of the necessary connection between a cause A and an effect B. To justify the existence of this causal relation, we have to assume that the future will be conformable to the past, i.e., that inductive generalization is a valid form of inference.

Hume's problem can now be formulated. Human knowledge, both in everyday life and in science, seems to go beyond the limits of immediate sense experience. All inferential beliefs have to be in some way based upon causal relations, but knowledge of such relations is derived from experience by using induction. But induction is reliable only if the world is uniform, i.e., the future resembles the past. This principle of the uniformity of nature is itself a general statement which can be justified only by induction. Therefore, the attempt to justify induction seems to be viciously circular.

2.2. Replies to Hume's Problem

Hume himself concluded that there are no necessary connections between causes and effects in nature: induction is only a habit of our mind to expect regular successions between ideas. But if such a habit has no justification, how can this position avoid scepticism? Hume's challenge has been faced in five basically different ways (see von Wright 1957; Foster and Martin 1966; Swinburne 1974).

First, Immanuel Kant claimed that the Universal Law of Causality, asserting that every event has a cause from which it necessarily follows, can be known *a priori*, since causality is one of the categories imposed by our mind on the phenomenal world. However, Kant's argument can be claimed to be again circular, as he derives his categories from the possibility of human experience, and this seems to presuppose that this possibility remains in the future. It is also unclear how the general Law of Causality could suffice to support particular inductive inferences. Moreover, the later development of natural science seems to indicate that the world is not deterministic, so that the Universal Law of Causality is false. Later attempts to find general principles which would reduce induction to deduction or serve as presuppositions making induction rational have not been successful: there are not sufficient a priori reasons for believing them, and the attempt to derive them from experience stumbles again on the circularity problem.

Secondly, Max Black (1954) accepts that attempts at justifying induction are circular, but still argues that the justification of some specific inductive methods can be increased by *self-supporting inductive arguments*. However, it has been pointed out that it is possible to give such self-supporting argument also to counter-inductive patterns of inference (Salmon 1966). Hence, the fact that the use of induction has been successful in science does not prove that it will be successful in the future (Barker 1957). However, weaker formulations of this success argument may be more plausible: theories and methods are both improving within scientific progress, since by using reliable methods we have gained more truthlike theories, and truthlike theories help us to design more reliable methods of inquiry (cf. Boyd 1984; Niiniluoto 1999a).

Thirdly, Karl Popper's (1959, 1963) *falsificationism* accepts Hume's message: induction is impossible. General statements or laws of nature cannot be proved to true or even probable, but they can be shown to be false by counterexamples. Thus, science does not employ, nor does it need, induction at all.

The fourth approach is to accept induction as part of human practice, but to reject the demand concerning its justification. For example, Peter Strawson (1952) claims that the justification of induction is a *pseudo-problem*, since "rational belief" means the same as "a belief with strong inductive support". Also G.H. von Wright (1957) argues that the impossibility of justifying induction is "a disguised tautology", since induction by definition is ampliative and non-deductive inference, but he suggests that it is important to consider the constructive problem of clarifying the nature of inductive inference.

The fifth approach is to accept the fallibility of induction, and to analyse the relation between its presises and conclusion in terms of *probability*. An attempt to show that induction leads to certainty is based upon "deductivist" and "foundationalist" prejudices: inductive inferences are uncertain, and their rationality has to be defended by accepting the fact that all of our empirical knowledge is at best probable. This kind of epistemological position was called *fallibilism* by Peirce. It can be combined with the classical definition of knowledge: X knows that p if and only if (i) X believes that p, (ii) p is true, and (iii) p is justified, where the third requirement is understood as not demanding complete justification or conclusive reasons but only some probabilistic condition.

Peirce argued that probabilistic induction can be justified *in the long run* as an inference that is guaranteed to approach to the truth when the investigation is indefinitely continued. This programme has led to Hans Reichenbach's (1938) pragmatic "vindication" of the Straight Rule of statistical generalization: if any method is successful for the task of predicting relative frequencies, then his rule is guaranteed to be successful as well. Similar ideas in a non-probabilistic framework have been studied in formal learning theory (see Earman 1992; Kelly 1996).

Other probabilistic explications of induction allow precise reformulations of the problem of induction in different kinds of situations. Rudolf Carnap's (1952, 1962) programme of inductive logic was initially based upon the hope that metalinguistic statement of probability are analytically true, but the further development of inductive logic rather suggests that the legitimacy of inductive inferences may have "local" or contextual presuppositions. Developed in this direction, probabilistic accounts of induction serve as a tool of coherentist, rather than foundationalist, theories of justification and knowledge.

3. INDUCTION AND PROBABILITY

Aristotle defined probability as that which "usually occurs". The medieval philosophers associated probability with opinions which are supported by many authorities (see Byrne 1968). These two aspects, frequencies of occurrence and grades of belief, have been the central ingredients of the concept of probability since the emergence of a mathematical calculus of chances in the seventeenth century (see Hacking 1975).

3.1. Interpretations of Probability

The classical definition of probability, developed by Blaise Pascal in the 1660s, assumes a framework of equally possible basic outcomes of an experiment (e.g., the six faces of a rolling dice), and defines the probability of an event as the number of favourable cases divided by the number of all cases (see Laplace 1951). Instead of events, this theory may be formulated for propositions which express the occurrence of events or other facts. The *probability* P(H) of a proposition H is a number between 0 and 1, where P(H) = 0 for an impossible proposition H and P(H) = 1 for a sure or necessary proposition H. Let H&G be the conjunction of H and G, and H$\vee$G their disjunction. Then, according to the Principle of Additivity, if H and G are mutually exclusive propositions, the probability of their disjunction P(H$\vee$G) equals the sum of P(H) and P(G). Further, the probability of ~H, the negation of H, is 1 – P(H). The *conditional probability* P(H/G) of H given G is defined as the ratio P(H&G)/P(G). Propositions H and G are probabilistically independent if the probability P(H&G) equals the product of P(H) and P(G), i.e., P(H/G) = P(H).

The applicability of the classical definition is severely restricted by the assumption that the basic cases have to be symmetric or "equally possible". Still, the classical theory of chance correctly identified three simple but most basic mathematical principles of probability:

(1) $P(H) \geq 0$ for all H
(2) $P(H) = 1$ if H is logically true
(3) $P(H \vee G) = P(H) + P(G)$ if H&G is logically false.

Further, it is assumed that probability behaves well with respect to logical entailment:

(4) $P(H) = P(G)$ if H is logically equivalent to G.

It follows from (2) and (4) that P(H/E) = 1 if E logically entails H.

The main addition in A. N. Kolmogorov's axiomatization in 1933 is the generalization of the additivity requirement (3) to any countable number of pairwise disjoint disjuncts (cf. von Plato 1994). This allows a precise proof of the *Law of Large Numbers*. Let $rf_n(A)$ be the relative frequency of event A in a series of n independent repetitions of an experiment. Then the Weak Law of Large Numbers (known already to Jacob Bernoulli in 1713) states that, for any $\epsilon > 0$, the probability

$$P\{|rf_n(A) - P(A)| > \epsilon\}$$

approaches the limit 0 when n grows to infinity. According to the Strong Law of Large Numbers (known to E. Borel in 1909), it is true with probability 1 (or "almost surely") that the value of $rf_n(A)$ approaches in the limit the value P(A) when n grows to infinity.

The classical symmetry assumption is not true e.g. for a loaded dice. The *frequency interpretation* proposes that probability is identified with the stable relative frequency of a chance event: for example, the number of tails in a sufficiently long series of tosses is close to ½. It would be arbitrary to identify probability as relative frequency within some finite series, but it could be defined as the limit of such relative frequencies when the series is repeated ad infinitum:

$$(5) \qquad P(A) =_{df} \lim_{n \to \infty} rf_n(A).$$

In this sense, probability is an idealization of observable long-run frequencies. An alternative hypothetical formulation says that the probability P(A/B) of an event A is the limit toward which its relative frequency would converge in an infinite series B. Some philosophers prefer to say that P(A/B) is the probability of attribute A in reference class B (cf. Salmon 1966). This kind of treatment of frequentist probabilities is also called the *ensemble* interpretation (cf. von Plato, 1994).

Proposed by R.L. Ellis in 1843, the first serious attempt to formulate the frequency interpretation was made by John Venn in 1866 in *The Logic of Chance.* Later attempts to make the frequency interpretation precise include the works of Richard von Mises (1951) and Hans Reichenbach (1949). The main technical difficulty is to characterize in a consistent way the "random sequences" or "collectives" relative to which the limits of relative frequencies should remain stable. For example, in a periodic sequence 10101010... the relative frequency of result '1' approaches ½, but it has an easily defined subsequence (every second term) in which this limit has the value 0. Alonzo Church proposed in 1940 that limits of relative frequencies should be stable in all subsequences that are defined by recursive place selection functions. In the 1960s, Kolmogorov developed a new approach to this problem in his theory of complexity: a sequence is random if a universal Turing machine needs an input of approximately the same length to calculate the sequence as an output (see Fine, 1973).

One forceful criticism of the frequentist definition notes that, according to the Strong Law of Large Numbers, the equality (5) of probability and the limit of relative frequency holds only "almost surely" or with probability 1, and it should not be made an analytic truth by stipulation (see Stegmüller, 1973).

Another criticism is that this interpretation applies probability only to repeatable event types, so that it does not make sense to speak of the probability of unique or singular events (e.g., the probability of rain in Hamburg on January 1, 2000) or of the probability of hypotheses (e.g., Einstein's theory of relativity). An attempt to handle these problems by the concept of "weight" was made by Reichenbach (1938) in his probability logic: treat the singular statement as a prediction and use as its "weight" the relative frequency with which this "posit" is true. However, the choice of the relevant reference class in which the single case is placed has remained a matter of controversy (cf. Salmon 1966). For example, if we wish to determine the probability that a certain person will live to the age of 70, the relative frequencies will be different depending on the choice of the reference class (e.g., the class of men, Finns, persons with a coronary disease, car-drivers, mountain-climbers, etc.). Salmon (1984) recommends the choice of the largest "objectively homogeneous reference class" which cannot be divided any more in statistically relevant ways (cf. Fetzer 1988).

Essentially the same problem arises, if we try to make inferences about an individual on the basis of statistical premises:

(6) 100p% of the Fs are G
This b is an F
Hence, probably, this b is an G.

Already Peirce noted that this kind of probabilistic inference can be assigned a truth-frequency of size p only if b is a "random" member of the class of Fs. But when this "direct inference" is used for the purposes of prediction or explanation (cf. Hempel 1965), we face the ambiguity that different classes F dive different probabilities to the conclusion Gb (see the discussion by Henry E. Kyburg and Isaac Levi in Bogdan 1982).

Another physical interpretation of probability starts from Leibniz's suggestion that probability should be understood as "degree of possibility". The idea that there are real possibilities in nature, independently of epistemic uncertainty, was discussed by A.A. Cournot and C.S. Peirce in the 19th century. Following the principles of his indeterministic "tychism", Peirce proposed in 1910 that probability should be understood as a dispositional "habit" or "would-be". This interpretation of physical probability as *propensity* was reintroduced by Karl Popper in his discussion of quantum mechanics (see Popper 1957).

According to the long-run propensity interpretation, probability is the disposition of a chance set-up to produce series of events with characteristic relative frequencies (cf. Hacking 1965). This formulation does not yet solve the problem of unique events. The single-case propensity interpretation defines probability as the dispositional strength of a chance set-up to produce an outcome of a certain kind on a single trial of that set-up (see Fetzer, 1981, 1988; Suppes, 1984). Such propensities between 0 and 1 are thus "degrees of possibility" for events that are not completely determined by objective antecedent or causal conditions. Single-case propensity statements are theoretical claims that become testable by observable relative frequencies if there is a sufficient number of similar set-ups (e.g., atoms of the same radioactive substance).

Epistemic or doxastic interpretations take probability to be always relative to our knowledge. Laplace, who supported determinism, asserted that probability is an expression of our ignorance of the real causes of events. According to his Principle of Indifference, two events should be treated as "equally possible" if we do not know of any reason to prefer one to another (see Laplace 1951). Later Bayesians, like Stanley Jevons in 1873, defined probabilities as rational degrees of belief (see Kyburg and Smokler 1964).

According to the *subjective* or *personal* interpretation, the probability P(H/E) of a hypothesis H given available evidence E is the *degree of belief* in the truth of H warranted by E. The tool for studying such probabilities is Bayes's Theorem which states that the posterior probability P(H/E) is proportional to the product of the prior probability P(H) of H and the likelihood P(E/H) of H relative to E:

$$P(H/E) = P(H)P(E/H)/P(E). \tag{7}$$

Psychological studies show that the actual intensities of beliefs of human agents do not always behave in the manner of mathematical probability. However, as Frank Ramsey and Bruno de Finetti showed in the 1920s, assuming some rationality conditions among the agent's comparative judgments and preferences, it can be proved that rational degrees of belief can be represented by numerical values that satisfy the axioms of probability (see Ramsey 1950; de Finetti 1972). De Finetti considered conditions which concern relations of comparative probability (the agent regards proposition H more probable than proposition G); later representation theorems of measurement theory have shown under what rationality conditions such relations can be consistently expressed by quantitative probability measures. Ramsey considered preference relations between lotteries which yield outcomes with specific physical probabilities; in this case, the Representation Theorem should guarantee the existence of well-behaved subjective probability measures and utility functions. Ramsey's results were later generalized in the Bayesian decision theory which accepts the Principle of *Subjective Expected Utility* as its main decision rule (see Savage 1954; Gärdenfors and Sahlin 1988). More precisely, if $a_1, ..., a_n$ are alternative acts, $s_1, ..., s_k$ are alternative states of nature, a_i leads to outcome o_{ij} when the state of nature is s_j, and $P(s_j)$ is the subjective probability of s_j and $u(o_{ij})$ is the utility of outcome o_{ij}, then the rational Bayesian agent should choose that action a_i which maximizes the expected utility

$$\sum_{i=1}^{k} P(s_j)u(o_{ij}). \tag{8}$$

De Finetti's theorems also characterize rational degrees of belief as coherent *betting quotients*: it can be shown that my betting quotients satisfy the axioms of probability if and only if they do not allow a Dutch Book against me (i.e., a system of bets where I necessarily loose). For example, if I bet on both H and ~H with the ratio 3/4 and the stake is 100 dollars, then I must give 75 + 75 = 150 dollars for the bet, but I win only one of the bets and thus gain 100 dollars back, thereby losing 50 dollars (see Skyrms 1975).

De Finetti's Representation Theorem shows that, under conditions guaranteeing the "exchangeability" of a series of tosses of a coin (i.e., invariance of probabilities with respect to the order of tosses), the subjective probabilities are weighted averages of binomial probabilities with a fixed probability p of tails, where $0 \leq p \leq 1$. Thus, the talk about an "unknown" objective probability of tails can be replaced by talk about a "second-order" subjective probability distribution over the numerical values from 0 to 1. When tosses with the coin are repeated, the second-order distribution will usually become concentrated around some fixed value of p. For de Finetti, this shows that objective probabilities do not exist, while I.J. Good (1983) interprets these results as showing how objective probabilities can be estimated by subjective ones.

Some philosophers, like John Maynard Keynes (1921), have tried to show that there are enough rationality constraints to make degrees of belief or "degrees of confirmation" unique. In this view, *logical probability* is a generalization of the entailment relation between propositions: P(H/E) tells what proportion of the models of E are also models of H. The "objective Bayesians" usually base their suggestions upon some principle of epistemic indifference, informational equality, or maximum entropy (cf. Jeffreys 1939; Rosenkrantz 1977). Rudolf Carnap applied in the 1940s formal tools to construct a system of inductive logic, where the probabilities of statements in a simple first-order language with individual names and one-place properties can be determined (see Carnap 1962). In the 1950s he generalized this approach to a continuum of inductive probability measures (see Carnap 1952).

Carnap understood logical probabilities as degrees of *partial entailment* between propositions. One difficulty with this view is that such degrees seem to depend on parameters which express context-dependent regularity assumptions. Hence, logical probabilities are not completely objective, but relative to some empirical or subjective assumptions. Another problem for Carnap is that in his system all genuinely universal generalizations (such as 'All ravens are black' where the domain is not restricted to any finite number of objects) have the probability 0 given any finite singular evidence. The alleged the zero-probability of universal laws was used as an argument against the programme of inductive logic by Popper (1959, 1963).This problem was solved in 1964 by Jaakko Hintikka whose system of inductive logic allows universal generalizations to receive non-zero probabilities (see Hintikka and Suppes 1966).

Some philosophers are probabilistic "monists" in the sense that they try to reduce all usage of this concept to only one interpretation. Other philosophers favour "pluralism". For example, Carnap argued that frequentist and logical probabilities both exist in different contexts independently of each other. Another kind of pluralism would be to accept the single-case propensity interpretation for the concept of probability in scientific laws, and the personal Bayesian concept for the treatment of epistemic uncertainty within scientific inference.

For the pluralist, it is important to study the interrelations of objective and epistemic probabilities. David Lewis's (1980) Principal Principle states that the rational "credence" of a proposition A given the evidence that the "chance" of A is r equals r:

$$C(A/P(A) = r) = r.$$

This is one way of formalizing arguments like (6) whose premises contains a statistical probability statement and the relation between premises and conclusion is interpreted in terms of epistemic probability.

3.2. Frequentist Approaches to Scientific Inference

Influenced by Venn, Charles S. Peirce defined in 1867 the probability of an argument as a *truth-frequency*, i.e., as the relative number of cases where the argument leads from true premises to a true conclusion. Valid deduction has the truth-frequency one, while inferences from a statistical premise may have a truth-frequency between 0 and 1 (cf. (6)).

Probability as truth-frequency does not qualify the conclusion of an inference but rather the mode of inference. Moreover, the frequentist probability concerns only the repetitions of the inferential pattern. For these reasons, it is questionable whether knowledge of such long-run frequencies indicates anything epistemically important about the short-run or a given single case (cf. Hacking 1965). Reichenbach's and Salmon's attempts to treat the single case face the problem that there does not seem to be a general non-arbitrary way of choosing the relevant reference class.

The frequency definition has, nevertheless, been the common background assumption of the main approaches to statistical inference in the twentieth century. R.A. Fisher (1956) criticized the use of Bayes's Theorem, since the choice of the prior probability cannot be justified on the frequentist basis: statistical tests should be based on upon the concept of likelihood (cf. Seidenfeld 1979). Some of Fisher's followers have understood the likelihood P(E/H) of a hypothesis H relative to observed data E (or the logarithm of this value) as a measure of the inductive support given by E to H (see Edwards 1972).

The "orthodox" Neyman-Pearson theory of statistical estimation and testing also operates with frequentist likelihoods. But Jerzy Neyman (1977) argues that the basic concept in this connection is not "inductive inference" but rather "inductive behaviour": statistics does not attempt to find out the epistemic credentials of a hypothesis or an estimate, but to give recommendations of acting as if some hypothesis or estimate were true; when a series of such decisions are made by following a statistical method, it should lead to successful results in the long run. This idea seems to be applicable in contexts like industrial quality control where the same test is repeated again and again, but its viability in assessing our trust in a scientific hypothesis in a particular single case is problematic. Still, Ron Giere (1979) and Deborah Mayo (1996) are among philosophers who argue that essentially the Neyman-Pearson type of inference is needed in the comparative evaluation of scientific hypotheses.

3.3. Bayesian Approaches to Scientific Inference

In the classical theory of probability, Bayes's Theorem was used for calculating the "inverse" probabilities of unknown causes given their known effects. Laplace gave his famous *Rule of Succession*: given m successes in a series of n experiments, the probability that the next case will be a success is (m + 1)/(n + 1). But the critics, among them George Boole in the 1850s, claimed that in the Bayes's formula

$$(9) \qquad P(H/E) = \frac{P(H)\,P(E/H)}{P(H)\,P(E/H) + P(\sim H)\,P(E/\sim H)}$$

the prior probability P(H) of hypothesis H and the probability P(E/~H) of evidence E given the negation of H can be determined only in an arbitrary way.

Serious interests in the Bayesian approach was reborn in the 1920s with Ramsey and de Finetti who showed in detail under what conditions degrees of belief satisfy the axioms of probability theory (see also Savage 1954). De Finetti's proved by his Representation Theorem that subjective probabilities are not arbitrary but behave in a reasonable intersubjective manner: if two persons agree that a sequence of events is exchangeable but start from different non-zero prior probabilities, their posterior probabilities given the same evidence will converge toward each other.

Another trend was the theory of logical probability from Keynes and Jeffreys to Carnap and Hintikka. In *inductive logic*, inductive probabilities are determined by symmetry assumptions concerning the underlying language (Carnap 1962; Hintikka and Suppes 1970; Niiniluoto and Tuomela 1973). Carnap's one-time favourite measure c* is a generalization of Laplace's rule of succession. But in Carnap's λ-continuum the probabilities depend on a free parameter λ which indicates the weight given to logical or language-dependent factors over and above purely empirical factors (observed frequencies), and in Hintikka's 1965 system one further parameter α is added to regulate the speed in which positive instances increase the probability of a generalization.

More precisely, let $Q_1,..., Q_K$ be a K-fold classification system with mutually exlusive predicates, so that every individual in the universe U has to satisfy one and only one Q-predicate. A *state description* relative to individuals $a_1,..., a_m$ tells for each a_i which Q-predicate it satisfies. A *structure description* tells how many individuals satisfy each Q-predicate. Every sentence within this framework can be expressed as a disjunction of state descriptions. Let e describe a sample of n individuals in terms of the Q-predicates, and let $n_i \geq 0$ be the observed number of individuals in cell Q_i (so that $n_1 + ... + n_K = n$). Carnap's λ-continuum takes the probability $P(Q_i(a_{n+1})/e)$ that the next individual a_{n+1} will be of kind Q_i to be

(10) $$(n_i + \lambda/K)/(n + \lambda).$$

The choice $\lambda = K$ gives Carnap's measure c*, which allocates probability evenly to all structure descriptions. The choice $\lambda = 0$ gives Reichenbach's Straight Rule. The choice $\lambda = \infty$ would give the range measure proposed in Wittgenstein's *Tractatus*, which divides probability evenly to state descriptions, but it makes the inductive probability (10) equal to 1/K which is independent of the evidence e and, hence, does not allow for the learning from experience.

Hintikka's λ-α-system solves the problem of universal generalization by dividing probability to constituents. A *constituent* tells which Q-predicates are non-empty and which empty in universe U. Every generalization (i.e., a quantificational sentence without individual names) can be expressed as a finite disjunction of constituents. When α grows without limit, Hintikka's measures approach in the limit the Carnapian values. When α is small, the probability of universal generalizations grows rapidly. In this sense, the choice of α is an index of boldness of the investigator, or a regularity assumption about the lawlikeness of the relevant universe U. In Hintikka's system, there is one and only one constituent C* which has asymptotically the probability one when the size of the sample e grows without limit; this is the constituent which states that the universe U instantiates precisely those Q-predicates which are exemplified in the sample e.

The Carnap-Kemeny axiomatization of Carnap's λ-continuum was generalized in 1974 by Hintikka and Niiniluoto, who allowed that the inductive probability of the next case being of type Q_i depends on the observed relative frequency of kind Q_i and on the number of different kinds of individuals in the sample e. The latter factor expresses the variety of evidence e. In this way, a system of inductive probability measures is obtained where Carnap's λ-continuum is the only special case with zero probabilities for universal generalizations (see Hintikka and Niiniluoto 1980; Kuipers 1978).

Further developments of inductive logic include its modification to problems concerning analogical reasoning where the distances between Q-predicates play a role in inference (see Helman 1988).

An important assumption of most Bayesian theories of induction is the model of *conditionalization* for revising degrees of belief: if the probability of H at time t is $P_0(H)$, and between t and t+1 a new piece of evidence E is found, the new probability P(H) of H at time t+1 is the conditional probability $P_0(H/E)$. It has been debated whether the principle of conditionalization can be justified by dynamic Dutch Book arguments (see Skyrms 1987; Earman 1992; Howson 1995).

One way of generalizing the simple model of conditionalization is to allow indeterminate (e.g., interval valued) prior and posterior degrees of belief (cf. Levi 1991). Another idea is Richard Jeffrey's (1965, 1992) probability kinematics for cases with uncertain evidence: if we see object a dimly by candlelight, and $P(Q_i(a))$, $i = 1,..., K$, are our new probabilities about the color of a after the observation, and the rigidity condition $P(H/Q_i(a)) = P_0(H/Q_i(a))$ holds for statement H, then

$$(11) \qquad P(H) = \sum_{i=1}^{K} P(Q_i(a))P_0(H/Q_i(a)).$$

If we learn that a is really Q_m, so that $P(E) = 1$ for $E = Q_m(a)$, then the Jeffrey formula (11) reduces to ordinary conditionalization: $P(H) = P_0(H/E)$.

A general axiomatic treatment of belief revision has been developed by Peter Gärdenfors (1988). Besides the expansion of a system of beliefs, it also takes into account contractions and revisions due to inconsistencies between old beliefs and new data.

3.4. Confirmation and Acceptance

Following Keynes, Hempel, and Hosiasson-Lindenbaum, Carnap called the inductive probabilities of his system *degrees of confirmation*. But there was also another idea of confirmation which was important in the debate whether the positive instances of a generalization "confirm" or "support" a universal generalization.

According to Jean Nicod's criterion, only positive instances of the form {Fa, Ga} confirm the generalization $\forall x(Fx \rightarrow Gx)$, while negative instances of the form {Fa, ~Ga} disconfirm it, and the cases {~Fa, Ga} and {~Fa, ~Ga} are neutral with respect to it. Carl G. Hempel argued in 1937 that, as 'All ravens are black' and 'All non-black things are non-raven' are logically equivalent, both black ravens and white handkerchiefs should be understood to confirm the hypothesis about the color of ravens. Janina Hosiasson-Lindenbaum (1940) gave the first Bayesian analysis of this "raven paradox" by arguing that an observed black raven increases *more* the inductive probability of the generalization than any non-black non-raven.

Karl Popper argued in the 1950s against Carnap that inductive logic is inconsistent and impossible (see Popper, 1959, 1963). As a reply to Popper's criticism, Carnap (1962) distinguished "degrees of confirmation" in two senses: as the posterior probability (i.e., P(H/E)) and as the increase of probability of H due to E (i.e., P(H/E) – P(H)). The qualitative concept of confirmation corresponding to these two alternatives can be defined by *high probability* (i.e., P(H/E) is larger than some fixed threshold value) or by the *positive relevance* condition (i.e., P(H/E) > P(H)). The corresponding comparative conceptions (cf. Lakatos, 1968) "E confirms H_1 more than H_2" can then be defined either by $P(H_1/E) > P(H_2/E)$ or by $P(H_1/E) - P(H_1) > P(H_2/E) - P(H_2)$. These definitions can also be relativized to background knowledge.

The high probability criterion satisfies the entailment condition: if E logically entails H, then E confirms H, and the special consequence condition: if E confirms H and H entails H', then E confirms H'. On the other hand, the positive relevance criterion satisfies the converse entailment condition: if H entails E, then E confirms H. It is known that these conditions cannot be satisfied together (Hempel 1965). Moreover, neither of the accounts of confirmation satisfy the conjunction condition: if E confirms H_1 and E confirms H_2, then E confirms $H_1\&H_2$.

Other proposals for the degree of confirmation of H given E are usually normalizations of the difference P(H/E) – P(H) which is equal to

$$(12) \qquad [P(E/H) - P(E)]P(H)/P(E).$$

They include I.J. Good's 1950 measure for the "weight of evidence"

$$(13) \qquad \log P(E/H) - \log P(E/{\sim}H),$$

Kemeny's and Oppenheim's 1952 measure for "factual support"

$$(14) \qquad \frac{P(E/H) - P(E/\sim H)}{P(E/H) + P(E/\sim H)}$$

Popper's 1954 formula for "degrees of corroboration", and Hintikka's formula for the information transmitted by E on H (see Hintikka 1968; Niiniluoto and Tuomela 1973). All these measures are greater than 0 if and only if E is positively relevant to H. On the other hand, if degrees of confirmation are defined by

$$(15) \qquad P(H/E)/P(H) = P(E/H)/P(E),$$

then comparative confirmation satisfies the Likelihood Criterion that E confirms more H_1 than H_2 if and only if $P(E/H_1) > P(E/H_2)$.

An important paradox of confirmation was suggested by Nelson Goodman's (1955) odd predicate 'grue' (i.e., green if examined before 2100, and blue otherwise). According to Goodman, all the evidence so far on the color of emeralds supports equally well the two incompatible hypotheses ‘All emeralds are green’ and ‘All emeralds are grue’. Why are we willing to “project” for the future the predicate ‘green’ rather than ‘grue’? Goodman concludes that induction depends on such pragmatic factors as our actual familiarity in using some predicates in our language. Goodman’s argument shows that inductive probabilities are not defined by purely syntactical conditions, but this is compatible with the idea that inductive logic may involve extra-logical or contextual parameters.

Clark Glymour (1980), who proposes to replace Bayesianism with his “bootstrap method”, has presented the “problem of old evidence” against the positive relevance account of confirmation. Suppose that evidence E is known at time t when theory H is introduced. Then at time t we have P(E) = 1, P(E/H) = 1, and by (7) P(H/E) = P(H). Hence, old evidence cannot confirm a new theory. But this is counterintuitive in the light of many examples from the history of science (cf. Howson and Urbach 1989). There is no agreement of the best way to handle this problem (see Earman 1992). As Glymour himself noted, this problem is related to the idealized assumption (4) that degrees of belief are invariant under logical equivalence. In this sense, they are probabilities for a logically omniscient scientist, and in a more realist treatment they should be replaced by some kind of “surface probabilities” which allow that the discovery of new deductive relations (e.g., that hypothesis H entails the old evidence E) may influence inductive probabilities.

Some Bayesians (like Carnap and Jeffrey) think that the theory of induction only tells how the probabilities of hypotheses are determined; these probabilities can then be employed in rational decision making using the formula (8) (Jeffrey 1965). In this account, scientists are themselves decision makers or advisers of decision makers. The values relevant to decisions, related to the good and bad consequences of actions, are practical utilities defined by the employer or the society. This approach, which leads to Bayesian decision theory, resembles Neyman’s conception of inductive behaviour: according to L.J. Savage’s “behaviouralism”, accepting a hypothesis means only that we are ready to act as if it were true.

Against Savage, Isaac Levi’s (1967) “cognitivism” argues that scientists tentatively accept hypotheses as parts of the evolving body of scientific knowledge. Levi and Hintikka have formulated inductive rules for the tentative *acceptance* of hypotheses on the basis of evidence (see Hilpinen 1968). One variant of inductive acceptance is the so-called inference to the best explanation: among rival hypotheses, it recommends the acceptance of the hypothesis that gives the best explanation of the given data.

The notion of acceptance clearly differs structurally from probability: for example, if hypothesis H is acceptable on evidence E, then ~H cannot be acceptable on E, even though both H and ~H may have non-zero probabilities given E. L.J. Cohen (1989) has proposed a non-Bayesian treatment of inductive support, but Levi argues that Cohen’s “Baconian probabilities” (which Cohen contrasts with the ordinary “Pascalian” probabilities) are variants of “degrees of confidence of acceptance” (cf. also Shafer 1976, 1996).

A powerful reformulation of the problem of induction has been given in cognitive decision theory by Levi and Hintikka. This theory shows that scientific induction can be treated in decision-theoretical terms, but then the relevant cognitive goals to be used in (8) are defined by *epistemic utilities*.

If the aim of our inquiry is truth, and nothing but the truth, the epistemic utility of accepting a hypothesis H on evidence E can be taken to be equal to its truth value (1 for truth, 0 for falsity). Then the expected utility of accepting H is simply P(H/E).1 + P(~H/E).0 = P(H/E). The rule of maximizing expected epistemic utility leads to the conservative principle of accepting only trivially true tautologies or hypotheses logically entailed by the evidence.

But if our aim is truthful information, Levi (1967) points out that we have to "gamble with truth" in order to gain other epistemic utilities. Levi assumes that the relevant hypotheses in an inductive decision problem are disjunctions of mutually exclusive and jointly exhaustive alternatives h_i constituting an "ultimate partition" B, and the information content c(H) of a hypothesis H depends on the number of elements of B excluded by H. More precisely, let $|H|$ be the number of alternatives in B allowed by H, and $|B|$ be the total number of elements in B. Letting $0 < q \leq 1$ to be an index of boldness, which tells how willing the scientist is to risk error in the attempt to relieve from agnosticism, Levi suggests that the epistemic utility of accepting H is $1 - q|H|/|B|$ when H is true and $-q|H|/|B|$ when H is false. This is essentially a weighted average of the truth value of H and the content c(H) of H. Levi's choice leads to the expected utility

(16) $$P(H/E) - q|H|/|B|,$$

and the following rule of acceptance: reject all elements h_i of B with $P(h_i/E) < q/|B|$, and accept the disjunction of all unrejected elements of B as the strongest on the basis of E.

If the information content of H is measured by

(17) $$\text{cont}(H) = 1 - P(H),$$

as suggested by Carnap, Bar-Hillel, and Popper, our gain in accepting H is cont(H) when H is true and our loss is cont(~H) when H is false, so that the expected utility is P(H/E) − P(H) (see Hintikka and Suppes, 1966; Hilpinen, 1968). This can be again written as the sum of P(H/E) and cont(H). These formulas show that it is possible to combine and balance the Popperian demand that science strives for bold (informative, a priori improbable) hypotheses and the traditional Bayesian demand for well-supported (a posteriori probable) hypotheses. The simple rule of maximizing (17) directly would lead, instead, to the unsatisfactory recommendation of accepting a logical contradiction.

Similar results are obtained, if cont-measure is replaced by the *systematic* (explanatory or predictive) *power* of H relative to E. Some measures of explanatory power are directly variants of formulas (13) – (16), so that E gives the highest degree of confirmation to that H which best explains E (cf. Hintikka 1968; Niiniluoto and Tuomela 1973). Hempel's (1965) proposal for systematic power is

(18) $$\text{syst}(H,E) = P(\sim H/\sim E).$$

Again the rule of maximizing (18) would recommend the acceptance of a logical contradiction. But if our gain is taken to be syst(H,E) if H is true and -syst(~H/E) if H is false, then the best hypothesis H is one which maximizes P(H/E) – P(H) (see Pietarinen 1970; Niiniluoto 1999a). These results can be understood as formalizations of the idea that induction can be treated as an *inference to the best explanation* (see Harman 1965).

Another way of combining Popperian and Bayesian elements in the theory of scientific inference is to view science as an attempt to maximize expected *verisimilitude*, where verisimilitude or truthlikeness is a measure of the "closeness" of a hypothetical theory to some interesting and informative truth (see Niiniluoto 1987). The mini-sum-measure of Niiniluoto can be understood as a generalization of Levi's assignment of epistemic utility: for a disjunctive hypothesis, truth value is replaced by the minimum distance from the truth, and information content by a normalized sum of the distances of disjuncts from the truth. If all false basic alternatives are equally distant from the truth, this measure reduces to Levi's proposal. When the truth is unknown, truthlikeness can be estimated by calculating the expected value of this distance. The main difference to probabilistic measures of confirmation and corroboration is then the possibility that a hypothesis H which is known to be refuted by evidence E may nevertheless be judged to be highly truthlike.

Some typical methods of Bayesian decision theory can be reinterpreted in terms of maximization of expected verisimilitude (see Niiniluoto, 1987; Festa, 1993). For example, a point estimate θ_0 of a real valued parameter θ should be chosen so that the posterior loss

$$(18) \qquad \int_R |\theta - \theta_0| p(\theta/e) d\theta$$

is minimized, where $p(\theta/e)$ is the posterior probability distribution of θ given evidence e. Here (18) is clearly the expected distance of θ_0 from the truth. The same treatment can be generalized to interval hypotheses.

Closeness to the truth is an important ingredient of *curve-fitting* problems. Assume that we are investigating the lawlike interrelation between two quantities x and y, and let $<x_1,y_1>$, ..., $<x_n,y_n>$ represent n points obtained by measuring the values of x and y. According to Reichenbach's (1938) formulation, the simplest curve that goes through these points expresses the most probable law of the form y = f(x). In practice, however, the statistical regression methods seek sufficiently simple curves (such as linear and quadratic functions) such that the distances of the observed points from the curve is as small as possible. The traditional Method of Least Squared Differences is an example of such an approach. In other words, given a class of simple functions, the least false or inaccurate among them is the best one (see Niiniluoto 1999a). (For the concept of simplicity, see Foster and Martin 1966; Hesse, 1974.) Recent work in statistics suggests how the demands of simplicity and accuracy can be combined and balanced in curve-fitting (see Forster and Sober 1994).

4. THE ROLE OF INDUCTION IN SCIENCE

Our common sense conceptions and beliefs are largely learned from experience, and thus rely in some way or another on inductive inference. The role of induction in science has been the subject of a lively debate.

4.1. Inductivism

According to Aristotle, *epagoge* or induction is the method of reaching the axioms or first principles of each science, and the route from axioms to theorems goes via deductive syllogisms. In spite of his leaning toward some kind of empiricism, he was not able to make clear how the process of induction depends on sense experience. Modern rationalists claimed instead that the axioms of science are obtainable by pure reason: their self-evidence is based upon "clear and distinct ideas", as Descartes put it (see Blake, Ducasse, and Madden 1965).

Francis Bacon criticized the Aristotelian conception of science (see Bacon, 1960), since in his view scientific inference proceeds gradually from singular observations to more general truths, and the most general axioms are reached only at the last stage. Instead of simple induction by enumeration, inductive generalization is based upon a method of elimination. This process is a routine or mechanical method, and its application helps the scientist to avoid the deceiving errors of observation or common prejudices. Therefore, eliminative induction leads to results which are conclusively certain. In the Baconian tradition, followed still by John Stuart Mill in the nineteenth century, induction was assumed to be both a method of discovery and a method of proof of scientific laws.

Today the doctrine that scientific theories are discovered by induction is known as *inductivism*. This view is a form of naive empiricism which assumes that science starts by collecting large amounts of observational data and then makes generalizations from them. The view that "facts speak for themselves" ignores that scientists need background theories to guide the collection of data and to interpret them. Science starts from some cognitive problems, and observations are relevant to a problem only when some initial hypothesis has been formulated as the object of study. Moreover, the relevant hypotheses are usually not derivable from the data by any mechanical method, but they are rather discovered to explain the observed facts (see Hempel 1966). Scientific theories contain theoretical terms which seem to refer to unobservable entities (see Laudan 1981). As Pierre Duhem pointed out in 1906, there are important examples where the new theory is inconsistent with the initial observations: Newton's theory corrects Kepler's laws. Such discovery presupposes creativity (Whewell 1860). Even in cases, where the discovery of hypothesis is suggested by induction and analogy, the hypothesis has to be tested by new independent observations.

4.2. Hypothetico-Deductive Method

Inductivism is opposed by the *hypothetico-deductive* (HD) conception of science: scientific statements and theories are free creations, hypotheses, that are tested by deducing empirical predictions from them. For an instrumentalist, such theoretical statements are uninterpreted schemas without a truth value, but the HD-method is usually associated with a realist view which takes theories to be genuine statements about reality (see Niiniluoto 1999a). A negative test result refutes the hypothesis by modus tollens, and a positive result gives confirmation or inductive support to the hypothesis. Induction is thus a part of the method of science, but it belongs exclusively to the context of testing and justifying hypotheses. The path of discovery is irrelevant in the assessment of the merits of the hypothesis, and there is no logical reconstruction of scientific discovery.

In the HD-method, a hypothesis is usually required to satisfy some initial conditions: it should be logically consistent, compatible with background theories, exactly formulated, testable in principle, informative, and simple. Further, it should solve the initial problem of explaining the observed facts. In Bayesian reconstructions of the HD-method, these "plausibility" conditions are usually built into assumptions concerning the prior probability P(H) of the hypothesis (cf. Salmon 1966). Indeed, such prior probabilities are decisive for the comparison of hypotheses which entail the evidence: if H and H' logically entail E, then P(H/E) > P(H/E) if and only if P(H/E) – P(H) > P(H'/E) – P(H') if and only if P(H) > P(H'). A plausible hypothesis is then testworthy, and subsequent tests are needed to decide whether it is also trustworthy. As Whewell (1840) argued, a good hypothesis should also foretel phenomena which are different from the ones that it already is known to explain. Again a Bayesian formulation is possible: if H entails a contingent observational statement E, and E turns out to be true, then E confirms H by the positive relevance criterion. The increase of the probability of H is the greater, the less probable E itself is.

4.3. Falsificationism

Karl Popper's (1963) *falsificationism* basically accepts the hypothetico-deductive model: science proceeds by proposing bold conjectures and by putting them in severe tests. This schema of learning from mistakes (i.e., from a problem to a tentative theory, and via error elimination to a new problem) is common to an amaeba and Einstein (Popper 1972). However, Popper denies that the HD-method has an inductive element: for him, theories always remain as conjectures, and they are never accepted as true or probable on the basis of observations. Hypotheses can be "corroborated", but only in the sense that they may for some time pass the most severe tests.

The critics claim that Popper in practice cannot avoid some inductive elements in his account of the growth of knowledge: modus tollens + corroboration = induction, as Salmon (1966) puts it. Popper defends his position by claiming that theories are "accepted" in science only for the purpose of testing them. However, it is difficult to see how Popper could account for the rationality of acting on the basis of the best-tested theory (see Popper 1972; Miller 1994) without some inductive assumption (cf. Niiniluoto and Tuomela 1973). Popper also claims that corroboration gives no prediction that the hypothesis will survive tests also in the future. However, he has also suggested that corroboration is an indicator of verisimilitude (see Popper 1972), but again the attempts to make this idea precise seem to involve some element of induction (cf. Niiniluoto 1987).

Further, it can be argued that the falsificationist's basic goal of refuting bold conjectures yields only small gains, if epistemic utilities are measured by the cont-function (17): if H turns out to be false, we gain the information content of ~H, but this value cont(~H) = 1 – P(~H) = P(H) is small, if H is a bold hypothesis. Our gain is large, instead, if we accept a bold hypothesis.

4.4. Alternative Views

One famous difficulty, known as the *Duhem-Quine problem*, for the method of hypothesis arises from the fact that scientific theories entail observational consequences only together with some auxiliary assumptions. But if H and A together entail E, then the falsity of E does not refute H any more, as the mistake may lie in A (Hempel 1966). This difficulty seems to show that the refutation of scientific hypotheses is no more conclusive than their confirmation. Scientists thus need some methodological principles or conventions for protecting some parts of their theoretical assumptions. Thomas Kuhn's normal science and Imre Lakatos' methodological research programmes are attempts to deal with this problem (see Lakatos and Musgrave 1980).

The HD-method has been criticized also for its failure to say anything interesting about the discovery of hypotheses. Already Peirce argued that there is a third mode of inference besides deduction and induction: *abduction* is the process of adopting a hypothesis which would explain some surprising facts. Thus, there might be at least a partial "logic of discovery" after all. George Polya (1945) has shown that inductive generalization and analogy may play an important role in heuristic reasoning in mathematics, and it is possible to teach even a computer programme to make discoveries on the basis of some given data (see Langley *et al.* 1987; Gabbay and Smets 1998). The "friends of discovery" are currently studying the question whether there are inductive or non-inductive logics of discovery (Nickles 1980; Magnani, Nersessian, and Thagard 1999).

Ilkka Niiniluoto
University of Helsinki

REFERENCES

Bacon, F.: 1960, *The New Organon*, The Bobbs-Merrill Company, Indianapolis.

Bar-Hillel, Y.: 1964, *Language and Information*, Addison-Wesley, Reading, Mass.

Barker, S.: 1957, *Induction and Hypothesis*, Cornell University Press, Ithaca.

Black, M.: 1954, *Problems of Analysis*, Routledge, London.

Blake, R. M., C. J. Ducasse, and E. H. Madden: 1960, *Theories of Scientific Method: The Renaissance Through the Nineteenth Century*, The University of Washington Press, Seattle.

Bogdan, R. J.: 1976, *Local Induction*, D. Reidel, Dordrecht.

Bogdan, R. J. (ed.): 1982, *Henry E. Kyburg & Isaac Levi*, Profiles, D. Reidel, Dordrecht.

Boyd, R.: 1984, 'The Current Status of Scientific Realism', in J. Leplin (ed.), *Scientific Realism*, University of California Press, Berkeley, pp. 41-82.

Butts, R. E. (ed.): 1968, *William Whewell's Theory of Scientific Method*, University of Pittsburgh Press, Pittsburgh.

Byrne, E. F.: 1968, *Probability and Opinion: A Study in the Medieval Presuppositions of Post-Medieval Theories of Probability*, Martinus Nijhoff, The Hague.

Carnap, R.: 1952, *The Continuum of Inductive Methods*, The University of Chicago Press, Chicago.

Carnap, R.: 1962, *The Logical Foundations of Probability*, 2nd ed., The University of Chicago Press, Chicago.

Cohen, L. J.: 1989, *An Introduction to the Philosophy of Induction and Probability*, Oxford University Press, Oxford.

Earman, J.: 1992, *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*, The MIT Press, Cambridge, Mass.

Edwards, A. W. F.: 1972, *Likelihood*, Cambridge University Press, Cambridge.

Festa, R.: 1993, *Optimum Inductive Methods*, Kluwer, Dordrecht.

Fetzer, J. H.: 1981, *Scientific Knowledge: Causation, Explanation, and Corroboration*, D. Reidel, Dordrecht.

Fetzer, J. (ed.): 1988, *Probability and Causality*, D. Reidel, Dordrecht.

Fine, T.: 1973, *Theories of Probability*, Academic Press, New York.

de Finetti, B.: 1972, *Probability, Induction, and Statistics: The Art of Guessing*, Wiley, New York.

Fisher, R. A.: 1956, *Statistical Methods and Scientific Inference*, Oliver and Boyd, Edinburgh.

Forster, M. and E. Sober: 1994, 'How to Tell when Simpler, More Unified, or Less *ad hoc* Theories will Provide More Accurate Predictions', *The British Journal for the Philosophy of Science* **45**, 1-35.

Foster, M. and M. Martin (eds.): 1966, *Probability, Confirmation, and Simplicity*, The Odyssey Press, New York.

Gabbay, D. M. and P. Smets (eds.): 1998, *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, vol. I-IV, Kluwer, Dordrecht.

Gärdenfors, P.: 1988, *Knowledge in Flux: Modeling the Dynamics of Epistemic States*, The MIT Press, Cambridge, MA.

Gärdenfors, P. and N.-E. Sahlin (eds.): 1988, *Decision, Probability and Utility*, Cambridge University Press, Cambridge.

Giere, R.: 1979, *Understanding Scientific Reasoning*, Holt, Rinehart, and Winston, New York.

Glymour, C.: 1980, *Theory and Evidence*, Princeton University Press, Princeton.

Good, I. J.: 1983, *Good Thinking*, University of Minnesota Press, Minneapolis.

Goodman, N.: 1955, *Fact, Fiction, and Forecast*, Bobbs-Merrill, Indianapolis.

Hacking, I.: 1965, *The Logic of Statistical Inference*, Cambridge University Press, Cambridge.

Hacking, I.: 1975, *The Emergence of Probability*, Cambridge University Press, Cambridge.

Harman, G.: 1965, 'Inference to the Best Explanation', *The Philosophical Review* **74**, 88-95.

Helman, D. H. (ed.): 1988, *Analogical Reasoning*, Kluwer, Dordrecht.

Hempel, C. G.: 1965, *Aspects of Scientific Explanation*, The Free Press, New York.

Hempel, C. G.: 1966, *The Philosophy of Natural Science*, Prentice-Hall, Englewood Cliffs.

Hesse, M.: 1974, *The Structure of Scientific Inference*, Macmillan, London.

Hilpinen, R.: 1968, *Rules of Acceptance and Inductive Logic*, Acta Philosophica Fennica, North-Holland, Amsterdam.

Hintikka, J.: 1968, 'The Varieties of Information and Scientific Explanation', in B. van Rootselar and J. E. Staal (eds.), *Logic, Methodology and Philosophy of Science III*, North-Holland, Amsterdam, pp. 151-171.

Hintikka, J.: 1973, *Logic, Language-Games, and Information: Kantian Themes in the Philosophy of Logic*, Oxford University Press, Oxford.

Hintikka, J. and I. Niiniluoto: 1980, 'An Axiomatic Foundation for the Logic of Inductive Generalization', in Jeffrey, 1980, pp. 157-181.

Hintikka, J. and P. Suppes (eds.): 1966, *Aspects of Inductive Logic*, North-Holland, Amsterdam.

Hintikka, J. and P. Suppes (eds.): 1970, *Information and Inference*, D. Reidel, Dordrecht.

Hosiasson-Lindenbaum, J.: 1940, 'On Confirmation', *The Journal of Symbolic Logic* **5**, 133-148.

Howson, C.: 1995, 'Theories of Probability', *The British Journal for the Philosophy of Science* **46**, 1-32.

Howson, C. and P. Urbach: 1989, *Scientific Reasoning: The Bayesian Approach*, Open Court, La Salle.

Hume, D.: 1902, *An Enquiry Concerning Human Understanding*, Clarendon Press, Oxford.

Jeffrey, R.: 1965, *The Logic of Decision*, McGraw-Hill, New York.

Jeffrey, R. (ed.): 1980, *Studies in Inductive Logic and Probability*, vol. II, University of California Press, Berkeley.

Jeffrey, R.: 1992, *Probability and the Art of Judgement*, Cambridge University Press, Cambridge.

Jeffreys, H.: 1939, *Theory of Probability*, Oxford University Press, Oxford.

Kelly, K.: 1996, *The Logic of Reliable Inquiry*, Oxford University Press, Oxford.

Keynes, J. M.: 1921, *A Treatise on Probability*, Macmillan, London.

Kuipers, T.: 1978, *Studies in Inductive Probability and Rational Expectation*, D. Reidel, Dordrecht.

Kyburg, H. E. and H. Smokler (eds.): 1964, *Studies in Subjective Probability*, John Wiley, New York.

Lakatos, I. (ed.): 1968, *The Problem of Inductive Logic*, North-Holland, Amsterdam.

Lakatos, I. and A. Musgrave (eds.): 1970, *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge.

Langley, P., H. A. Simon, G. L. Bradshaw, and J. M. Zytkow: 1987, *Scientific Discovery: Computational Explorations on the Creative Processes*, The MIT Press, Cambridge, MA.

Laplace, P. S.: 1951, *A Philosophical Essay on Probabilities*, Dover, New York.

Laudan, L.: 1981, *Science and Hypothesis*, D. Reidel, Dordrecht.

Levi, I.: 1967, *Gambling With Truth: An Essay on Induction and the Aims of Science*, Alfred A. Knopf, New York.

Levi, I.: 1991, *The Fixation of Belief and Its Undoing: Changing Beliefs through Inquiry*, Cambridge University Press, Cambridge.

Lewis, D.: 1980, 'A Subjectivist's Guide to Objective Chance', in Jeffrey, 1980, pp. 263-293.

Mackie, J.: 1974, *The Cement of the Universe: A Study of Causation*, Oxford University Press, Oxford.

Magnani, L., N. Nersessian, and P. Thagard (eds.): 1999, *Model-Based Reasoning in Scientific Discovery*, Kluwer/Plenum, New York.

Mayo, D.: 1996, *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

Mill, J. S.: 1843, *A System of Logic*, Longmans, Green, and Co., London.

Miller, D.: 1994, *Critical Rationalism: A Restatement and Defence*, Open Court, Chicago and La Salle, Ill.

Neyman, J.: 1977, 'Frequentist Probability and Frequentist Statistics', *Synthese* **36**, 97-131.

Nickles, T. (ed.): 1980, *Scientific Discovery, Logic, and Rationality*, D. Reidel, Dordrecht.

Niiniluoto, I.: 1984, *Is Science Progressive?*, D. Reidel, Dordrecht.

Niiniluoto, I.: 1987, *Truthlikeness*, D. Reidel, Dordrecht.

Niiniluoto, I.: 1995, 'Hintikka and Whewell on Aristotelian Induction', *Grazer philosophische Studien* **49**, 49-61.

Niiniluoto, I.: 1999a, *Critical Scientific Realism*, Oxford University Press, Oxford.

Niiniluoto, I.: 1999b, 'Defending Abduction', *Philosophy of Science* **66** (*Proceedings*), S436-S451.

Niiniluoto, I. and R. Tuomela: (1973), *Theoretical Concepts and Hypothetico-Inductive Inference*, D. Reidel, Dordrecht.

Peirce, C.S.: 1931-35, *Collected Papers*, vols. 1-6, C. Hartshorne and P. Weiss (eds.), Harvard University Press, Cambridge, Mass.

Pietarinen, J.: 1970, 'Quantitative Tools for Evaluating Scientific Systematizations', in Hintikka and Suppes, 1970, pp. 123-147.

Polya, G.: 1945, *How to Solve It: A New Aspect of the Mathematical Method*, Princeton University Press, Princeton.

Popper, K.: 1957, 'The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory', in S. Körner (ed.), *Observation and Interpretation*, Butterworth Scientific, London, pp. 65-70.

Popper, K.: 1959, *The Logic of Scientific Discovery*, Hutchinson, London.

Popper, K.: 1963, *Conjectures and Refutations: The Growth of Scientific Knowledge*, Routledge and Kegan Paul, London.

Popper, K.: 1972, *Objective Knowledge* (2nd ed. 1979), Oxford University Press, Oxford.

Ramsey, F.: 1950, *The Foundations of Mathematics*, Routledge and Kegan Paul, London.

Reichenbach, H.: 1938, *Experience and Prediction*, The University of Chicago Press, Chicago.

Reichenbach, H.: 1949, *The Theory of Probability*, University of California Press, Berkeley.

Rosenkrantz, R.: 1977, *Inference, Method and Decision: Toward a Bayesian Philosophy of Science*, D. Reidel, Dordrecht.

Ross, W.D.: 1949, *Aristotle's Prior and Posterior Analysis*, A Revised text with Introduction and Commentary, Clarendon Press, Oxford.

Salmon, W.: 1966, *The Foundations of Scientific Inference*, University of Pittsburgh Press, Pittsburgh, 1966.

Salmon, W.: 1984, *Scientific Explanation and the Causal Structure of the World*, Princeton University Press, Princeton.

Savage, L. J.: 1954, *The Foundations of Statistics*, Wiley, New York.

Seidenfeld: 1979, *Philosophical Problems of Statistical Inference: Learning from R. A. Fisher*, D. Reidel, Dordrecht.

Shafer, G.: 1976, *A Mathematical Theory of Evidence*, Princeton University Press, Princeton.

Shafer, G.: 1996, *The Art of Causal Conjecture*, The MIT Press, Cambridge, MA.

Skyrms, B.: 1975, *Choice and Chance: An Introduction to Inductive Logic*, 2nd ed., Dickenson, Belmont.

Skyrms, B.: 1987, 'Dynamic Coherence and Probability Kinematics', *Philosophy of Science* **54**, 1-120.

Stegmüller, W.: 1973, *Personelle und statistische Wahrscheinlichkeit*, Springer-Verlag, Berlin.

Strawson, P.: 1952, *Introduction to Logical Theory*, Methuen, London.

Suppes, P.: 1984, *Probabilistic Metaphysics*, Blackwell, Oxford.

Swinburne, R. (ed.): 1974, *Justification of Induction*, Oxford University Press, Oxford.

Tarski, A.: 1956, *Logic, Semantics, Metamathematics*, Oxford University Press, Oxford.

Venn, J.: 1866, *The Logic of Chance*, Macmillan, London.

von Mises, R.: 1951, *Probability, Statistics, and Truth*, 2nd rev. ed., Allen and Unwin, London.

von Plato, J.: 1994, *Creating Modern Probability*, Cambridge University Press, Cambridge.

von Wright, G. H.: 1951, *A Treatise on Induction and Probability*, Routledge and Kegan Paul, London.

von Wright, G. H.: 1957, *The Logical Problem of Induction*, 2nd ed., Blackwell, Oxford.

Whewell, W.: 1840, *The Philosophy of the Inductive Sciences*, John W. Parker and Sons, London.

Whewell, W. 1860, *On the Philosophy of Discovery*, John W. Parker and Sons, London. (Reprinted by Lenox Hill, New York, 1971.)

Wittgenstein, L.: 1922, *Tractatus Logico-Philosophicus*, Routledge and Kegan Paul, London.

PART IV: EPISTEMOLOGY AND AREAS OF KNOWLEDGE

PAUL HUMPHREYS

SCIENTIFIC KNOWLEDGE

INTRODUCTION

A discussion of scientific knowledge requires paying special attention to the distinctive methods that science has developed to acquire and evaluate knowledge. There is no prima facie reason why, in the light of these methods, the epistemology of science should, to any great extent, resemble traditional epistemology, the study of knowledge gained by unassisted humans. Indeed, we shall see that not only are those traditional epistemological concerns often rather remote from those that are relevant to science, but insisting on applying constraints from older epistemological traditions can seriously distort our assessment of scientific knowledge.

A second issue that lies at the heart of scientific knowledge is the need to find an appropriate balance between the 'in principle' interests of traditional epistemology and the 'in practice' demands of a realistic science. Much epistemology has concerned itself with idealized epistemic agents operating with perfect data sources. While this kind of idealization and the resulting interest in what it is possible in principle to know is entirely appropriate – most notably when one is interested in what it is not possible to know, even for such an idealized agent – such scenarios are far removed from the actual processes of acquiring knowledge in science.

Perhaps the most obvious feature of scientific activity is the way in which unaided human epistemic abilities have been vastly expanded by specifically scientific techniques, both mathematical and instrumental. The traditional empiricist/rationalist division[1] is ill-suited to account for this expansion.

The division in traditional epistemology between a priori and a posteriori knowledge was based on the origin of the knowledge concerned, roughly whether the justification for the knowledge claim required sensory input or not[2]. (A similar division might be made for certain sorts of instruments – does the instrument require detectors that sense its environment or is the justification independent of such inputs?[3])

Although a priori knowledge is, of course, employed both in constructing and in using the mathematical representations that occur in many scientific theories, few would now be willing to argue that purely mathematical considerations serve as the criteria of theory choice even in the most rarified of physical theories. Symmetry considerations in quantum theory, for example, are ultimately constrained by what is actually the case. In other areas where a priori methods have sometimes seemed acceptable, such as the development of rational choice theory, there has been a steady accumulation of evidence that these theories are descriptively false of much of human decision making.(See e.g. Kahneman et al., 1982). That is, although a priori methods can legitimately ground normative standards in these areas, these

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 549-569.
© 2004 *Kluwer Academic Publishers.*

standards cannot serve as the basis of a scientific theory of economic or of psychological behaviour because a scientific theory must conform, within certain limits, to the best empirical data, and whatever purpose a priori considerations may serve in initially suggesting a theory, these should always be overridden by empirical considerations when the theory is used scientifically. Alternatively, it can be claimed that humans use a variety of alternative strategies for dealing with the world and that different criteria of rationality are appropriate in different circumstances, these criteria sometimes being discoverable only a posteriori. (See e.g. Gigerenzer et al., 1999.)

A third place where a priori criteria have been employed is with so-called pragmatic criteria for theory choice, where constraints such as simplicity are used to select a preferred candidate from a number of empirically equivalent alternatives. Yet it is easy to overstate the importance of these a priori criteria. Theories that are exactly empirically equivalent are rare, and even in those cases where one can give real (as opposed to artificial) examples, such as with structurally different causal models in the social sciences and epidemiology that produce identical correlation matrices, substantive scientific knowledge about biological plausibility will often rule out all but one of the rivals. (See Hill, 1965) for various criteria of this kind.) In other cases, theory choice has been presented as something that is forced upon a scientific community whereas in fact it is usually feasible, and indeed proper, to allow a number of more or less empirically equivalent rival theories to continue until further evidence distinguishes between them. To take a well known example, even supposing that the Ptolemaic and Copernican theories of planetary motion were actually empirically equivalent in 1543[4], the 1609-10 Galilean observations that showed evidence of Venusian phases and the nineteenth century evidence from Foucault's pendulum later provided good empirical evidence against traditional geocentric accounts. Of course this evidence is not conclusive, but the alternative hypotheses needed to construct a new rival theory are usually either unsupported by the contemporary scientific evidence and hence ad hoc, or are counter-indicated by it. Theory choice is rarely instantaneous and it is no flaw in science to postpone a decision until more empirical evidence is available.

Within the a posteriori realm, requiring human sensory experience to be the ultimate arbiter of what counts as scientific knowledge is an unrealistic constraint for the purposes of science. One of the most important features of science is its use of instrumentation to extend our native sensory equipment and its development of computational devices, both theoretical and physical, to enhance our natural computational abilities. Despite these developments there is a curious bias in much of the empiricist literature, because it is willing to consider idealized epistemic agents and limit science yet it ordinarily refuses to relinquish its commitment to basic observation statements and to conform to successful contemporary practice. I shall have more to say about this later.

In the later part of this century there has been a pronounced move from viewing epistemology as best approached through logical reconstructions to a wider perspective within which individual psychological factors and sociological factors are held to play a role in knowledge acquisition and evaluation. The logical approach was exemplified by such exercises as Carnap's inductive logic (Carnap, 1962) and the study of various epistemic logics (Hintikka, 1962). The use of psychology has evolved from the simple behaviourism of early and mid-career Quine (Quine, 1958, 1960) and of Kuhn (Kuhn, 1974) to the more realistic but still elementary studies of cognitive scientists. (See e.g. Churchland and Churchland, 1998, §III)[5]. The move towards an emphasis on psychology and sociology has often been called a move to a naturalized epistemology, but this is a term that has no fixed meaning. (See e.g. Almeder, 1990, also Kitcher, 1992). The great dividing line in all of these areas is between normative standards and descriptive studies, whether historical or contemporary. The normative standards are closely connected with what is possible in principle versus what is possible in practice, with competence versus performance, with ideal versus real epistemic agents, with limit science versus current science and so on. These are not all the same but they are motivated by similar distinctions and they all fall, albeit roughly, into the categories I have earlier mentioned of the 'in principle' versus the 'in practice'.

This is not to say that the concerns of traditional epistemology have no overlap with those of scientific epistemology. Evil demons and brains in a vat are not serious scientific hypotheses, but viewed in a wider context they are examples of how available empirical data can underdetermine our choice of hypotheses (see Stroud, 1984, Chapter VI) – here the rival hypothesis is the existence of the external world in the way our common sense thinks of it – and the underdetermination of theoretical hypotheses by empirical data is one of the dominant themes of twentieth century philosophy of science. Nevertheless, the concerns about knowledge that are of proper interest to a philosopher of science are usually of a different kind than are the concerns of a traditional epistemologist, such as: have unknown confounders been omitted from an epidemiological model so that claims to causal knowledge are thereby undermined?[6] Or what criteria are used to decide whether something counts as a datum in an experiment? (See e.g. Galison, 1987).

I shall thus focus here on those issues that are, at least prima facie, specific to the philosophy of science and replace the traditional a priori/a posteriori dichotomy by the divisions between a) knowledge that is theoretically based, b) knowledge that is drawn from observation, c) knowledge that is drawn from experiment, and d) knowledge that is based on models. Within c) there is a further division between the knowledge originating from laboratory experimentation and that gained from field experiments, whether randomized or not. These divisions have at best imprecise correlates in traditional epistemology.

2. KNOWLEDGE VIA THEORIES

For more than two millennia, at least as far back as Greek astronomical theories, scientists have viewed theoretical knowledge as ideally represented within some formal representational system. One particularly influential version of this approach has been the insistence that the theory must be presented as an axiomatized system, with a set of basic postulates from which all other claims of the theory can be derived with the aid of at most non-creative definitions of non-primitive terms. Such axiomatic systems differ in their degree of rigour, from the relatively informal systems of Newton's *Principia Mathematica* to extremely abstract axiomatizations such as von Neumann's axiomatization of quantum theory (von Neumann, 1955), modern measure-theoretic axiomatizations of probability (e.g. Loève, 1960), and the axiomatic development of theories of measurement (Krantz et al., 1971-90). Axiomatic approaches lend themselves naturally to foundational accounts, with the entire content of the theory being implicitly contained in its axioms. It has to be acknowledged that axiomatically presented theories are not common outside the basic physical sciences, that the axiomatizations are reconstructions of the more tractable representations used in practice by scientists, and that the kind of unity that an axiomatization gives to an area of scientific knowledge is thought by some to be inappropriate for areas such as biology or anthropology which are less amenable to reductivist approaches and which have significant historical aspects. (Although see e.g. Woodger, 1937, Williams, 1970, and Lloyd, 1988) for examples of formal treatments of some biological areas.)

Axiomatizations are ideally suited to hypothetico-deductive approaches, which have the advantage that prediction, explanation and confirmation can all be based on the deduction of consequences from more basic assumptions. What all axiomatic approaches have in common is the idea that scientific knowledge can be represented explicitly. The notion that the axioms should be self-evidently true has long ago been abandoned as an inappropriate constraint on empirical theories[7]. Instead what are considered to be the fundamental laws of the subject matter are to be used as the axioms or, more pragmatically, the principles that provide the most economical organization of the knowledge in the given area.

In the contemporary philosophical literature there has emerged a different way that axiomatic treatments are viewed. For those interested in limit science, axioms are often viewed as organizational devices: when all the evidence is in, those propositions that most efficiently allow the deduction of the remaining true propositions are designated as the axioms. Within this approach, the axioms are laws by fiat – they play a pragmatic role rather than one of representing fundamental truths about the subject matter. This view has been advocated by Ramsey (1931) and Lewis (1994, §3,4).

The older approach, which is the one that inevitably must be used in practice, chooses the axioms before all the knowledge in the area is available. Although, as mentioned above, the idea that the axioms of empirical theories must be self-evidently true has disappeared, there is still a strong sense that axioms must capture the core content of the subject matter. This may mean a representation of the fundamental laws of the subject, as with axioms for quantum field theory, or incontestable truths about the subject, as with classical probability theory. In neither case need one have the most efficient axiomatization. In fact efficiency is often at odds with transparency in that excessively compact axiomatizations frequently provide little understanding of the subject matter. (For examples, see Humphreys, 1993a) Such considerations are relevant to projects of unifying a subject area through axiomatization.

These two perspectives on axiomatization – retrospective and prospective – are thus likely to lead to different axiomatic presentations, even though both would be complete in the sense of being able to derive all truths about the subject matter from them.

Such axiomatic approaches, whether developed with the resources of logic or of mathematics, can be presented in one of two ways. The first, a syntactic axiomatization (also sometimes called a Hilbert style axiomatization), lays out the fundamental postulates of the theory in some precisely specified formal or formalized language with an explicit syntax. The theory may be identified either with the set of axioms or with the set of all logical consequences of those axioms. The 'in principle' approach has favoured the latter and has abstracted from alternative, logically equivalent, axiomatizations on the grounds that they are all axiomatizations of the same underlying theory. The important content of the theory is said to be captured by the set of its theorems (which trivially includes any axioms). While this is useful for certain purposes, such as proving the equivalence of the Heisenberg and Schrödinger representations of non-relativistic quantum theory, or the Lagrangian and Hamiltonian formulations of classical mechanics, there are significant differences in the way that different representations of the same theory can be applied to particular cases, especially where computational concerns are paramount, and the 'in practice' concern with how a given theory is applied shows that the particular representation used is indeed relevant. For example, the interaction representation is easier to use for certain S-matrix calculations of collision phenomena in quantum mechanics than is the Schrödinger representation, but the Heisenberg representation is preferable for certain problems with non-local interactions. (See Humphreys, 1995/6 for details and references.)

In the axiomatization, the non-logical parts of the language may or may not be interpreted. In many treatments the latter, more abstract, approach has been preferred and can result in otherwise distinct interpretations being associated with the same set of axioms, such as is often claimed to be the case with probability theory[8], for example. There are good reasons to doubt, however, whether relative frequencies, subjective degrees of rational belief, propensities, and logical probabilities do indeed have identical axiomatizations. Limiting relative frequencies are not additive, subjective degrees of probability are not plausibly countably additive, and propensities do not satisfy Bayes' Theorem.[9] Leaving the non-logical terms uninterpreted also has the serious drawback that there are multiple unintended interpretations of the same syntactic theory, and these purely syntactic objects are thus incapable of fully capturing the specific content of the theory.

A classic example of this kind of approach can be found in Carnap (1956), within which the argument is made that the choice of ontology for a theory is entirely pragmatic, and this line has been taken up by, amongst others, Quine (1969), Goodman (1978), and Putnam (1980). Such arguments can be traced back to various so-called basis theorems of early model theory, the best known of which are the Lowenheim-Skolem theorems. The latent Pythagoreanism inherent in such arguments (i.e. that we could, for all the theory says, choose as its interpretation a model the domain of which is the natural numbers) is best construed as an indication that the approach of stripping a theory of its original interpretation and reimposing a different semantics on the resulting formal structure is fundamentally defective. If a representational device cannot distinguish between cockroaches and natural numbers, this should be regarded as a serious reason to doubt the adequacy of the representational apparatus, not as an argument for anti-realism. Indeed it is the idea that formal representations are presented to us devoid of any antecedent interpretation that is at fault. It should be remembered that Tarski insisted that in the case of mathematical theories it was incoherent to consider an uninterpreted theory[10] and the point holds, perhaps with even more force, in the case of scientific theories. In gutting theories of their semantic content in order to provide a fully abstract representation, such purely syntactic axiomatizations are improperly representing the content of the original theory.

The dominance of logical reconstruction over much of this century has given such strategies more credence than they deserve: formal methods are undeniably useful for certain purposes, but it is an ineliminable aspect of extensional formal semantics that it is incapable of representing intrinsic properties, be they first or higher order. In standard set-theoretic semantics, all that matters is the numerical distinctness of the members of the domain; their intrinsic properties are irrelevant to the truth conditions for the theory. The true denizens of the domain of formal semantics are haecceities.[11]

The other principal mode of axiomatization is commonly known as the semantic or model-theoretic approach.[12] Here, rather than focussing on the syntax, the class of models that satisfy a syntactic axiomatization is identified with the theory[13]. This has the significant advantage of directly accounting for how the theory represents the part of the world to which it applies, for it is the existence of structure preserving maps, usually isomorphisms or homomorphisms, between the model-theoretic structures and (a substructure of) the real system that explains why the former represents the latter. To take a simple example, a partial ordering on a domain of rational numbers can represent the preference ordering possessed by an individual on commodity bundles because there is an isomorphism between the former and the latter. This is also the reason why quantitative measurement structures can be devised for systems that satisfy certain well defined structural conditions. (See e.g. Krantz et al., 1971-90). The semantic approach and its close cousin the structuralist approach has attracted many followers in recent years but without supplementation it also suffers from the defect that the content of a theory can be captured only up to isomorphism, a defect that is integral to the approach and one that has once again been used to argue that a realist interpretation of theories is impossible. In addition, what is claimed to be one of its chief virtues, that of abstracting from the linguistic representation, so as to eliminate inessential syntactic aspects of the axiomatization, means that it cannot directly deal with issues of the kind we discussed earlier where the syntactic form does make a difference to the possibility of applying the theory to specific systems.

3. Knowledge via Observation

Empiricism has always exerted an attraction on philosophers of science. Yet its traditional forms, whether in the versions espoused by the British empiricists Locke, Hume, or Mill, or the later forms advocated by the German, Austrian, and American logical empiricists, are curiously remote from the scientific enterprise. To take one example, the unaided senses, which for traditional empiricism play a central role in grounding evidential claims, have a small and diminishing role in scientific epistemology.

This emphasis on unaided sensory perception as the court of final epistemic appeal is appealing for empiricists because until recently, science has always been science carried out by humans. Yet we need to ask how assessments of scientific knowledge will change as the collection, evaluation, and presentation of data, hypothesis generation, experimental controls, and so on become automated and increasingly instrument dependent. In fact, if one looks at even commonplace data collection via the human sensory apparatus, it is seen to have a remarkably limited range of application. Consider temperature, for example. Why do we not use a human's ability to sense temperature as a reference point? The human body is at best a crude sensor of ordinal temperature scales and even inexpensive mercury thermometers do much better at providing a precise ordering relation on ambient temperatures of everyday objects. In fact, the human senses are involved only at these stages of calibrating mercury thermometers: a) identifying the fixed points of the scale (e.g. the triple point of an ice/water mix and the boiling point of water as well as identifying the substance involved as water), b) marking the fixed points and dividing the scale, c) observing the location of the mercury column on subsequent occasions (The last two can easily be automated). Although in many applications the identification involved in a) is made by reference to standard instruments, the instrumental epistemic regress is halted in this case by the fact that the fixed points involved are phase transition points which not only can be stipulatively determined but can, at least roughly, be directly observed by humans.

Of course background theories are involved as well – the regular linear expansion of mercury with temperature has to be established, for example – but for simple cases this can itself be empirically tested in a reasonably direct manner. Once humans have calibrated the instrument, the instrument itself becomes the standard and the need to appeal to human sensory abilities is abandoned, the output can be automated, and its reliability will be far higher even in this simple case than are human sensory abilities. In fact, once one begins to articulate the conditions that would make a human a reliable reference source for temperature, it is far from clear whether the resulting idealized circumstances are easier to achieve in the case of humans than they are for automated instruments. A human must not have been exposed to an ambient temperature significantly different from that to be detected (recall Berkeley's famous experiment in the First Dialogue between Hylas and Philonous involving the bowl of water whose temperature is estimated by a hot hand and a cold hand), the individual must not be suffering from a fever, the human must not yet have reached an age when temperature sensitivity is diminished, if the substance involved is air it must have a predetermined level of relative humidity, substances that are good insulators make better substances for human detection than do materials that rapidly conduct heat away from the skin, and so on. Even a common mercury thermometer is free from all of these defects.

The fact is that humans are reliable detectors of certain phenomena, such as whether two objects spatially coincide, and unreliable detectors of others, such as temperature. That is why many instruments present their output in a form that plays to humans' epistemic strengths, such as the coincidence of the end of a mercury column and a gradation on a glass tube, or the coincidence between a pointer and a mark on a dial, or a digital read-out. The moral to be drawn is that humans should be used as the ultimate reference source only for those data on which they are the most reliable detectors. This is a very small subset of the range of detectable phenomena and one can happily defer in the other cases to instruments the reliability of which is much greater than that of humans.[14]

For any empiricist, the division between empirical and non-empirical knowledge is crucial. The empiricist enterprise is driven primarily by the need for epistemic security and certain kinds of directly accessible empirical knowledge have been seen as possessing a high level of security. A great variety of candidates for grounding knowledge has been proposed, amongst which perhaps the best known are the psychologically oriented ideas of the seventeenth and eighteenth century British empiricists; the sense data accounts of G. E. Moore and Bertrand Russell (for whom sense data were objective, interpersonal entities); and the appeal to observational predicates and sentences by logical empiricists such as Carnap and Hempel. The great division within scientific languages for this last group was between observational and theoretical terms and a considerable amount of energy has been expended in trying to find criteria that would classify such terms in the correct manner. It has often been noted that this division between the observational and the theoretical is improper and should be replaced by two dichotomies; one between observable and non-observable entities, the other between theoretical and non-theoretical terms. This eminently sensible suggestion is ignored almost as often as it is made, but it can be quite misleading to argue on the basis of this incorrect classification.

The two most notorious difficulties blocking the formulation of criteria that distinguish between observable and non-observable entities are, first, the continuity of cases between the obviously observable and the uncontroversially unobservable, apparently making any division arbitrary and, secondly, arguments to the effect that there are no purely observational terms, that all putative observational vocabulary has some degree of theoretical content. The first kind of argument is provided by such cases as observing an insect with the naked eye, with a magnifying glass, with a low power optical microscope, with a high power optical microscope, with an electron microscope, and so on.[15] This argument is designed to establish two things; that drawing the line between the observable and the unobservable at the limit of unaided human sensory abilities is arbitrary, and that it does not correspond to any interesting ontological division. As it stands the argument is unsuccessful on the first count, for although the dividing line between the humanly observable and the humanly unobservable is not precise, it does not follow from that alone that it is arbitrary; indeed, the line is drawn at exactly the place where traditional uncertainty sets in. However, as we shall see shortly, a related argument can be given which shows that traditional empiricism is incorrect in universally drawing the line where it does if its primary concern is with epistemic security. The argument is more successful on the second count, because although the appeal to human observers may be convenient, it is clearly contingent.

Although this line of argument is frequently used in favour of realism in science, in the sense that if one is a realist about observable entities then there is no principled reason to withhold reality from entities 'observed' with sophisticated instruments[16], it can equally well be taken as showing that the division between the observable and the non-observable is (literally) an artificial one. The project of demarcating the observable from the unobservable is standardly taken as one that provides a permanent dividing line, one that is historically and technologically invariant, and the demarcation line is assumed to be assessed using a priori criteria. For scientific purposes this project is inappropriate. Scientific epistemology is not the subject of unaided human knowledge but of knowledge gained by humans with the aid of enormously sophisticated instruments that greatly enlarge our domain of access to the world. What counts as scientifically observable changes with scientific and technological progress: bacteria, viruses, macromolecules, stars of the tenth magnitude, distant galaxies and so on are now routinely observed although they were unobservable in the fifteenth century. That is, what counts as observable is a function of technological progress: the moons of Jupiter were unobservable for Copernicus, but any of us could have observed them on television sets displaying pictures transmitted by the Voyager I space probe in 1979.

4.DEMARCATION CRITERIA

Inherent in our treatment of scientific knowledge has been the assumption that there is something intrinsically different about the processes by which science produces knowledge. Thus far, the differences between scientific knowledge and everyday knowledge have been made clear. Now we need to consider how to demarcate scientific knowledge from claims to knowledge produced by pseudo-science and other suspect activities.[17] Much has been written on this topic and we cannot possibly treat the issue in its full complexity.[18] Nevertheless, although the debates about the demarcation issue demonstrated that separating scientific knowledge from pretenders to that title is not a simple matter, the difficulties and perhaps the impossibility of so doing have been exaggerated.

The three chief lines of objection to formulating a demarcation criterion are a) the conventionalist objection, b) the historical objection, and c) the analytic/synthetic objection. The conventionalist objection has formed the principal criticism of Karl Popper's famous falsificationist criterion for scientific status. Scientific knowledge for Popper is inherently fallible and results from the provisional acceptance of corroborated hypotheses – those hypotheses that have survived attempts to falsify them by comparing predictions drawn from the theory with empirical evidence statements. Critics of the falsificationist criterion have repeatedly pointed out something which Popper himself admitted from the earliest days of the falsificationist programme – that in principle, any given hypothesis can be protected against falsification because hypotheses are not tested in isolation. The hypothesis can be brought into contact with empirical data only by assuming the truth of a large number of background hypotheses of both a theoretical and an empirical kind. The correct working of an experimental apparatus, it is said, often has to be justified on the basis of a sophisticated theory; observation reports have to be taken as veridical; deductions of testable predictions from the hypothesis of interest require the use of other theoretical assumptions; and so on. The blame for a false prediction can thus be shifted to some other, dispensable, statement, and any given hypothesis can therefore be shielded from falsification. If a statement is scientific just in case it is falsifiable, and any given statement can be rendered unfalsifiable by the choice of appropriate conventions to shield it from falsification, then what counts as scientific and what does not is a result of adopting certain conventions.

The historical objection is often tacitly linked with the conventionalist objection. Case studies from contemporary science or from the history of science are used to establish the claim that as a matter of fact, some, perhaps many, eminent scientists have engaged in eminently unscientific practices. Newton was an enthusiastic alchemist, Kepler held that the dimensions of the planetary orbits were the result of conforming to the five perfect Platonic solids, Charles Darwin perhaps colluded in preventing Alfred Russel Wallace from receiving his share of the credit for the theory of natural selection, contemporary accusations of scientific fraud are brought or supported by jealous colleagues, and so on. The existence of such shady activities is not particularly damaging to efforts to establish a demarcation criterion. Descriptive studies can only undermine a normative criterion if the practices described are dominant and commonly agreed to be scientifically acceptable. There is no doubt that some things Newton did were weird – he was, even compared to other geniuses, something of an outlier. But his experiments with prisms were solidly scientific and millions of students have successfully replicated them. Kepler had to support himself as a court astrologer, but his fitting of data to the orbit of Mars was a model of careful scientific work. Who deserves credit for the theory of natural selection is a different issue than whether that central claim of evolutionary theory is true. It would be more damaging to the prospects of formulating a successful demarcation principle if such unscientific practices could not be separated from the legitimate scientific activities of these individuals. It is sometimes suggested that scientific fraud or subtly unscientific practices are standard in contemporary science, but the evidence does not support this idea and an obvious selection bias lies behind such claims.

The analytic/synthetic objection is most famously due to Quine, but its denial also lies behind some of the more important work of Kuhn.[19] If all definitions, which are traditionally viewed as analytically true, have some empirical content and can be abandoned when attitudes towards them change, and any apparently empirical statement can be converted, if so desired, into a statement that is true by stipulation, then the role of empirical testing in science becomes quite fluid and the difference in status between the conservation of energy principle and various definitions of energy, for example, is difficult to establish: 'My present suggestion is that it is nonsense, and the root of much nonsense, to speak of a linguistic component and a factual component in the truth of any individual statement'(Quine, 1951, p.42). Such claims have had a profound effect on the demarcation enterprise, not the least because the two dogmas of empiricism – the second being, roughly, a verificationist account of meaning – are, famously, claimed by Quine to be the same[20], and they are notoriously difficult to counter. Yet, despite the widespread acceptance of Quine's arguments over the past fifty years, there has always been segment of the philosophical community which has held that his arguments are not as destructive to the analytic/synthetic distinction as is commonly thought.[21] Kuhn's arguments, which are based on an historical study of principles used by Galileo, are more obviously flawed[22].

The attempts to undermine various demarcation criteria have served a useful purpose, and some morals can be drawn from the debates. Demarcation criteria typically try to combat two anti-scientific tendencies, dishonesty and sloppiness. The former comes in at least two varieties, failure to admit that one is wrong and outright fraud. Scientific methods are much better at detecting fraud and sloppiness than they are at forcing a determined truth-resister to concede error. It is revealing that the more serious of the three objections to formulating a satisfactory demarcation, the first and the third, rest on the possibility of these being unreasonably exploited by truth-resisters. The moral to be drawn here is that methods by themselves cannot ensure that scientific practices are being followed. There have to be in addition appropriate psychological attitudes possessed by those who use the methods. What do I mean by this? Consider the case of the legal system. It is commonly held that the primary reason for needing a system of laws to regulate human behaviour is the fact that humans are not always capable of behaving reasonably towards one another. Divorce, contract disputes, personal injuries, slander, wills; these and other conflicts tend to make humans behave badly. The hope is that imposing reasonable rules of procedure on antagonists will result in a correct resolution of the matter at hand. The democratic rule of law is clearly superior to other procedures but as anyone familiar with the practice of law knows, these rules can be manipulated by ill-intentioned practitioners, often with grotesque results.[23] And so it can be in science.[24] In this sense science is not fully self-correcting via its methods alone.

Such considerations require us to refine our criteria for what distinguishes scientific practice from other forms of human behaviour. (It is worth considering how science carried out by automata would differ from our current human practice.) This is not the place to provide a detailed account, but a few suggestions should make the point clear. Science comes in degrees – some activities are more scientific then others, and few are wholly and ideally scientific. The subject matter makes a difference and variations in the subject matter of the individual sciences produce differences in the effectiveness of scientific methods. For example, the replicability of experimental data is a central tool in detecting fraud and sloppiness, but it is far easier to replicate data in turbulence simulations than it is to replicate epidemiological studies. Multiple alternative explanations for a given data set tend to be easier to find in anthropology than in chemistry. Mere adherence to scientific procedures is insufficient to detect unscientific behaviour; intelligence, insight, and a critical attitude must all be present. These are not easily codified and they must be accompanied by knowledge acquired practising the science involved. One must also recognize that following scientific procedures does not guarantee arriving at the truth. Ptolemaic astronomers were respectably scientific, in fact astonishingly so for their time, but they did not have access to data which made it clear that their representations of the planetary system were false.

To supplement a codification of scientific practices with an appropriate scientific psychology is not easy, and the very enterprise will be viewed with deep suspicion by those who believe that eliminating psychological content from epistemology constitutes one of the great achievements of twentieth century science. Yet the prescriptive basis of scientific psychology ought to be congenial to those who seek a normative, rather than a descriptive, basis for scientific method. Its formulation will not resolve all the difficulties faced by demarcation criteria, but excluding it will make the objections outlined above insuperable.

5. KNOWLEDGE VIA EXPERIMENT

The role of experiment in producing scientific knowledge tends to be in two areas. The first is in producing knowledge about causes and the second is as the basis of arguments in favour of the existence of certain kinds of unobservable entities. It is the ability to manipulate factors in experimental situations rather than merely passively observing associations that lies behind the power of experiment to detect causal relations. Notoriously, the fact that two variables covary does not by itself provide good grounds for deciding that one is the cause of the other, most particularly because their covariance can be the result of variations in a third, unobserved, factor of which the observed variables are joint effects. By virtue of intervening so that changes in one of the two original variables, A, are forced, a subsequent lack of variation in the other variable, B, will rule out A as a direct cause of B. Without such intervention, one must eliminate alternative explanations for the covariance of A and B and this involves significant background knowledge.

An argument for scientific realism using experimental manipulations has been advanced by Ian Hacking (Hacking, 1983)[25]. Hacking's criterion asserts that if a putative entity can be used to manipulate other entities whose existence has already been established, one should accept the manipulator as real. The motivation for formulating this criterion is to avoid the dependence upon theoretical descriptions to pick out entities. (The distinction between so-called 'entity realism', within which one is committed to the existence of entities referred to within a theory, and so-called 'theoretical realism' within which one is also committed to the truth of various theoretical claims about those entities is important here. Hacking is an entity realist but not a theoretical realist.)

6. KNOWLEDGE VIA MODELS

Scientific knowledge has, for the majority of this century, been produced by the interplay of theory and observations. A more recent trend that promises a more realistic illumination adds models as intermediaries between abstract theory and concrete data. Scientific models[26] are constructions derived from theoretical assumptions augmented by idealizations and approximations. Abstract theory in most sciences consists of general schemata that are applicable to a wide variety of systems. The clearest examples of this can be found in the formal sciences: Newton's Second Law, Schrödinger's Equation in quantum mechanics, Poisson models of stochastic phenomena and so on. These are applied to concrete systems by specifying a specific model for that system, the model representing in certain respects salient features of the system being modelled. The model is arrived at by a) idealizing the real system by neglecting many of its features in order to restrict attention to the dominant ones b) approximating various quantities in the real system by, for example, considering a discrete quantity as continuous, or vice versa, truncating an infinite series by neglecting small orders of magnitude, considering the nucleus of an atom to be stationary rather than subject to small perturbations, and so on. These idealizations and approximations are frequently made in order to aid computational tractability but they may also be made simply as an aid in analogical reasoning.

Because these models falsify reality, sometimes dramatically, the question arises of how they can give us knowledge of the world. The answer to this is fairly straightforward and it illuminates the relation between the context of discovery and the context of justification. For the most part, the models are not adopted on an ad hoc basis, but are the result of a deliberate construction process within which many of the idealizations and approximations are explicitly known not to hold of the real system. Thus, when the model delivers incorrect predictions about the real system, it is usually already known which aspects of the model require adjustment in order to bring it into closer approximation with reality. The choice of how to adjust the model in the face of its failure to exactly fit the data is thus a process that is guided by criteria that are in place before the testing of the model even begins, and falsification of the model does not result in simply rejecting the model but in elaborating the model structure, tightening the approximations previously used, and so on.

6a.Kuhnian Paradigms

In marked contrast to the emphasis on objective knowledge discussed thus far are the views of Thomas Kuhn. These are well enough known not to need an extensive summary. Briefly, the focus of Kuhn's position is on the knowledge of a scientific community rather than on that of any given individual practitioner. Kuhn's position thus marked a transition from the logical to the social nature of knowledge, rather than the psychological, and it has had a significant influence on the development of the field known as social epistemology. The community's knowledge is incorporated into a paradigm or, as Kuhn later preferred, a disciplinary matrix, the components of which are (pace Kuhn, 1970, Postscript; Kuhn, 1974]) i) symbolic generalizations ii) metaphysical commitments, iii) values, iv) exemplars, and v) models. Kuhn's perspective on symbolic generalizations is quite different from the abstract account given by formalizable knowledge. For him, 'Though uninterpreted symbolic expressions are the common possession of the members of a scientific community, ..., it is not to the shared generalization that these tools are applied but to one or another special version' (Kuhn, 1974, p.465). It is this focus on particularity that is important, for the knowledge is contained not in the abstract 'laws', such as Newton's Second Law, or Schrödinger's Equation, but in knowing how to apply such schema to special systems. It is here that Kuhn's use of exemplars is particularly useful.

As Kuhn noted (1970, p.189) one of the most important skills acquired during the training period scientists undergo is the ability to first apply the basic principles of the science to stock examples and then to extend that ability to new applications that resemble the exemplars in key respects. Unlike the much better known, but defective thesis of incommensurability (see below), a thesis that unfortunately appealed in some versions to the use of exemplars within a behaviorist process of learning the meaning of the terms, the use of exemplars in applying theories is one of Kuhn's great positive insights and it anticipated some of the later work on models as the focus of much applied science.[27]

One of the most controversial areas of Kuhnian thought is the assertion of the incommensurability of different paradigms, the idea that either the standards of justification or the meanings of terms occurring within paradigm-dependent theories are so dependent on the paradigm that cross-paradigmatic comparison is impossible. This, if true, would at least make impossible the objective comparison or evaluation of claims to knowledge by various scientific paradigms. There is now good reason to think that the thesis of incommensurability took on an importance that was not warranted. As has been noted (see, for example, Laudan, 1984, McMullin, 1993), once the structure of a paradigm as a disciplinary matrix is made clear, it is historically inaccurate to claim that all the components of a paradigm are abandoned simultaneously in changing to a new paradigm. The retention of what is often the majority of a disciplinary matrix's components within a rival matrix provides sufficient overlap of either methods or of basic beliefs that objective standards of comparison can be applied. With resect to the supposed incommensurability of meaning, the rise of direct reference theories, which explicitly allow for a description (and hence theory) independent denotational process, avoids entirely the theory dependent meaning that in many versions of meaning incommensurability is the source of the trouble. (See Humphreys and Fetzer, 1998 for accounts of the development of such theories.) That is, a common argument for meaning incommensurability rests on the view that the theoretical structure within which a term is embedded provides an implicit definition for that term and hence different theoretical structures (i.e. disciplinary matrices) will bestow different meanings on their constituent terms, incommensurability being an inevitable result of this. The application of direct reference to avoid this problem will almost always be to natural kind terms rather than to individuals, and in some versions of the approach, appeal will be made once again to exemplars as stereotypical cases of the kind.

In the wake of Kuhn's work, there has been an increasing emphasis in recent years upon the role that non-rational procedures play in knowledge acquisition. This has two forms, one of which asserts that certain forms of inference engaged in by humans is irrational by traditional standards, the other of which attempts to show that certain sub-optimal methods subject to bounded rationality or heuristic methods are at work. There is a clear sense in which all model building involves such approaches. In deciding which variables to include in a multivariate regression analysis of economic data, for example, the so-called 'idealizations' are in fact epistemically just the opposite, for they involve moving away from an idealized optimal representation of the phenomena towards a simpler but more realistic approach to the material. Despite the success of such strategies, one should keep in mind that they are driven by an acknowledgement of the limitations of human cognitive abilities and the increasing use of computational aids pulls us in the opposite direction. There is less need to deal with simple models of phenomena when computationally assisted multivariate analysis is possible, for example.

6b.Non-Propositional Knowledge

Standard analyses of knowledge focus on a special kind of propositional attitude – x knows that p. There is a great deal of excellent literature on this (e.g. Bonjour, 1985) but it scarcely exhausts the kinds of knowledge of interest to scientists. Conceiving of knowledge as a propositional attitude is illuminating for some purposes, but the adherents of this view have often committed themselves to a quite extreme form of linguistic representationalism within which all knowledge must be represented within a precisely specified language. Valuable as this may be for certain purposes, it is too restrictive for scientific purposes because visual images are a key source of knowledge in many areas of science. Much scientific data is presented in graphical or pictorial form, and in many cases this is not simply from convenience but from epistemic necessity. The sheer quantity of data gathered from, for example, a far gamma ray telescope makes visual displays the only manageable way to present that data to an observer. In other areas, the use of directed graphs in sociology and economics to represent causal associations between variables constitutes a rich source of information for scientists. It is unreasonable to argue that in principle such sources could be coded into a language with an explicit syntax, for the fact is that propositional forms of representation are far less efficient sources for human perceptual mechanisms than are the kinds of non-propositional representations mentioned above and it is cognitively impossible for humans to assimilate millions of items of data presented in propositional form. The tension here between in principle ideals and in practice possibilities reflects the tension outlined earlier within empiricism. Science is, from the standpoint of traditional empiricism, science done by humans. Yet important human cognitive capacities, such as a facility for dealing efficiently with visual images, are often ignored by empiricists because of contemporary empiricism's emphasis on propositional entities, even though there is no evidence that humans actually process all information propositionally. The moral to be drawn once again is: the fact that transformations can be carried out in principle does not entail the epistemic equivalence of the data representations on either side of the equivalence. The mode of representation does matter. It has to be said as a caveat, however, that the move to put humans back at the centre of epistemology has been part of the naturalist movement but in so doing, it has once again made the mistake of focussing too much on human abilities and of excluding enhancers of those natural abilities.[28] There is now a growing body of literature concerning visualization in science, logic, and mathematics (see e.g. Wolff and Yeager, 1993, Shin, 1994) that promises to greatly enrich our understanding of how such knowledge is acquired and processed.

A second source of knowledge that is often claimed not to be amenable to representation in propositional form is what has variously been called physical (or biological, chemical, etc) intuition, tacit knowledge, or experimental know-how. Such knowledge is highly subject specific, comes from years of experience in the subject, is immediate rather then inferential (hence 'intuition'), is sometimes but not always knowledge how (so-called 'practical knowledge') and is characteristically supposed not to be formalizable (hence 'tacit'). A common area in which tacit knowledge is said to play a role is with ceteris paribus conditions attached to laws or to causal claims, with the attendant claim that such conditions are not finitely statable.

It is easy to overstate such claims. Much of the success of physics, for example, comes from its ability to find abstractly characterizable variables that include a myriad of conditions previously thought to be independent of one another. To adapt one of Bertrand Russell's examples, a vending machine's operations can be affected by conditions that include fat men jostling the machine, an earthquake rattling its works, the nearby detonation of an underground charge, a customer shaking it to retrieve his coins, and so on. Yet, contrary to Russell's belief, all of these conditions can be covered by descriptions of certain accelerative forces on the machine, these descriptions abstracting from the particularities of the specific cases.

Underlying these disputes is a question that is of central concern to the philosophy of science: how subject independent are the forms in which knowledge is represented to practitioners of science? The stock-in-trade logical and mathematical apparatus of twentieth century philosophers of science is predicated on the topic-neutrality of the representational apparatus but in order to justify the use of a particular representation in a given application one ordinarily needs to know some subject matter specific information.

6c. Reductive Knowledge

The axiomatic approach discussed earlier incorporates a special kind of reductive knowledge. In a wider sense, the issue arises about whether there is subject specific knowledge, or whether in some sense, our knowledge of non-physical entities, strictly construed, is illusory. It has long been maintained that there is a special mode of knowing that operates in the human sciences, one that goes under the various names of empathic understanding and Verstehen. Within this perspective, one is supposed to abandon the distanced, objective stance that one has towards the subject matter and, in the case of sociological or anthropological studies, to immerse oneself in the culture being studied so as to understand the subject matter from within. A contemporary version of this method is the so-called simulation studies in psychology, wherein one projects oneself into the position of the subject so as to imaginatively reconstruct how one would behave were one in the subject's position. This is to be contrasted with the more objective (and appallingly named) 'theory theory' approach within which a theory about the subject's mental states is constructed from an external perspective.

More generally, we can contrast various unity of science perspectives, within which there is claimed to be a single method for obtaining scientific knowledge, be it the kind espoused by the logical empiricists, falsificationists, Bayesians, Baconian inductivists, or other universalist approaches, and various disunity claims suggesting that each science has its own ways of gaining knowledge. Trivially, of course, there is no denying that chemistry requires different experimental methods than do elementary particle physics and population biology; that experimental economics engages in different methods than does archaeology. The unification theses are pitched at a much higher level of abstraction than this, however: the origin of the data is simply taken as a given and there is then claimed to be a universal way of evaluating the data as a contribution to knowledge (which is ordinarily taken to be propositional in form). Despite this, there is some clear sense in which the subject matter influences how we acquire knowledge of it.

Furthermore, there is good reason to suppose that subject specific knowledge is required for model construction and evaluation. For example, if we were asked to evaluate an epidemiological model of how levels of various airborne pollutants affect morbidity and mortality levels, a considerable amount of meteorological knowledge is required to decide how representative are the levels at the measurement site of average levels in a given area; of biological knowledge to decide how the various pollutants are affected by various respiratory processes; of social knowledge to know how much time individuals in the area spend outdoors absorbing the ambient levels rather than indoor levels; of miscellaneous knowledge such as levels of smoking in the region, and so on. To take a different example, to construct and modify a spin flips dynamics Ising model of ferromagnetism, the plausibility of various idealizations must be assessed on physical grounds when deciding either how realistically to interpret the model or how best to adjust the model to obtain a better fit to data.

It is a significant defect of purely logical approaches to scientific theories that no criteria are given for theory adjustment beyond vaguely pragmatic criteria such as overall simplicity or maintaining progress. Informed subject specific knowledge plays a central role in decisions of how to improve theories because it is usually known in advance just how the modelling procedure has misrepresented the system being modelled. It is therefore often not surprising when the model fails to fits the data exactly and the adjustment procedure is not based on simplicity grounds, but on contentful criteria.

A second area within which the issue of reductive knowledge is important is in the area of interlevel reduction – whether chemistry can be reduced to physics, biology to chemistry, and so on. The autonomy of certain sciences has been argued for by various anti-reductionists on a variety of grounds. It has been maintained that psychology cannot be reduced to biology because of the inherently intentional natures of some psychological states, or that the natural kinds employed in psychological laws (if such exist) do not naturally match those employed in biological laws. Conversely, if chemistry is reducible to physics, then in what sense is chemical knowledge distinct from physical knowledge? On the deductive-reductive view, it is in no way different, because the ability to deduce various theoretical claims of chemistry about energy levels in molecules from quantum theory renders the situation no different than one in which we are deriving theorems of physics from an axiomatically formulated theory. In practice, these ab initio calculations are often impossible to carry out and this again leaves certain kinds of specifically chemical knowledge distinct from physical knowledge.

It has been said (e.g. Sarkar [1998]) that reduction consists in explaining phenomena at the higher levels in terms of phenomena at lower levels. This is different from ontological reduction, within which entities at the higher levels are shown to be 'nothing but' complexes composed of entities from lower levels. With explanatory reduction, we know why the higher level states behave as they do in terms of the properties of lower level states. One can see the essentially epistemic orientation of this view by comparing it to similar explanatory understanding gained through using earlier states of a system to explain later states (for simplicity assume here that we are dealing with a deterministic system.) In the latter, there is no temptation to construe the later states as 'nothing but' the earlier states and it is possible to construe the explanatory reduction as having provided explanatory knowledge.[29]

Paul Humphreys
University of Virginia

NOTES

[1] Augmented perhaps by pragmatism.

[2] The division could not be made on the grounds of whether the justification was internal or external, for the justification for knowing that one has a headache is internal (and even mental) but is a posteriori for all that.

[3] Here a decision about the appropriate boundary between the instrument and its environment is obviously necessary.

[4] Rather than in principle, which is true but quite different – it is provable that every holomorphic function can be approximated to an arbitrarily high degree of accuracy by a linear combination of epicyclic motions. Ellipses can be represented by holomorphic functions, so the motions of planets described by modern Keplerian version of Copernicus's theory can be simulated by Ptolemaic devices.

[5] See Glymour (1992) for a candid evaluation of some of these efforts.

[6] See Rosenbaum and Rubin (1983) for an examination of such issues.

[7] It is by no means obvious that it should be imposed on mathematical theories either.

[8] This claim is made by Carnap (1962) amongst others.

[9] See e.g. van Fraassen (1980), Suppes (1970), de Finetti (1972), Humphreys (1984).

[10] 'It remains perhaps to add that we are not interested here in 'formal' languages and sciences in one special sense of the word 'formal', namely sciences to the signs and expressions of which no material sense is attached. For such sciences the problem here discussed has no relevance, it is not even meaningful.' (Tarski, 1956, p.166).

[11] And possibly not even those in the case of indistinguishable particles such as electrons.

[12] For the original source of this see Suppes (1970), which is unfortunately still unpublished. A more recent approach that uses the semantic view is van Fraassen (1980).

[13] This is not to say that such a syntactic axiomatization need exist. Many semantic approaches directly define the class of appropriate structures without detouring through a syntactic theory.

[14] For a detailed account of the role played by calibrating instruments in physics, see Franklin (1986).

[15] The classic source of this argument is Maxwell (1962)

[16] For arguments pro and con along these lines see Shapere (1982), Hacking (1983), and van Fraassen (1980).

[17] This is not the same issue as demarcating scientific knowledge from non-scientific knowledge, at least because a great deal of non-scientific knowledge requires the same critical attitude towards the supposed evidence for its justification as does scientific knowledge.

[18] For classic sources see, Popper (1959), Kuhn (1970), and Lakatos and Musgrave (1970). For more recent assessments, see Horwich (1993), Miller (1994).

[19] See Quine (1951); Kuhn (1964).

[20] Equally famously, Popper denied that falsificationism was about meaning, but we are not solely concerned here with Popper's attempts at demarcation.

[21]For an assessment of the current status of the debate, see e.g. Boghossian (1997) and Hintikka (unpublished).

[22] See Cargile (1987), Humphreys (1993b)

[23] I am speaking here primarily of the adversarial system used in British and American courts and not of the truth-oriented system of the French courts, although the latter is also, of course, open to abuse.

[24] Of course there are significant differences between science and the law: there is no statute of limitations in science, it is far more difficult to exclude intelligent participants from the equivalent of juries, there is no final court of appeal, there are differences in personality traits between lawyers and scientists, and so on.

[25] David Miller has pointed out to me that the essence of the criterion goes back at least to Alfred Landé. As cited in Popper (1982), p.46: '...by and large I regard as excellent Landé's suggestions that we call physically real whit is 'kickable' (and able to kick back if kicked).'

[26] Not to be confused with the objects of study in model theory in logic.

[27] This is not to say that Kuhn was alone in this. Both Hesse (1953) and Campbell (1920) had earlier suggested that we pay more attention to the use of models in science. Contemporary treatments of modelling include Wimsatt (forthcoming), Hartmann (1995), and Humphreys (1991).

[28] I note that the term 'naturalism' has become too diffuse to capture all of the ways in which it is applied. No doubt there are advocates of naturalized epistemology who would allow the extended sense I have suggested here.

[29] I am grateful to David Miller for drawing my attention to some errors in an earlier draft.

REFERENCES

Almeder, R.: (1990), 'On Naturalizing Epistemology', *American Philosophical Quarterly* **27**, 263-81.

Boghossian, P.: (1997), 'Analyticity', in B. Hale and C. Wright, *A Companion to the Philosophy of Language*, Blackwell Publishers, Oxford, pp. 331-368.

Bonjour, L.: (1985), *The Structure of Empirical Knowledge*, Harvard University Press, Oxford.

Campbell, N. R.: (1920), *Physics, The Elements*, Cambridge University Press, Cambridge.

Cargile, J.: (1987), 'Definitions and Counterexamples', *Philosophy* **62**, 179-93.

Carnap, R.: (1956), 'Empiricism, semantics, and ontology', in R. Carnap, *Meaning and Necessity* (2nd Ed.), University of Chicago Press, Chicago, pp. 205-21.

Carnap, R.: (1962), *The Logical Foundations of Probability*, University of Chicago Press, Chicago.

Churchland, P. M. and P. S. Churchland: (1998), *On the Contrary*, MIT Press, Cambridge Mass.

de Finetti, B.: (1972), *Probability, Induction and Statistics*, J. Wiley and Sons, New York.

Franklin, A.: (1986), *The Neglect of Experiment,* Cambridge University Press, Cambridge.

Galison, P.: (1987), *How Experiments End,* University of Chicago Press, Chicago.

Gigerenzer, G. and P. Todd: (1999), *Simple Heuristics that Make Us Smart*, Oxford University Press, Oxford.

Glymour, C. G.: (1992), 'Invasion of the Mind Snatchers', in R. Giere (ed.), *Cognitive Models of Science, Minnesota Studies in the Philosophy of Science, Vol XV*, University of Minnesota Press, Minneapolis, pp. 465-471.

Goodman, N.: (1978), *Ways of Worldmaking*, Hackett Publishing Company, Indianapolis.

Hacking, I.: (1983), *Representing and Intervening,* Cambridge University Press, Cambridge.

Hartmann, S.: (1995), 'The World As a Process', in R. Hegselmann (ed.), *Simulation and Modeling in the Social Sciences from the Philosophy of Science Point of View*, Kluwer Academic Publishers, Dordrecht.

Hesse, M.: (1953), 'Models in Physics', *British Journal for the Philosophy of Science* **4**, 198 ff.

Hesse, M.: (1966), *Models and Analogies in Science*, University of Notre Dame Press, Notre Dame.

Hill, A. B.: (1965), 'The Environment and Disease: Association or Causation?' *Proceedings of the Royal Society of Medicine* **58**, 295-300.

Hintikka, J. (unpublished): 'A Distinction Too Few or Too Many: A Vindication of the Analytic/Synthetic Distinction'

Hintikka, J.: (1962), *Knowledge and Belief*, Cornell University Press, Ithaca.

Horwich, P. (ed.): (1993), *World Changes*, MIT Press, Cambridge, Mass.

Humphreys, P.: (1984), 'Why Propensities Cannot Be Probabilities', *Philosophical Review*, **94**, 557-570.

Humphreys, P.: (1991), 'Computer Simulations', in A. Fine, M. Forbes, and L. Wessels (eds.), *PSA 1990, Volume 2*, Philosophy of Science Association, East Lansing, pp. 497-506.

Humphreys, P.: (1993a): 'Greater Unification Equals Greater Understanding?', *Analysis* **53**, 183-188.

Humphreys, P.: (1993b), 'Seven Theses on Thought Experiments', in J. Earman, A. Janis, G. Massey, and N. Rescher (eds.), *Philosophical Problems of the Internal and External Worlds*, University of Pittsburgh Press, Pittsburgh, pp. 205-227.

Humphreys, P.: (1995/6), 'Computational Empiricism', *Foundations of Science* **1**, 119-130.

Humphreys, P. and J. Fetzer (eds.): (1998), *The New Theory of Reference*, Kluwer Academic Publishers, Dordrecht.

Kahneman, D., P. Slovic, and A. Tversky: (1982), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, Mass.

Kitcher, P.: (1992), 'The Naturalists Return', *Philosophical Review* **101**, 53-114.

Krantz, D., R. Luce, P. Suppes, and A. Tversky: (1971-90), *Foundations of Measurement*, Volumes 1, 2, 3, Academic Press, San Diego.

Kripke, S.: (1980), *Naming and Necessity*, Harvard University Press, Cambridge, Mass.

Kuhn, T.: (1964), 'A Function for Thought Experiments', reprinted in T. Kuhn, *The Essential Tension*. University of Chicago Press, Chicago, 1977.

Kuhn, T.: (1970), *The Structure of Scientific Revolutions* (2nd Edition), University of Chicago Press, Chicago.

Kuhn, T.: (1974), 'Second Thoughts on Paradigms', in F. Suppe (ed.), *The Structure of Scientific Theories*, University of Illinois Press, Urbana, pp. 459-482.

Lakatos, I. and A. Musgrave (eds.): (1970) *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge.

Laudan, L.: (1984), *Science and Values*, University of California Press, Berkeley.

Lewis, D.: (1994), 'Humean Supervenience Debugged', *Mind* **103**, 473-90.

Lloyd, E.: (1988), *The Structure and Confirmation of Evolutionary Theory*, Greenwood Press, New York.

Loève, M.: (1960), *Probability Theory,* (2nd Edition), van Nostrand, Princeton.

Maxwell, G.: (1962), 'The Ontological Status of Theoretical Entities', in H. Feigl and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Scence. Volume 3*, University of Minnesota Press, Minneapolis, pp. 3-15.

McMullin, E.: (1993), 'Rationality and Paradigm Change in Science', in P. Horwich (ed.), *World Changes: Thomas Kuhn and the Nature of Science*, MIT Press, Cambridge, Mass. pp. 55-78.

Miller, D.: (1994), *Critical Rationalism,* Open Court, La Salle.

Popper, K.: (1959), *The Logic of Scientific Discovery*, Basic Books, New York.

Popper, K.: (1982), *Quantum Theory and the Schism in Physics,* Rowman and Littlefield, Totowa, N.J.

Putnam, H.: (1980), 'Models and Reality', *Journal of Symbolic Logic* **45**, 464-82.

Quine, W. V. O.: (1951), 'Two Dogmas of Empiricism', *Philosophical Review* **60**, 20-43, reprinted in W. V. O. Quine, *From a Logical Point of View,* Harvard University Press, Cambridge, Mass.,1953.

Quine, W. V. O.: (1958), 'Speaking of Objects', *Proceedings and Addresses of American Philosophical Association* **31**, 5-22.

Quine, W. V. O.: (1960), *Word and Object*, MIT Press, Cambridge, Mass.

Quine, W. V. O.: (1969), 'Ontological Relativity', in his *Ontological Relativity and Other Essays*, Columbia University Press, New York, pp. 26-68

Ramsey, F. P.: (1931), *The Foundations of Mathematics and Other Essays*, Open Court, La Salle.

Rosenbaum, P. and D. Rubin: (1983), 'The Central Role of the Propensity Score in Observational Studies for Causal Effects', *Biometrika* **70**, 41-55.

Sarkar, S.: (1998), *Genetics and Reductionism*, Cambridge University Press, Cambridge.

Shapere, D.: (1982), 'The Concept of Observation in Science and Philosophy', *Philosophy of Science* **49**, 231-267.

Shin, S.-J.: (1994), *The Logical Status of Diagrams*, Cambridge University Press, Cambridge.

Stroud, B.: (1984), *The Significance of Philosophical Scepticism*, The Clarendon Press, Oxford.

Suppes, P.: (1970), *Set-Theoretical Structures in Science,* mimeod manuscript, Institute for Mathematical Studies in the Social Sciences, Stanford University.

Tarski, A.: (1956), 'The Concept of Truth in Formalized Languages', in his *Logic, Semantics, Metamathematics*, The Clarendon Press, Oxford, pp. 152-278.

van Fraassen, B.: (1980), *The Scientific Image*, The Clarendon Press, Oxford.

von Neumann, J.: (1955), *Mathematical Foundations of Quantum Mechanics*, Princeton University Press, Princeton.

Williams, M.: (1970), 'Deducing the Consequences of Evolution: A Mathematical Model', *Journal of Theoretical Biology* **29**, 343-385.

Wimsatt, W.: (forthcoming), *Piecewise Approximations to Reality*, Harvard University Press, Cambridge.

Wolff, R. and L. Yaeger: (1993), *Visualization of Natural Phenomena*, Springer-Verlag, New York.

Woodger, J. H.: (1937), *The Axiomatic Method in Biology,* Cambridge University Press, Cambridge.

ROMAN MURAWSKI

MATHEMATICAL KNOWLEDGE*

Since its very beginnings mathematics played a special and distinguished role in the human knowledge. It was close to the ideal of a scientific theory, even more, it established such an ideal and served as a pattern of a theory. It has played an important role also in the development of the epistemology. In fact mathematics has been through ages a pattern of any rational knowledge and the paradigm of a priori knowledge. Hence the importance and meaning of philosophical and methodological reflections on mathematics as a science. Such reflections have accompanied mathematics since ancient Greece.

In philosophical reflections on mathematics one can distinguish two principal groups of problems: ontological and epistemological. Among main questions of the first group are the following ones: what is the subject of mathematics, in particular what is the nature of mathematical objects, where and how do they exist, what are the criteria of their existence, what is the source and origin of mathematical objects, what is the nature and properties of the mathematical infinity.

Epistemological problems concerning mathematics (which are the main subject of the present article) can be divided into four groups:

- the problem of cognitive methods used and accepted in mathematics. In particular one considers here the problem of sources and origin of mathematical knowledge, the problem of the process of arriving at new results, the problem of methods of justifying mathematical statements and theorems, the problem of the validity of such methods, the problem of the nature of mathematical proofs, of criteria of distinguishing correct and incorrect proofs, of the justification of the axiomatic-deductive method and of the range of its applicability as well as of the axioms and of their origin, problems of decidability and the question whether there are any (and what) limits or bounds of mathematical knowledge. Here belong also the problems whether deduction is the only legitimate method in mathematics or should it be combined with induction and generally with empirical methods? Or is the method of proofs and refutations the proper method of establishing new results? What is the role of intuition in mathematical knowledge? Should nonconstructive methods be allowed in mathematics or one should restrict mathematics to constructive methods only?
- the problem of the type of mathematical knowledge. One asks here in particular whether mathematical knowledge is a priori or an empirical knowledge, whether mathematical theorems are analytic or synthetic statements, what is the value of mathematical statements, does a mathematician discover or create mathematical reality and its properties, and

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 571-606.

© 2004 *Kluwer Academic Publishers.*

consequently mathematical knowledge. If one has to do with discovering in mathematics then which methods can be used here, and similarly in the case of the alternative answer. One also considers here the problem of the relations between pure and applied mathematics, in particular the fundamental question why abstract mathematical theorems can be applied to the description of physical phenomena of the external world.

- the problem of a systematization of mathematical knowledge, and in particular the problem of the unification of mathematics,
- the problem of the dynamics and the development of mathematics as well as the problem of the place and the role of mathematics in the whole (system) of culture and especially in relation to other domains of human scientific knowledge.

Note that this list of problems and questions is not complete and particular items of it overlap each other.

Both aspects of the philosophical reflection on mathematical knowledge distinguished above ontological and epistemological are interconnected. Answers to some questions induce and imply (or even force) solutions to other problems. Nevertheless – to accord with the subject of this article – we shall concentrate here on epistemological problems (being conscious the whole time of the fact that one cannot escape some ontological solutions and decisions).

Philosophy of mathematics and in particular the epistemology of mathematics are of course connected with other branches of philosophy and with mathematics itself. The development of mathematics, the development and changes of the subject of study and of methods of mathematics lead to the development of the philosophical reflection on mathematics and to the change and revision of previous doctrines. As an example, perhaps the most striking one, can serve the introduction and development of the non-Euclidean geometries in the nineteenth century. On the other hand new mathematical problems induce new philosophical questions and problems (for example the recent use of computers in proving mathematical theorems). Hence the importance of the history of mathematics to the philosophy of mathematics (one can even say, paraphrasing I. Kant, that the history of mathematics without philosophy is blind and the philosophy of mathematics without history is empty). Also the development of logic, especially of the mathematical logic at the turn of the nineteenth century, as well as of the mathematical studies of mathematics as a science (metamathematics) played a great role for the epistemology of mathematics making possible the precise formulation of various problems and notions (such as proof, truth, consistency) as well as their solution (indicating for example some limitations and bounds of the axiomatic-deductive method) – cf. Gödel's incompleteness theorems, Löwenheim-Skolem theorems or Tarski's theorem on the undefinability of truth).

Philosophy (and in particular the epistemology) of mathematics plays a double role with respect to mathematics: on the one hand it describes and codifies the methods actually used in mathematics (one should distinguish here of course between the context of discovery and the context of justification) and on the other it plays a normative role establishing and justifying the legitimate and correct methods of mathematics.

From various possible ways of presenting the main doctrines in the epistemology of mathematics we have chosen the historical one, because, as L. Kolakowski wrote, "All that is really important in the philosophy, is being discovered by learning its history; great philosophers sensibilize us to the plurality of perspectives from which the world can be considered as well as to the plurality of languages in which it can be described".[1] Hence the article is organized in the following way: At the beginning the predecessors of the contemporary doctrines are presented. Next the main modern conceptions in the epistemology of mathematics (connected with logicism, intuitionism and formalism) are considered. Finally recent trends in the philosophy of mathematical knowledge are described.

1 PREDECESSORS OF THE CONTEMPORARY DOCTRINES

The real reflection on mathematics as a science began by Plato (427-347 B.C.). His philosophy of mathematics grew out of his theory of ideas. He claimed that the subject of mathematics are mathematical (arithmetical and geometrical) ideas (or forms).[2] They are real entities conceived as being independent of perception and being apprehended, as being capable of absolutely precise definition and as being absolutely permanent, that is to say timeless or eternal. Hence a mathematician does not create mathematical objects and their properties but does discover them. Consequently the mathematical knowledge is based on the reason and the proper method of mathematics is the axiomatic method – Plato was probably the first who introduced it. Mathematics is very close to Plato's ideal of knowledge because it abstracts from changeable phenomena and concentrates on unchangeable, timeless, mind-independent and definite objects and relations between them. Plato admitted that a mathematician uses in his research practice observations and drawings or perform constructions but they serve only the process of remembering the proper mathematical objects (ideas) and not the creation of them (Plato refers here to his theory of *anamnesis*). Hence mathematics is a science whose aim is the description of timeless, mind-independent and definite mathematical objects (ideas) and their mutual relations. Consequently all mathematical propositions are necessarily true. Their necessity is independent of their being apprehended by a mathematician, independent of any formulation and thus of any rules governing a natural or artificial language. Mathematical theorems can be applied to the description of the objects of sense-experience because the latter are to a certain degree similar to, or better, approximate the ideas (Plato says here that, for example, one apple participates in the arithmetical idea One).

Aristotle (384-322 B.C.), the disciple of Plato, developed his philosophy of mathematics partly in opposition to that of Plato and partly independently of it. He rejected Plato's theory of ideas claiming that mathematical objects are forms of things, are idealizations obtained by the process of abstraction. Hence they do not exist timelessly and independently of things but are in a sense in things. Consequently mathematical propositions as being only idealizations cannot be necessarily true. The necessity cannot be found in any single statement about mathematical objects but in hypothetical statements saying that if a certain proposition is true then a certain other proposition is also true. Hence using today's terminology we can say that for Aristotle the necessity of mathematics was that of logically necessary hypothetical propositions. Aristotle paid much more attention to the structure of whole theories in mathematics than to isolated propositions (cf. *Physics* II, 9, 200^{a}, 15-19; *Metaphysics* 1051^{a}, 24-26). According to him the base of any knowledge is formed by general notions which do not need to be defined and by general propositions which do not need to be proved. All other notions should be defined and all other statements should be proved. He distinguished in any theory four basic components (cf. *The Posterior Analytics* I, 10, 76^{a}, 44-77^{a} 3): (1) the principles which are common to all sciences (Aristotle called them axioms, they correspond to logical axioms and axioms of identity in today's terminology), (2) the specific principles which are taken for granted by the mathematician engaged in the demonstration of theorems (Aristotle called them postulates, they correspond to non-logical axioms in the terminology of today's formal logic), (3) definitions (add that Aristotle did not assume that what is defined exists) and (4) existential hypotheses assuming that what has been defined exists independently of our perception and thought (note that according to Aristotle such hypotheses seem not to be required for pure mathematics). Aristotle introduced also the distinction between potential and actual infinity (cf. *Physics*, Book III). He claimed that the potential infinity suffices in mathematics and the actual one is in fact superfluous. It is worth adding that Aristotle saw in mathematics also some aesthetic elements, even more, he claimed that they play an important role in the development of mathematical knowledge. In fact mathematics says, though not explicitly, about the beauty and reveals some of its elements (cf. *Metaphysics* 1078^{b}, 52-1078^{b}, 4). Note that similar ideas can be found also by Proclus, a neoplatonic philosopher living in the fifth century, or by Henri Poincaré, French mathematician and philosopher living in the nineteenth century.

Plato's philosophy of mathematics and Aristotle's ideas concerning the structure of a scientific theory and in particular of a mathematical theory found their deepest application and realization in *Elements* by Euclid (365 (?)-300 (?) B.C.). In fact this work established a paradigm in mathematics prevailing up until the end of the nineteenth century called today Euclidean paradigm.

The *Elements* were on the one hand the presentation of results obtained by Greek mathematicians in the last 300 years before Euclid and on the other it gave a firm basis for the future development of mathematics. They consisted of 13 books: books I-IV were devoted to the plane geometry, book V to Eudoxus' theory of proportions in its purely geometrical form, book VI to the similarity of plane figures, books VII-IX to arithmetic (the ancient number theory), book X to incommensurable magnitudes and books XI-XIII to solid geometry. Every book began by defintions of new notions and by a list of axioms and postulates (one can see here the influence of Aristotle). Note that the *Elements* contained no list of primitive undefined terms, but, on the contrary, Euclid attempted to define all the terms he used (eventually those "definitions" were rather explanations of notions than proper definitions in the strict sense). It was possible because he, as Aristotle, did not distinguish between the language of the considered theory and the colloquial language – in fact the language of a theory was not separated from the natural language.

The postulates, axioms and definitions supplied the starting point for Euclid's proofs. His aim was to prove all principles by showing that they follow necessarily from the basic assumptions. In this way he wanted on the one hand to strengthen the mathematical knowledge by increasing the rigor with which already known laws could be proved and on the other to extend this knowledge by proving new and hitherto unknown laws. He wanted to organize mathematics (first of all geometry) in a systematic deductive form. The exact analysis of Euclid's proofs however indicates that there were certain gaps in them.

Nevertheless the *Elements* established a pattern of a scientific theory and in particular a paradigm in mathematics. Since Euclid till the end of the nineteenth century mathematics was developed as an axiomatic (in fact rather a quasi-axiomatic) theory based on axioms and postulates. Proofs of theorems contained several gaps – in fact the lists of axioms and postulates were not complete, one freely used in proofs various "obvious" truths or refered to the intuition. Consequently proofs were only partially based on axioms and postulates. Almost no attention was paid to the precization and specification of the language of theories – in fact the language of the theories was simply the unprecise colloquial language.

Add that the Euclid's approach (connected with Platonic idealism) to the problem of the development of mathematics and the justification of its statements (which found its fulfilment in the Euclidean paradigm), i.e. justification by deduction (by proofs) from explicitly stated axioms and postulates, was not the only approach and method which was used in the ancient Greek (and later). The other one (call it heuristic) was connected with Democritean materialism. It was applied for example by Archimedes who used not only deduction but any methods, such as intuition or even experiments (not only mental ones), to solve problems. Though the Euclidean approach won and dominated in the history one should note that it formed rather an ideal and not the real scientific practice of mathematicians. In fact rigorous, deductive mathematics was a rather rare phenomenon. On the contrary, intuition and heuristic reasoning were the animating forces of mathematical research practice. The vigorous but rarely rigorous mathematical activity produced "crises" (for example the Pythagoreans' discovery of the incommensurability of the diagonal side of a square, Leibniz's and Newton's problems with the explanation of the nature of infinitesimals, Fourier's "proof" that any function is representable in a Fourier series, antinomies connected with Cantors imprecise and intuitive notion of a set).

There was a significant problem connected with the axiomatic-deductive method, namely the problem of the choice of postulates and axioms. For Plato they were simply necessary truth. Aristotle spoke as though he felt that every science had its own definite principles (which should function as postulates) and its own definite primitive terms (for every definite term there was just one correct way of defining it). Euclid expressed no opinion on such questions. Proclus (410-485), the neoplatonic author of *Commentary to the First Book of Euclid's "Elements"* claimed that the common feature of axioms and postulates is the fact that they need no justification or proof, they can be accepted as known. The difference between them is similar to that between theorems and problems. Axioms contain facts which are immediately obvious and do not make any trouble for our thought; in postulates on the other hand one tries to find facts and properties which can be easily established and by which no sophisticated procedures or constructions are needed.

He accepted Plato's ideas of the origin of mathematical notions and said that they have their source in the soul which contains their patterns. This induces also the method of mathematics – in fact the proper method of mathematics is not intuition but the discursive method consisting of deduction from the premisses.

It is worth noting here (anticipating the development of the events in the epistemology of mathematics and in the mathematics itself) that in fact till the end of the nineteenth century mathematicians were convinced that axioms and postulates should be simply true statements, hence sentences describing the real state of affairs in the mathematical reality. Only the development of non-Euclidean geometries in the nineteenth century called the attention to the possibility that this is not necessary, that one can develop theories based on any consistent set of axioms. But the way to the full consciousness of this was long and not direct.

Middle Ages did not bring new important ideas to the philosophy of mathematics. The views and theses of Plato, Aristotle and Euclid were developed and commented and mathematics was developed along the lines established by the *Elements.* Only in the seventeenth century some new ideas appeared. The intensive development of natural sciences and of mathematics brought new problems which should be solved. One looked also for some principles which would unify the whole edifice of human scientific knowledge.

As a founder of modern philosophy one usually considers René Descartes (1596-1650). He was the first philosopher whose outlook was profoundly affected by the new physics and astronomy. He was a philosopher and a mathematician. As a mathematician he is known first of all as the inventor of the analytic (coordinate) geometry (though not quite in its final form) – it was in fact based on the application of algebra to geometry and contributed very much to the unification of those two branches of mathematics which since antiquity were developed as two separate parts of mathematics. Descartes contributed also significantly to the methodology of mathematics. To explain it one should start from his principle claiming that all things that we conceive very clearly and very distinctly are true. Hence the criterion of certainty in science is based on clear and distinct ideas. Descartes proclaimed a programme of universal rational knowledge build along the principles similar to those of mathematics. According to him only mathematicians are constructing proofs and therefore only mathematics provides an unfailing and secure knowledge. It has its sources in the fact that only quantitative properties are considered in mathematics. Hence Descartes' idea of bounding every scientific theory to such considerations and his idea of creating a universal analytical and mathematical theory called by him *mathesis universalis.* In mathematics itself – being the pattern of any other science – only analytic methods should be applied. Descartes allowed in it only intuition and deduction. Axioms of mathematics were for him unfallible and indubitable truths. The analytic method should enable us to discover the simple components of thoughts. And this what was simple, was for Descartes clear and distinct, hence certain. In *Discours de la méthode* (1637) he enumerates some rules which suffice in every scientific theory. According to them one should not accept any statement which is not clear and distinct, one should apply the analytic method and "decompose" any problem into so many components that are enough to find a solution and finally to deduce more complex truths from simple truths, i.e., from axioms.

Blaise Pascal (1623-1662), French philosopher and mathematician one generation younger than Descartes, was not so "bewitched" by the power of reason. He distinguished two parts, two realms: the realm of the reason and that of a heart. Reason cannot help us to solve existential problems. Descartes' clear and distinct ideas are of no help here. He wrote: *"Le coeur a ses raisons, que la raison ne connait pas"*. There is of course a question what should be understood under "le coeur"? Various interpretations have been provided. One of them identifies it with the human ability to know the supernatural things, other one with the intellectual intuition.

In the realm of reason mathematics, and in particular geometry, was – according to Pascal – the pattern and ideal. Geometry is the only domain of human knowledge which provides unfallible and indubitable proofs. The new ideal scientific method based on geometry should be founded on two rules: (1) one should not use any term whose meaning has been not exactly explained, and (2) all statements should be proved. Pascal was of course conscious of the fact that in scientific practice one cannot define all terms and prove all statements. Hence he allowed to accept without definition some terms which are clear by the "natural light" and to accept some clear initial principles (axioms) on which proofs can be based. Among such clear primitive terms are the notions of space, time, movement, number or equality. They are clear because the very nature gave us the understanding of them. Similarly for the case of axioms – they are clear by *le coeur*. But the deduction of theorems from the axioms proceeds in the realm of reason and according to its rules. Both are certain and secure though they take place on two different levels. The common feature of Descartes' and Pascal's ideas was the conviction of the universal character of mathematics as a pattern of a scientific theory – just mathematics was for them an ideal of human knowledge and its methods could (and should) be applied in all domains. The reason for that was the fact that only mathematics and its methods can lead to a secure and unfallible knowledge (only mathematics can give a real justification of its statements and claims). On the other hand they proposed several conditions which should be fulfilled to obtain a valuable theory.

Descartes' idea of constructing a universal science based on the patterns of mathematics and giving a frame for any scientific knowledge was further developed by Gottfried Willhelm Leibniz (1646-1716). He proclaimed the idea of a universal logical calculus. The latter was connected with his idea of treating logic in a mathematical way, i.e., by methods characteristic for mathematics. Leibniz attempted first of all to design a universal symbolic language, *characteristica universalis*. It was supposed to be a system of signs fulfilling the following conditions: (i) there is to be a one-one correspondence between the signs of the system (provided they are not signs of empty places for variables) and ideas or concepts (in the broadest sense), (ii) the signs must be chosen in such a way that if an idea (thought) can be decomposed into components then the sign for this idea will have a parallel decomposition, (iii) one must devise a system of rules for operating on the signs such that if an idea M_1 is a logical consequence of an idea M_2, then the 'picture' of M_1 can be interpreted as a consequence of the 'picture of M_2 (this is a sort of completeness condition).

According to these conditions all simple concepts corresponding to simple properties ought to be expressed by single graphical signs, complex concepts by combinations of signs. This was based on a fundamental general assumption that the whole possible vocabulary of science can be obtained by combinations of some simple concepts. The method of constructing concepts was called by Leibniz *ars combinatoria.* It was a part of a more general method – a calculus – which should enable people to solve all problems in a universal language. It was called *mathesis universalis, calculus universalis, logica mathematica logistica.* Leibniz hoped that *characteristica universalis* would, in particular, help to decide any philosophical problem.[3] He claimed that "he owed all his discoveries in mathematics exclusively to his perfect way of applying symbols, and the invention of the differential calculus was just an example of it" (cf. L. Couturat, *La logique de Leibniz,* Paris 1901, pp. 84-85). Note that Leibniz's idea was in fact an idea of a universal logic (based on mathematics) and the idea of mechanization of reasonings (which should be reduced to manipulations of symbols according to certain rules referring only to the form and not to the contents of statements written in the appropriate symbolic language). He proposed using mechanical calculations in aid of deductive reasoning (which meant in particular the introduction of calculations into logic).

Leibniz did not succeed in realizing his idea of *characteristica universalis.* One of the reasons was that he treated logical forms intensionally rather than extensionally. This could not be reconciled with the attempt to formalize logic completely and transform it into a universal mathematics of utterly unqualified generality. Another source of difficulty was his conviction that the combination of symbols must be a necessary result of a detailed analysis of the whole of human knowledge. Hence he did not treat the choice of primitive fundamental notions as a matter of convention. His general metaphysical conceptions induced a tendency to search for absolutely simple and primitive concepts (an analogue of monads), the combinations of which would lead to the rich variety of notions. As a partial realization of Leibniz's idea of *characteristica universalis* one can treat mathematical logic developed on the turn of the nineteenth century (see below).

The idea of a universal logical calculus which should form a framework for any valid reasoning in mathematics and generally in any scientific theory was not his only contribution to the epistemology of mathematics. The other one was his distinction between truths of reasoning and truths of fact on the one hand and the distinction between primitive and derived truths on the other. Primitive truths are truths known by intuition, they do not need any justification because they are clear by themselves and cannot be deduced from anything simpler and more sure. Derived truths are truths which can be reduced to primitive ones; they form the demonstrative knowledge. Truths of reasoning and truths of fact were characterized by Leibniz in the following way: "Truth of reasons are necessary and their opposite is impossible; truths of fact are contingent and their opposite is possible. When a truth is necessary, its reason can be found by analysis, resolving it into more simple ideas and truths, until we come to those which have primacy ..." (cf. *Monadology).* Truths of reason are grounded in the 'principle of contradiction' (which Leibniz took as covering the principle of identity and of the excluded middle). Facts can neither justify nor refute them. They are not based on facts and do not concern facts – they concern only the possibility. Hence they are true not only in the actual world but in all possible worlds, similarly as the logical laws. They do not say about any specific type of objects. Truths of fact are based on facts which can either justify or reject them. They are true only in the actual world.

Leibniz claimed that not only trivial tautologies but all the axioms, postulates, definitions and theorems of mathematics are truths of reason. This means that they are necessary and eternal, they are not based on facts or experience and are true in all possible worlds.

Leibniz's distinction between truths of reason and truths of fact found its development and in a sense a fulfilment by Immanuel Kant (1724-1804). All later analyses of the problem refered to him.

Starting from the assumption that the general form of a judgement is "*A* is *B*" Kant distinguished in *Kritik der reinen Vernunft* between analytic and synthetic judgements. A judgement is analytic if and only if nothing but reflection upon the concepts in the judgement and upon the form of combination of these concepts is needed to enable us to know whether the judgement is true. A judgement is synthetic if and only if mere reflection upon the concepts in it and upon their combination is not sufficient to enable us to know whether it is true; one must appeal to something further (experience or intuition or both are required). Consequently analytic judgements are uninformative, they tell us nothing we did not already have to know just to understand them. Observe that if one assumes additionally that analyticity of a judgement follows from the fact that its subject is included in its predicate then by definition all analytic judgements are true. On the other hand synthetic judgements can be either true or false, they are informative. Note that Kant's terminology was a bit psychologistic – but one can eliminate this difficulty and rephrase his definitions by replacing the term "judgement" by "statement". Kant claimed that all laws of formal logic are analytic but there are also other analytic statements.

Kant as the first linked up the analyticity and syntheticity with the property of being a priori or a posteriori. Roughly speaking one can say that a statement is a posteriori if it is empirical, i.e., based on experience and requiring justification from experience, and it is a priori if it is attainable prior to experience and its justification does not need any reference to experience. Kant formulated the following famous problem: how are synthetic a priori statements possible? He claimed that there are such statements in our knowledge, for example theorems of arithmetic or geometry. His answer to this fundamental question was revolutionary for the epistemology. He claimed that synthetic a priori statements are possible because there are a priori components in our knowledge. In fact space and time as forms of our intuition *(Anschauung)* and categories as forms of reason[4] are such a priori elements. They are necessary conditions of any knowledge. In particular the a priori intuition of time is a basis for arithmetic and the intuition of space – for geometry.

Kant divided synthetic a priori statements into two classes: intuitive and discursive. The former are connected with the structure of perception and perceptual judgements, the latter with the ordering function of general notions. An example of a discursive synthetic a priori proposition is the principle of causality. Kant claimed (contrary to Leibniz!) that all propositions of pure mathematics belong to the intuitive class of synthetic a priori statements. They cannot be empirical because they are necessary. On the other hand they are synthetic because they are about the structure of space and time as revealed by what can be constructed in them. They are a priori because pure space and time are a priori conditions of any perception of physical objects. On the other hand propositions of applied mathematics are either synthetic a posteriori (if they concern the empirical contents of sense perceptions) or synthetic a priori (if they concern the structure of space and time). Pure mathematics considers the structure of space and time free from empirical material and applied mathematics has for its subject matter the structure of space and time together with the material filling them.

It is necessary to stress here the distinction which Kant made between the thought of a mathematical concept and its construction. The former requires merely internal consistency while the latter requires that perceptual space should have a certain structure. Consequently one can postulate the existence of, for example, 5-dimensional sphere but one cannot construct it. Hence Kant did not deny the possibility of consistent geometries other than Euclidean one. So the development of non-Euclidean geometries in the nineteenth century did not refute Kant's philosophy of mathematics.

There is still one point in Kant's philosophical ideas concerning mathematics which should be indicated here. It is his theory of the actual infinity. Following Aristotle he distinguished between potential infinity and actual infinity. But he did not claim, as Aristotle did, that the latter is logically impossible. His idea was to treat it as an idea of reason, i.e., as an internally consistent notion which is however inapplicable to sense experience since instances of it can be neither perceived nor constructed. On the other hand it is needed in mathematics. This approach to the actual infinity will be later used by David Hilbert in his formalistic programme (see below).

Kant's apriorist thesis concerning mathematics has been criticized in the nineteenth century by empiricists. One of them was John Stuart Mill (1806-1873). He developed the methodological version of the empiricism and attempted to justify it using logic. He applied it also to mathematics and attempted to argue that mathematics is in fact an empirical science. In particular he claimed that the source of mathematics is the reality perceived by senses. Mathematical concepts have been simply abstracted from the reality by omitting some properties of real objects and by generalizing and idealizing other properties. Hence mathematical propositions are not necessarily true. Their necessity can be reduced only to the fact that they are logical consequences of assumptions. But the very assumptions are far from being necessary and certain, on the contrary, they are only hypothesis and in fact one can adopt any propositions as assumptions. Hence in mathematics the necessity can be attributed only to logical connections between propositions and not to very propositions themselves. Consequently mathematical propositions are necessary truths only in such a degree as the axioms are. But the latter can be any sentences, even more, in practice axioms are usually false statements because they do not describe the real world but are only idealizations and generalizations of its properties. One can easily see here a similarity of Mill's views and the views of Aristotle (though the latter did not claim that axioms of mathematics can be any sentences).

2 MODERN DOCTRINES IN THE EPISTEMOLOGY OF MATHEMATICS

Modern philosophy of mathematics was dominated by three schools: logicism, intuitionism and formalism. They emerged in the last quarter of the nineteenth century and the first thirty years of the twentieth century. They refered of course to earlier doctrines, e.g., to Plato, Aristotle, Leibniz and Kant. Their origin just in the period between 1870 and 1930 was connected with the origin and the intensive development of mathematical logic on the one hand and with the crisis in the foundations of mathematics at the end of the nineteenth century on the other.[5] It was connected with the discovery of antinomies in Cantors set theory, i.e., with the discovery of pairs of mutually inconsistent propositions each of which could be justified with the same degree of certainty. They are called today logical antinomies (to distingusih them from semantical antinomies in which the concepts of meaning and reference are involved). Their source was the imprecise notion of a set used by Cantor. Examples of those antinomies are Burali-Forti antinomy of the greatest ordinal (it was known to Cantor), Cantons antinomy of the set of all sets and Russell's antinomy of the irreflexive classes (called today simply Russell's antinomy). Attempts to overcome difficulties revealed by the antinomies stimulated the researches also on the philosophical level and led to the development of logicism, intuitionism and formalism. This would not be possible without the modern mathematical logic developed at the end of the nineteenth century (this is true especially in the case of logicism and formalism).

Mathematical logic was a partial realization of Leibniz's idea of *characteristica universalis*. Its development was connected with the process of mathematization of logic and with the extention of Aristotelian logic (which was in fact restricted to the logic of names). Among the pioneers of the modern logic one should mention August de Morgan, George Boole, Charles Sanders Peirce and Ernst Schröder. They developed the so called algebra of logic. Beside it there has been also developed the non-algebraic form of logic by Gottlob Frege and Bertrand Russell. Both trends provided the necessary technical background which enabled the development of logicism and formalism. In particular Frege's fundamental logical work *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens* (1879) opened the new period in the history of formal logic (though its reception by the contemporaries, especially by the representatives of the algebraic trend, for example by E. Schröder or J. Venn, was not enthusiastic, on the contrary, Frege's work was not accepted and quickly forgotten and the reasons for that concerned not only the merits – for example Frege used a very complicated and "hard" symbolism – but also of personal nature). *Begriffsschrift* contained the (first in the history of logic) formal axiomatic system, more exactly the formal system of propositional calculus with implication and negation as the only connectives, and provided for the first time full analysis of the notion of a quantifier by giving a suitable system of axioms for them. Frege's formal system of propositional calculus was a system in which the deduction of theorems from axioms was complete and without gaps, i.e., rules of inference and the very notion of a formal proof have been precisely defined at the very beginning. In this way Frege's work contributed in an essential manner to the clarification and making precise of the basic notion of the epistemology of mathematics, i.e., of the notion of proof. Recall that since Plato, Aristotle and Euclid the axiomatic-deductive method was treated as the best method of mathematics, but the very notion of a proof was in fact understood rather intuitively and no precise definition of it was given. Having fixed and described in a precise way the rules on inference, Frege defined a proof as a sequence of sentences in a fixed precisely described formal language such that every element of this sequence is a result of applying one of the rules of inference to formulas appearing earlier in this sequence and the rules of inference could refer only to the form of formulas and not to their meaning or sense. In this way the intuitive notion of a consequence (being in fact of a rather psychological than of mathematical or logical character) has been replaced by a formal logical notion of a proof and consequence based on the notion of a rule of inference. This approach has been developed later (among others by Bertrand Russell, David Hilbert and Kurt Gödel) and contributed very much to the establishing of a new paradigm in mathematics, namely of the logico-set-theoretical paradigm.

Discussing the sources of modern doctrines in the philosophy of mathematics one should mention also the development of the non-Euclidean geometries in the thirties of the nineteenth century and the axiomatization of geometry at the end of the nineteenth century. Their impact can be seen first of all in the philosophy of geometry but they influenced the epistemology of mathematics in general as well. In particular they contributed to the discussion of the problem whether mathematical knowledge is an a priori or a posteri knowledge, whether mathematics is an analytic or synthetic science. The fact that non-Euclidean geometries have been shown to be as logically consistent as Euclidean geometry shook the conviction that a mathematical theory should be based on true axioms, that geometry is a description of properties of the real space and that consequently the choice of primitive concepts and axioms is in a sense determined by the physical reality.

The idea of axiomatizing geometry in a complete way appeared simultaneously in works of several mathematicians. The most important were here the contributions of Moritz Pasch, Giuseppe Peano and the Italian school (G. Veronese, M. Pieri, F. Enriques) and David Hilbert. Pasch in his *Vorlesungen über neuere Geoemetrie* (1882) formulated the idea that if geometry is to be really a deductive science then the process of deducing theorems in it must be independent of any meaning of geometrical notions not determined by axioms. The final separation of geometry from the empirical reality was done in Hilbert's *Grundlagen der Geometrie* (1899). In this way geometry become pure mathematical theory. Axioms were not treated any longer as evident and necessary statements. The question about their truth lost its meaning and sense. As axioms any sentences could be adopted. The main problem was now the problem of consistency of given axioms. Geometrical deductive systems became uninterpreted axiomatic systems various interpretations of which are possible. In this way the traditional philosophical view which regarded geometrical knowledge as synthetic a priori knowledge of our world has been decisively refuted.

2.1 Logicism

The main thesis of logicism states that mathematics can be reduced to logic, i.e., mathematics is a part of logic and logic is an epistemic ground of all mathematics. This thesis can be formulated as the conjunction of the following three theses: (1) all mathematical concepts (in particular all primitive notions of mathematical theories) can be explicitly defined by purely logical notions, (2) all mathematical theorems can be deduced (by logical deduction) from logical axioms and definitions, (3) this deduction is based on a logic common for all mathematical theories, i.e., the justification of theorems in all mathematical theories referes to the same basic principles which form one logic common for the whole of mathematics (here is also included a thesis that any argumentation in mathematics should be formalized). Hence theorems of mathematics have uniquely determined contents and it is a logical contents. Moreover, mathematical theorems are analytic (similarly as logical theorems). Note that the founders of logicism did not state precisely what is the character of logical laws. Frege for example claimed only that they are not rules of nature but "rules of rules of nature" and that they are not laws of thinking but "laws of truth".

The genesis of logicism can be seen in the philosophical controversy between rationalism and empiricism concerning the character of mathematical propositions (judgements). Logicism was connected with Locke's and Leibniz's view that mathematical propositions have tautological character and with the Leibniz's view about the possibility of algorithmizing all mathematical reasoning (and generally all scientific reasonings).

Logicism can be treated as an extension of the nineteenth century tendency to the unification of the classical mathematics (in particular of analysis) by reducing it to the arithmetic. Pioneers of this trend were Karl Weierstrass and Richard Dedekind. It consisted in showing that the theory of real numbers which forms the foundations for analysis can be developed on the basis of the arithmetic of natural numbers, i.e., that the notion of a real number can be defined in terms of natural numbers and some concepts of set theory and that all properties of reals can be deduced from theorems of the arithmetic of natural numbers. This task has been in fact done by R. Dedekind in his paper *Stetigkeit und irrationale Zahlen* (1872).

In this situation the problem of reducing the arithmetic of natural numbers to a simpler theory arose. It was solved by Gottlob Frege (1848-1925), the founder of logicism, in two of his books: *Grundlagen der Arithmetik* (1884) and *Grundgesetze der Arithmetik* (vol. – I 1893, vol. II – 1903). Frege opposed there both empiricism claiming that arithmetical laws are simply inductive generalizations and formalism which treated them as rules of operating on symbols as well as kantianism which viewed them as synthetic a priori propositions. Frege claimed that arithmetic (and consequently the whole of mathematics) is a part of logic and that its theorems are analytic and *a fortiori* they are a priori.[6] On the other hand Frege disagreed with Kant who stated that mathematical theorems do not extend our knowledge, on the contrary, Frege claimed that they are in fact informative. Frege's definition of analyticity said that a sentence *A* is analytic if and only if *A* is provable on the basis of logic and definitions alone. He claimed that all arithmetical notions can be defined by logical concepts and all arithmetical theorems can be deduced from laws of logic and definitions. He showed in the indicated works how this can be done. He applied here the notion of the equipollence of sets introduced by Cantor. In fact Frege did not use sets (which are not logical notions) but was talking about concepts and their extensions and about the equipollence of extensions. This enabled him to remain on the ground of logic only. In modern terminology one would say that Frege first defined a cardinal number as a class of equipollent sets and then defined the natural numbers in terms of the successor function. Note that the concepts were understood by Frege in an absolute, platonic way as existing independently of time, space and human knowledge. Mathematicians do not create them and their properties but are only discovering them.

Having defined the notion of a natural number in terms of logical notions Frege proved in *Grundgesetze* several properties of them. Unfortunately it turned out that the system of logic used by him was inconsistent. This was discovered by Bertrand Russell (1872-1970) in 1901 and in 1902 communicated to Frege. In fact Russell observed that one can construct in Frege's system an antinomy of irreflexive classes called today Russell's antinomy.[7] This antinomy was published and analysed by him in the book *The Principles of Mathematics* (1903) where he expressed (and defended) several views on the foundations of mathematics close to those of Frege.

Russell undertook the task of reducing mathematics to logic in the monumental work *Principia Mathematica* written together with Alfred North Whitehead (1861-1947) (vol. I – 1910, vol. II – 1912, vol. III – 1913). One finds there a new completely reconstructed system of logic called the ramified theory of types. It was based on the general assumption that the totality of properties which can be considered forms an infinite hierarchy of types: properties of the first type are properties of individuals, properties of the second type are properties of properties of the first type, etc. This hierarchy does not contain properties which could hold of objects of different levels. To avoid the vicious circle of inpredicative definitions Russell and Whitehead introduced not only types but also orders of properties (orders depended on the form of a formula describing a considered object or property). By using those means they could eliminate the antinomy of irreflexive classes.

Properties, called by Russell prepositional functions, played in the theory of types the same role as concepts and their extensions by Frege. Hence Russell and Whitehead could simply adopt his definition of natural numbers. Some new problems appeared while proving properties of natural numbers. In particular it has turned out that to prove that for every natural number there exists a successor of it one needs an additional assumption being of a non-logical character, in fact the axiom of infinity stating that there exist infinitely many individuals is needed. Russell proposed a solution to this problem by suggesting to consider in the case of any such theorem not the theorem itself but rather an implication the antecedent of which is just the axiom of infinity and the succedent the considered theorem (by deduction theorem this implication can be proved in logic). In a similar way one can treat also other theorems whose proofs require additional assumptions such as for example the axiom of choice.

It has been shown in *Principia* how to reduce not only the arithmetic but the whole of mathematics to logic, i.e., to the ramified theory of types. It is worth noting here that – contrary to Frege – Russell and Whitehead treated concepts not in a platonic way but in a nominalistic way. Hence they did not postulate the independent existence of sets and all symbols for sets treated only as signs denoting nothing. Sets were reduced by them to propositional functions.

Russell claimed that the whole of mathematics is analytic.[8] His notion of analyticity can be given by the equivalence that a sentence A is analytic if and only if A is a tautology and a tautology can be characterized by three properties: (i) it is a priori, (ii) its negation is inconsistent and (iii) it is invariant with respect to logical constants (note the similarity of the latter to Bolzano's characterization of analyticity). On the other hand it is not clear whether Russell admitted that tautologies are "empty" or, on the contrary, they are informative – in fact he wrote in *Introduction to Mathematical Philosophy* that "logic is concerned with the real world just as truly as zoology, though with its more abstract and general features" (p. 169).

Russell's thesis about the analyticity of the whole of mathematics concerned also geometry (compare this with the views of Frege). In fact Russell distinguished two types of geometry: pure geometry in which one deduces consequences from the adopted axioms and does not ask whether they are true or not – this geometry is a priori and is not synthetic, hence it is analytic; and applied geometry being in fact a part of physics – it is an empirical science and is synthetic but not a priori.

The system of *Principia Mathematica* has been changed and modified later, in particular by Chwistek and F. P. Ramsey who introduced the so called simple theory of types. This was an extensional theory (in contrast with the original Whitehead and Russell's theory which was in fact intensional). The theory of types has been accepted as the best system for the foundations of mathematics, it became a basis to which other researches refered (cf. for example Tarski's theory of satisfaction and truth (1933) or Gödel's work on the incompleteness of first-order systems (1931)). It played this role till the fifties when its functions were taken up by the axiomatic set theory. The system of *Principia Mathematica* (and its later simplifications) was the first complete, consistent and natural system of logic. It was in a sense a synthesis of all earlier conceptions in the field of logic and the foundations of mathematics. It indicated the power and meaning of formal methods in logic and mathematics, in particular it showed that the formal principles of logic provide a sufficient tool for deduction of theorems from any given axioms.

Logicism played an important role in the development of the foundations of mathematics and of the philosophy of mathematics. Though various critical remarks concerning it (and in particular the theory of types) has been formulated one of its fundamental merits is to indicate the elegant way of a systematization of the mathematical knowledge and of making precise the intuitive notion of a mathematical proof. The logicists used here the results of the mathematical logic (and on the other hand contributed very much to its development). They underlined the universal and simultaneously fundamental character and role of logic in mathematics. Since it has turned that the reduction of mathematics to logic done by logicists was in fact the reduction to logic in a broader sense (several non purely logical assumptions were necessary in this reduction), the today's version of it says that the whole of mathematics is reducible to logic and set theory. There exists also another version of logicism, a methodological one, called if-thenism. It is based on the finitistic character of the operation of logical consequence and on the deduction theorem and claims that theorems proved in mathematical theories should be understood as implications whose antecedents are finite conjunctions of axioms; such implications are theses of logic. Add at the end that logicism is not necessarily connected with the thesis about the analyticity of mathematical statements. In fact logisism claims only that there is a certain "homogeneity" of logic and mathematics with respect to the partition of propositions into synthetic and analytic. One can of course imagine a version of logicism claiming that logic and mathematics are both synthetic a priori or a posteriori.

2.2 Intuitionism

Intuitionism as a doctrine in the philosophy of mathematics has been founded by the Dutch mathematician Luitzen Egbertus Jan Brouwer (1881-1966). Intuitionists saw as their predeccessors those philosophers and mathematicians who claimed that mathematics is a science possessing a definite contents and that the human mind is able to perceive directly mathematical objects and to formulate about them synthetic a priori judgements. Hence they willingly refered to I. Kant and to Paul Natorp. It seems that the ideas expressed later explicitely in works of Brouwer were in a sense in the air at the end of the ninteenth century.

For the first time the intuitionistic ideas appeared by German mathematician Leopold Kronecker (1823-1891) and his students in the seventies and eighties of the nineteenth century. In *Über den Zahlbegriff* (1887) Kronecker formulated a programme of "arithmetization" of algebra and analysis, i.e., the programme of founding those domains of mathematics on the most fundamental notion of a number.[9] He developed a unified theory of various types of numbers based on the primitive intuition of a natural number. His scientific credo has been summerized in the best way in his sentence: *"Die ganzen Zahlen hat der lieber Gott gemacht, alles andere ist Menschenwerk"*.[10] A consequence of this attitude was for example the fact that Kronecker admitted only those definitions of numbers which give a procedure of deciding whether a given number satisfies it or not. He accepted only "pure" existence proofs, i.e., proofs of existential theses giving constructions of the postulated objects.

Philosophical and methodological theses of Kronecker were a reaction to K. Weierstrass' attempts of applying methods of Cantors set theory (in particular the theory of infinite sets) to the number theory and to the theory of functions. Similar were the sources of the ideas of the group of French mathematicians called today the Paris school of intuitionism or French semi-intuitionists: R.L. Baire, E. Borel, H.L. Lebesgue and the Russian mathematician N.N. Luzin. Their considerations on the foundations of mathematics were mainly connected with the study of the role of the axiom of choice. They did not create a compact philosophy doctrine but formulated several general remarks on the margin of their mathematical investigations in the theory of functions. Their common feature is certain constructivistic tendency. Many of their views were later adopted by Brouwer.

Discussing the problem of forerunners of Brouwer's intuitionism one must mention the French mathematician Henri Poincaré (1854-1912). His philosophical attitude can be characterized by saying that he was an apriorist, intuitionist and constructivist as well as a founder of the conventionalism. According to Poincaré the main role in the human mathematical knowledge is played by the creative activity of the mind and by its ability to construct concepts. This creative activity is manifested in various ways. One of its manifestation is intuition which appears both in the unconscious as well as conscious work. It has spontaneous and rational character, it gives the consciousness of clearity and evidence. It does not need the evidence of senses. There are various kinds of intuition: generalization by induction, the intuition of pure number, reference to senses or imagination, preexisting ability to construct the concept of a group as a pure, and not sensory, form of knowledge (this idea was especially important in his philosophy of geometry). Poincaré claimed that intuition should be supplemented by a discursive knowledge, i.e., it should be proceeded and concluded by the conscious work in which the intuitive "revelations" are verified in a rational way.

One of the kinds of intuition is mathematical induction. It forms the base of number theory and of the whole of mathematics. According to Poincaré mathematical induction is a synthetic a priori judgement which extends our knowledge, it is the archetype of mathematical reasoning. Therefore theorems of mathematics (with the exception of geometry!) based on it are synthetic. He considered reasonings using induction as "the exact type of the a priori synthetic intuition" (cf. *La science et l'hypothèse,* English translation, p. 388). Induction is not empirical, it cannot be reduced to logic. Logic is tautological and cognitively empty, it enables us only to build analytical judgements, therefore mathematics cannot be reduced to logic only (as logicists maintain). Our trust in mathematical induction comes from the fact that "it is only the affirmation of the power of the mind which knows it can conceive of the indefinite repetition of the same act, when the act is once possible" (cf. *La science et I'hypothèse,* English translation p. 388).

Poincaré claimed that mathematical objects are being constructed by human mind, that there is no domain of mathematical knowledge which would be independent of the knowing subject. One of the consequences of this antirealistic attitude was the rejection of the actual infinity and the restriction of mathematics to objects which can be defined by finitely many words only.

Discussing Poincaré's philosophy of mathematics one must mention also his conventionalism. It was manifested mainly in his philosophy of natural science but also in his philosophy of geometry. In the latter it contributed to the overcoming of the traditional views according to which geometry is the description of the real empirical space. Poincaré claimed that axioms and postulates of geometry are (contrary to statements of arithmetic!) neither synthetic a priori judgements nor empirical facts. They are conventions and implicite definitions. The choice between various possible conventions is based on experimental facts but we are free in this choice – the unique restriction is to avoid inconsistency. The question which geometry is a true geometry (this problem was especially important after the introduction and development of the non-Euclidean geometries) is, according to Poincaré a wrong question. He wrote: "One geometry cannot be more true than the other; it can be only more convenient" *(loc. cit.).* And the Euclidean geometry turns out to be the most convenient one. Poincaré used the idea of F. Klein to characterize geometries as theories of invariants of appropriate groups of transformations and the very concept of a group, more exactly the possibility of constructing a concept of a group as a pure form of knowledge preexisting in our mind (it was treated by him as a part of intuition). In this way geometry was not any longer *"la science de la verité"* but it became *"la science de la consequence"* (Poincaré).

Brouwer presented his philosophical views for the first time in his doctoral dissertation *Over de Grondslagen der Wiskunde* (1907).[11] The main aim of him (and of the intuitionism) was to avoid inconsistencies in mathematics. Brouwer proposed means to do that which turned out to be very radical and led in the effect to the deep reconstruction of the whole of mathematics.

The main fundamental thesis of Brouwer's intuitionism is the rejection of Platonism in the philosophy of mathematics, i.e., of the thesis about the existence of mathematical objects which is independent of time, space and human mind. The proper ontological thesis is the conceptualism. According to Brouwer mathematics is a function of the human intellect and a free activity of the human mind, it is a creation of the mind and not a theory or a system of rules and theorems. Mathematical objects are mental constructions of an (idealized) mathematician. As a consequence one should reject the axiomatic-deductive method as a method of developing and founding mathematics. It is not sufficient to postulate only the existence of mathematical objects (as it is done in the axiomatic method) but one must first construct them. One must also reject the actual infinity. An infinite set can be understood only as a law or a rule of forming more and more of its elements, but they will never exist as forming an actual totality. Hence there are no uncountable sets and no cardinal numbers other than α_0.

The conceptualistic thesis of intuitionism implies also the rejection of any non-constructive proofs of the existential theses, i.e., of proofs giving no constructions of the postulated objects. In fact in intuitionism "to be" means "to be constructed". Brouwer claimed that just nonconstructive proofs were the source of all troubles and mistakes in mathematics. Since such proofs are usually based on the law of the excluded middle ($p \vee \neg p$) as well as on the law of double negation ($p \equiv \neg\neg p$), intuitionists could not any longer use the classical logic. Moreover they claimed that logic is neither a basis for mathematics nor a starting point of it. Brouwer said that logic is based on mathematics, that it is secondary and dependent on mathematics and not vice versa, i.e., logic is a basis for mathematics as the logicists assert. According to Brouwer and the intuitionism mathematics is based on the fundamental intuition of time. One sees here the connection with the philosophy of Kant. Indeed the intuitionism accepts Kant's thesis about the a priori time and rejects thesis about the a priori space. The intuition of time is a basis for the mental construction of natural numbers. Moreover this construction is the basic mathematical activity from which all other mathematical activity springs. One of the consequences of this is that arithmetical statements are synthetic a priori judgements.

A mathematical theorem is a declaration that a certain mental construction has been completed. All mathematical constructions are independent of any language. Hence there is no language (formal or informal) which would be safe for mathematics and would protect it from inconsistencies. It is a mistake to analyse the language of mathematics instead of analysing the mathematical thinking. Mathematics should be justified not "on paper" but "in the human mind". This intuitionists' thesis contradicts the thesis of the logicism and especially of the formalism about the meaning and importance of the formal reconstruction of mathematics and of the formal proof of the consistency of mathematical theories. It is also incomparable with views of all those philosophers and mathematicians since Plato who claimed that any abstract thinking is dependent of a language.

One of the consequences of Brouwer's theses described above and forming a base for the doctrine of intuitonism was the need for the reconstruction of the whole of mathematics according to the intuitionists' principles. Brouwer began to realize this task with his students in 1912. They reconstructed the concept of the continuum, the theory of point sets, theory of functions, theory of countable well orderings, the theory of complex functions, projective geometry, algebra, topology, measure theory, affine geometry and others. From the point of view of the epistemology of mathematics most important were the works of Brouwer's pupil Arend Heyting (1898-1980). Indeed he popularized the ideas of Brouwer and attempted to explain them in a language usually used in the reflection on mathematics – note that Brouwer expressed his ideas in a language far from the standards accepted by mathematicians and logicians and therefore not always understandable. It seems that without those attempts of Heyting the ideas of Brouwer would soon disappear. In particular Heyting constructed the first formalized system of the intuitionistic propositional calculus, i.e., of the propositional calculus satisfying principles of the intuitionism. This system has been never accepted by Brouwer. This sceptical attitude has its reasons in the thesis about the essential inexhaustibility of the totality of mental processes which can be accepted as valid. Consequently there exists no formal system which would adequately represent the human mathematical activity. Indeed the latter is always dynamic and not closed while the former is static and closed. Hence no formal system can be a complete and adequate description of the intuitionistic mathematics. The latter is more poor than the classical mathematics and, what more, much more complicated and consequently not so useful in applications. On the other hand intuitionistic logic has turned out to be a very useful tool in various parts of mathematics, in particular in the topos theory or in the theoretical computer science.

2.3 Constructivism

Intuitionism is one of the constructivistic trends in the foundations and philosophy of mathematics. Constructivism is a common name for various doctrines the main thesis of which is the demand to restrict mathematics to the consideration of constructive objects and to constructive methods only. Hence constructivism is a normative attitude the aim of which is not to build appropriate foundations for and to justify the existing mathematics but rather to reconstruct the latter according to the accepted principles and to reject all the methods and results which do not fulfil them. The constructivistic tendencies appeared in the last quarter of the nineteenth century as a reaction against the intensive development of highly abstract methods and concepts in mathematics inspired by Cantor's set theory.

There are various constructivist programmes and schools. They differ by their interpretation and understanding of the concept of constructivity. One of the most developed schools is intuitionism discussed above. Others are finitism, ultraintuitionism (called also ultrafinitism or actualism), predicativism, classical and constructive recursive mathematics, Bishop's constructivism. It is impossible to describe here all those doctrines in detail. Note that they are not based on one philosophical system, on the contrary, they accept different, not always compatible, philosophical, in particular ontological assumptions. In general one can distinguish four types of constructivism according to the accepted ontological basis: (1) objectivism claiming that objects of mathematics are objective results of constructive processes existing independently of the knowing subject which constructs them, (2) intentionalism which ascribes to mathematical objects being results of appropriate constructive processes the intentional existence (being characteristic for cultural entities), (3) mentalism claiming that objects of mathematics being products of mental acts exist only in those acts, (4) nominalism according to which mathematical objects are concrete and definite spatio-temporal objects.

Constructivistic tendencies in mathematics contributed very much (and are still contributing) to making precise various notions and ideas of mathematics. They are very important also for the computer science. On the other hand constructivistic mathematics is in fact much more poor than the classical one.

2.4 Formalism

The third main school in the philosophy of mathematics is formalism created by German mathematician David Hilbert (1862-1943). Hilbert was of the opinion that the attempts to justify and found mathematics undertaken hitherto, especially by the intuitionism, were unsatisfactory because they led to the restriction of mathematics and to the rejection of various parts of it, in particular those considering infinity. The aim of his programme formulated for the first time in his famous lecture at the Second International Congress of Mathematicians in Paris in 1900 was to save the integrity of the existing classical mathematics (dealing with the actual infinity) by proving that it is secure. Among twenty three main problems which should be solved Hilbert mentioned there as Problem 2 the task of proving the consistency of axioms of arithmetic (under which he meant number theory and analysis) (cf. Hilbert 1901). He has been returning to the problem of justification of mathematics in his lectures and papers, especially in the twenties, where he proposed a method of solving it called today Hilbert's programme. One should add that Hilbert saw the supramathematical significance of the whole issue writing that "the definite clarification of the nature of the infinite has become necessary, not merely for the special interests of the individual sciences but for the honor of human understanding itself" (cf. *Über das Unendliche).*

Hilbert's programme of clarification and justification of mathematics was Kantian in character. One can see here a turn in the direction of idealism. In Kant's philosophy ideas of reason, or transcendental ideas, are concepts which transcend the possibility of experience but on the other hand are answer to a need in us to form our judgements into systems that are complete and unified. Therefore we form judgements concerning an external reality which are not uniquely determined by our cognition, judgements concerning things in themselves. To do that we need ideas of reason.

In likening the infinite to a Kantian idea of pure reason, Hilbert suggested that it is to be understood as a regulative rather than a descriptive device. Sentences concerning the infinite, and generally expressions which Hilbert called ideal propositions, should not be taken as sentences describing externally existing entities. We use ideas of reason and ideal elements in our thinking because they allow us to retain the patterns of classical logic in our reasonings.

Hilbert distinguished between the unproblematic, 'finitistic' part of mathematics and the 'infinitistic' part that needed justification. Finitistic mathematics deals with so called real propositions which are completely meaningful because they refer only to given concrete objects (add that real propositions play the role of Kant's judgements of the understanding (Verstand)). Infinitistic mathematics on the other hand deals with the so called ideal propositions that contain reference to infinite totalities (they play the role of Kant's ideas of pure reason). Ideal propositions play an auxiliary role in our thinking, they are used to extend our system of real judgements. Hilbert believed that every true finitary proposition had a finitary proof. Infinitistic objects and methods enabled us to give easier, shorter and more elegant proofs but every such proof could be replaced by a finitary one (this is the reflection of Kant's views of the relationship between ideas of reason and the judgements of the understanding). Hilbert was also convinced that consistency implies existence and that every proof of existence not giving a construction of postulated objects is in fact a presage of such a construction.

Hilbert proposed to justify the infinitistic mathematics by finitistic methods because only they can give it security. He wanted to do it via proof theory *(Beweistheorie)*. Its main goal was to show that proofs which use ideal elements in order to prove results in the real part of mathematics always yield correct results. One can distinguish here two aspects: consistency problem (prove by finitistic method that the infinitistic mathematics is consistent) and conservation problem (show by finitistic methods that any real sentence which can be proved in the infinitistic part of mathematics can be proved also in the finitistic part, even more, that there is a finitistic method of transforming infinitistic proofs of real sentences into finitistic ones).

Hilbert's proposal to carry out this programme consisted in two steps. To be able to study seriously mathematics and mathematical proofs one should first of all define accurately the very concept of a proof. Hilbert used here the results of mathematical logic (G. Peano, G. Frege and B. Russell), in particular the idea of a formalized system in which a mathematical proof is reduced to a series of very simple and elementary steps, each of which consists in performing a purely formal transformation on the sentences which have been previously proved. Hence the first step proposed by Hilbert was to formalize mathematics, i.e., to reconstruct infinitistic mathematics as a big, elaborate formal system. The second step was to give a proof of the consistency and conservativeness of mathematics by considering formal proofs, i.e., strings of symbols of the appropriate artificial symbolic language. This was just the aim of the proof theory created by Hilbert (and called also metamathematics). It was a theory in which (formalized) mathematical theories and their properties are to be studied by mathematical methods.

One should note here that formalization was for Hilbert only an instrument used to prove the correctness of (infinitistic) mathematics. Hilbert did not treat mathematical theories as games on symbols or collections of formulas without any contents. Formalization was only a methodological tool in the process of studying the properties of the preexisting mathematical theories.[12] On the other hand Hilbert, contrary to the intuitionists, connected thinking with a language. He claimed that thinking, similarly to the process of speaking and writing, takes place by constructing and ordering sentences.

Hilbert represented a strongly antilogicistic attitude. He maintained that mathematics cannot be deduced from logic alone, logic does not suffice to justify mathematics – hence the attempts of Frege and Russell were in his opinion fruitless.

Hilbert and his school had scored some successes in realization of the programme of justifying infinitistic mathematics. In particular Wilhelm Ackermann showed by finitistic methods the consistency of a fragment of arithmetic of natural numbers. But soon something was to happen that undermined Hilbert's programme. In 1931 the Austrian mathematician Kurt Gödel (1906-1978) proved that arithmetic of natural numbers and all formal systems containing it are essentially incomplete provided they are consistent (and based on a recursive, i.e., effectively recognizable set of axioms).[13] He announced also a theorem stating that no such theory can prove its own consistency, i.e., to prove the consistency of a given theory T containing arithmetic one needs methods and assumptions stronger than those of the theory T. Hence in particular one cannot prove the consistency of an infinitistic theory by finitistic methods.[14]

Gödel's methods were used to indicate still another feature of axiomatic theories. The studies initiated by Alonzo Church and continued by others have shown that most theories interesting from the mathematical point of view are undecidable, i.e., there does not exist (and cannot exist) an effective method for deciding whether a given statement can be proved (justified) on the basis of a given system of axioms.[15] Gödel's results indicated certain limitations of the axiomatic-deductive method considered since antiquity to be the best method for mathematics. They showed that one cannot include the whole of mathematics in a consistent formalized system based on the first order predicate calculus – what more, one cannot even include in such a system all truths about natural numbers. In this way it has been also shown that the concept of a formal proof which was supposed to be the precization of the imprecise notion of a mathematical proof is not adequate. In the research practice mathematicians are using any (correct) methods to solve problems and to answer questions. The scope of the admissible methods is not fixed or bounded beforehand. They are chosen according to the needs and problems that appear. On the other hand there is no precise definition of a correct method in mathematics. Therefore the hopes that the precise notion of a formal proof based on (first order) logic (with fixed axioms and rules of inference) will provide such a definition. Gödel's incompleteness theorems indicated that this is not (and cannot be) the case. They revealed also the distinction between syntactical and semantical notions, in particular between provability and truth. Note that formalists considered formal provability to be an analysis of the concept of mathematical truth. Gödel showed that semantic truth cannot be adequately expressed by syntactical provability. In fact there is a "gap" between them, more exactly the notion of provability (for any first order formal theory) is definable in the language of arithmetic of natural numbers by a formula containing only one existential quantifier (such formulas a called $\Sigma^0{}_1$ formulas) while the notion of a true sentence of the arithmetic of natural numbers is not definable by an arithmetical formula (hence it is not arithmetic, indeed it is hyperarithmetic; similarly for other theories).

One can also prove that it is not only impossible to characterize mathematical structures, e.g., the structure of natural numbers, adequately by a (first order) formalized theory (because such theories are always incomplete), but that any description of a considered structure (such as the structure of natural numbers) by (first order) axioms is inadequate in the sense that the theory has a great variety of models, even models very different from the structure one wants to describe. Hence first order logic assumed to be the best tool in reconstructing mathematics is too weak. On the other hand higher order characterizations (e.g., by second order notions and second order logic) are not so regular and natural (from the logical, methodological and philosophical point of view).

Gödel's results struck Hilbert's programme. Did they reject it? This question cannot be answered definitely for the simple reason – Hilbert's programme was not formulated precisely enough, it used vague terms as finitistic, real, ideal which were never precisely defined. Both Hilbert and Gödel were ready after the incompleteness theorems to extend the scope of admissible methods by allowing some forms of infinitistic reasonings. Gödel doubted whether all correct proofs can be captured in a single formalized system.

The idea of extending the admissible methods and allowing general constructive methods instead of only finitistic ones was explicitely formulated by Paul Bernays. A motivation for this shift from the original Hilbert's programme could be Gödel's reduction (found independently also by Gerhard Gentzen) of classical arithmetic to the intuitionistic arithmetic of Heyting and Gentzen's proof of the consistency of arithmetic by transfinite induction (which was apparently accepted by Hilbert). But there arises a problem: what is meant by constructivity? This concept is in general much less clear than that of finitism. Nevertheless the broadening of original Hilbert's proof theory postulated by Bernays has become an accepted paradigm (it is usually called the generalized Hilbert's programme). Investigations were carried out in this direction and several results have been obtained.

Another consequence of Gödel's incompleteness results is the so called relativized Hilbert's programme. If the entire infinitistic mathematics cannot be reduced to and justified by finitistic mathematics then one can ask for which part of it is that possible. In other words: how much of infinitistic mathematics can be developed within formal systems which are conservative over finitistic mathematics with respect to real sentences? This question constitutes the relativized version of the programme of Hilbert. Recently results of the so called reverse mathematics developed mainly by H. Friedman and S.G. Simpson contributed very much to this programme.[16] In fact they showed that several interesting and significant parts of classical mathematics are finitistically reducible. This means that Hilbert's programme can be partially realized.

2.5 Logico-set-theoretical paradigm

Studies on the foundations of mathematics and on the philosophy of mathematics in the nineteenth century and in the first half of the twentieth century led to the establishing of the new paradigm of mathematics, called logico-set-theoretical paradigm. It replaced the Euclidean paradigm (described above) prevailing up until the end of the nineteenth century. Several events and achievements contributed to the establishing of the new paradigm. Among the most important are the origin and the development of set theory (G. Cantor), arithmetization of analysis (A. Cauchy and K. Weierstrass, R. Dedekind), axiomatization of the arithmetic of natural numbers (G. Peano), non-Euclidean geometries (N. I. Lobachewsky, J. Bolayi, C.F. Gauss), axiomatization of geometry (M. Pasch, D. Hilbert), the development of mathematical logic (G. Boole, A. de Morgan, G. Frege, B. Russell). Besides those "positive" factors there was also a "negative" factor, viz., the discovery of antinomies in the set theory (C. Burali-Forti, G. Cantor, B. Russell) and semantical antinomies (G.D. Berry, K. Grelling). They showed that the intuitive concept of a set is vague and a precise definition of it is needed. The latter was provided by axiomatizing set theory (E. Zermelo, T. Skolem). Semantical antinomies indicated the necessity of distinguishing between language and metalanguage.

The main features of the logico-set-theoretical paradigm can be characterized as follows: (1) set theory became the fundamental domain of mathematics, in particular some set-theoretical notions and methods are present in any mathematical theory and set theory is the basis of mathematics in the sense that all mathematical notions can be defined by primitive notions of set theory and all theorems of mathematics can be deduced from axioms of set theory, (2) languages of mathematical theories are strictly separated from the natural language, they are artificial languages and the meaning of their terms is described exclusively by axioms; some primitive concepts are distinguished and all other notions are defined in terms of them according to precise rules of defining notions, (3) all mathematical theories have been axiomatized,[17] (4) there is a precise and strict distinction between a mathematical theory and its language on the one hand and metatheory and its metalanguage on the other (the distinction was explicitely made by A. Tarski), (5) two crucial concepts for mathematics, i.e., the concept of a consequence and the concept of a proof have been precisely defined.[18]

One should emphasize here the significant role played by the mathematical logic and the foundations of mathematics in the development of the philosophy of mathematics. This has been evident especially after 1930. Results of those domains contributed to the process of making precise various philosophical problems and explaining crucial methodological concepts (such as proof, truth, consistency, etc.) and indicated several important properties of axiomatic systems which implied the necessity of the revision of some ideas of the epistemology of mathematics. In particular one should mention here the precise definition of truth and model (Tarski) and the so called limitation theorems, i.e., Gödel's incompleteness theorems, Churches theorem on the undecidability, Tarski's theorem on the undefinability of truth and Skolem-Löwenheim's theorems on the cardinality of models and on nonstandard models (indicating the impossibility of a unique characterization of structures by first order axiomatic systems).

Studies on the foundations of geometry and arithmetic and especially the metamathematical studies on the set theory pointed out some problems connected with the axiomatization of mathematics. Gödel showed that all richer (i.e., containing arithmetic of natural numbers) theories are essentially incomplete. Axiomatization of geometry and the development of non-Euclidean geometries threw some light on the problem which system of geometry is true. On the other hand it has turned out that some interesting and important (for various branches of mathematics) hypotheses of set theory, i.e., the axiom of choice and the continuum hypothesis, are independent of other accepted axioms for sets (K. Gödel proved in 1939 that the axiom of choice and the continuum hypothesis are consistent with the axioms of Zermelo-Fraenkel set theory and in 1963 P. Cohen proved that they are independent). Since set theory is a fundamental mathematical theory (in the sense explained above) this indicated that there is in fact no firm basis for mathematics fixed once and for ever, i.e., various set theories are possible (i.e., consistent) and can serve as a basis for mathematics. Which is the proper one? And what does it mean? Which axioms should and can be accepted in mathematics? What should decide of the acceptance or rejection of particular axioms? Which new axioms can be added to solve particular problems in mathematics, for example to solve the continuum hypothesis or the axiom of determinacy (which is inconsistent with the axiom of choice). What is the justification of axioms of large cardinals? Such problems were present in the philosophy of mathematics since its origins but now they are showed in a new light.

The indicated problems are especially pressing in the current set theory. Therefore we shall discuss them just on the example of this theory. One can distinguish three types of arguments used to justify the axioms: intrinsic, extrinsic and heuristic (those classes are not disjoint). The intrinsic arguments are based on the very notion of a set, they refer to the primitive intuition of a set (there is of course a problem what are the sources of this intuition). Extrinsic arguments are stated in terms of consequences or intertheoretic connections. In particular they refer to the fact that a considered axiom (or theorem) has been confirmed in special particular cases, that it implies unknown results of the lower level, that it provides new proofs of old results, that it enables a unification of new and old results in such a way that the old results become special instances of the new ones, that it enables to extend the patterns known for weaker theories, that it provides new strong methods of solving problems unsolved so far, that it enables us to solve various hypothesis or to establish some connections between theories. Heuristic arguments of justification are of an a priori character and are not uniform. They refer to various principles such as for example Cantorian finitism (infinite sets are similar to finite ones), limitation of size (there exist only sets which are not too "big" in comparison with sets already accepted), maximization (everything that can be a set, is a set), realism (based on the distinction between existence and definability, it rejects the restriction of existing sets only to definable ones), uniformity (the universe of sets is in fact uniform, i.e., the same properties and situations appear anew at higher levels).

3 CURRENT TRENDS IN THE PHILOSOPHY OF MATHEMATICS

The philosophy of mathematics after 1930 has been shaped by Gödel's incompleteness theorems and the consciousness of limitations of the axiomatic-deductive method revealed by them. It was characterized on the one hand by the dominance of the classical doctrines like logicism, intuitionism and formalism and on the other by the emergence of some new conceptions.

One should mention here Willard Van Orman Quine's holistic philosophy of mathematics and his indispensability argument according to which mathematics should be considered not in separation from other sciences but as an element of the collection of theories explaining the reality (cf. Quine 1951a, 1951b, 1953). Mathematics is indispensible there, in particular in physical theories, hence its objects do exist.

In this way Quine attacked the anti-realist and anti-empiricist approaches to the philosophy of mathematics. This cleared the way for empiricist approaches. One of them is the quasi-empiricism of Hilary Putnam who claimed (cf. Putnam 1975) that mathematical knowledge is not a priori, absolute and certain, that it is rather quasi-empirical, fallible and probable, much like natural sciences. He argues that quasi-empirical mathematics is logically possible and that ordinary mathematics has been quasi-empirical all along. In (1967) Putnam proposed a modal picture of mathematics according to which mathematics does not study any particular objects themselves but rather possibilities involving any objects whatsoever. Hence mathematics studies the consequences of axioms and asserts also the possibility of its axioms having models. The introduction of modalities opened the door to new epistemologies of mathematics.

Quine's-Putnam's indispensability argument was criticized by Hartry Field whose theory belongs to one of the most discussed proposals in the philosophy of mathematics in the recent years (cf. Field 1980 and 1989). Analysing the role of mathematics in the natural sciences, especially in physics, Field comes to the conclusion that it is not true that mathematics is indispensable in them and that science uses mathematics merely as a theoretically dispensable descriptive and inferential short-cut only. Mathematical objects play there another role than abstract theoretical objects. In fact the latter extend the purely observational theories while in theories using abstract mathematical objects one cannot prove more than in a theory which does not refer to such objects. In other words any statement that does not refer to mathematical objects, which is a consequence of a mathematical extension of Field-style theory, is already a consequence of the nonmathematical part of the theory. Field illustrated his programme of formulating versions of scientific theories that do not presuppose the existence of numbers and functions by developing an intrinsic version of Newtonian gravitation theory. This leads him to the nominalism – he claims that mathematics is only a useful auxiliary fiction, a set of propositions which enable us to formulate and to justify statements about the real world which itself has in fact no interpretation.

In contrast to Field, Charles Chihara and Philip Kitcher claim that natural sciences require something like the mathematical formalism to formulate and develop its theories. Chihara maintains (cf. Chihara 1990) that this formalism is not about mathematical objects but it concerns the possibility of taking open sentences.

Unfortunately he says little about the epistemology of those possibilities. Kitcher views mathematics as an idealizing theory – it describes how we would segregate, arrange and collect physical objects if we lived in an infinite world and had perfect memories, etc. In (1983) he attempts to show that the growth of mathematical knowledge is far more similar to the growth of scientific knowledge than is usually appreciated. He offers a picture of mathematical knowledge which rejects mathematical apriorism. It is shown how early mathematical theories described empirically based idealizations and how theory gave birth to the study of even more remote idealizations.

Some interesting philosophical ideas concerning mathematics can be also found by K. Gödel who formulated them especially in connection with some problems of set theory (cf. Gödel 1944 and 1947). His philosophy of mathematics can be characterized as Platonism. He claimed that mathematical objects exist in the reality independently of time, space and the knowing subject. He stressed the analogy between logic and mathematics on the one hand and natural sciences on the other. Mathematical objects are transcendental with respect to their representation in mathematical theories. The basic source of mathematical knowledge is intuition though it should not be understood as giving us the immediate knowledge. It suffices to explain and justify simple basic concepts and axioms. Mathematical knowledge is not the result of a passive contemplation of data given by intuition but a result of the activity of the mind which has a dynamic and cumulative character. Data provided by the intuition can be developed by a deeper study of mathematical objects and this can lead to the adoption of new statements as axioms.

In the eighties a naturalized version of Gödel's ideas has been developed by Penelope Maddy (cf. Maddy 1980, 1990a, 1990b). Gödel thought we can intuit abstract sets, Maddy claims that we can see sets of concrete objects whose members are before our eyes. We perceive sets of concrete physical objects by perceiving their elements (physical objects). Sets are located in the space-time real world. In this way we can know "simple" sets, i.e., hereditarily finite sets. More complicated sets are treated by Maddy as theoretical objects in physics – they and their properties can be known by metatheoretical considerations. One sees that in this conception Gödel's mathematical intuition has been replaced by the usual sensual perception. The advantage of Maddy's approach is that it unifies the advantages of Gödel's Platonism enabling us to explain the evidence of certain mathematical facts with Quine's realism taking into account the role that mathematics plays in scientific theories.

We must mention also Ludwig Wittgenstein (1889-1951). His ideas concerning mathematics can be reconstructed from his remarks made at various periods – hence they are not uniform, moreover they are even inconsistent (cf. Wittgenstein 1953, 1956). They grew out from his philosophy of language as a game. He was against logicism, and especially against Russell's attempts to reduce mathematics to logic. He claimed that by such reductions the creative character of a mathematical proof disappears. A mathematical proof cannot be reduced to axioms and rules of inference, because it is in fact a rule of constructing a new concept. Logic does not play so fundamental role in mathematics as logicism claims – its role is rather auxiliary. Mathematical knowledge is independent and specific in comparison with logic. Mathematical truths are a priori, synthetic and constructive. Mathematicians are not discovering mathematical objects and their properties but creating them. Hence mathematical knowledge is of a necessary character. One can easily see here the connections of Wittgenstein's philosophy of mathematics and Kant's ideas as well as the ideas of intuitionists.

In the sixties there appeared in the philosophy of mathematics a new anti-foundational tendency. It was the reaction to the limitations and one-sidedness of the classical views which are giving one-dimensional static picture of mathematics as a science and are trying to provide indubitable and infallible foundations for mathematics. They treat mathematics as a science in which one automatically and continuously collects true proved propositions. Hence they provide only one-sided reconstructions of the real mathematics in which neither the development of mathematics as a science nor the development of mathematical knowledge of a particular mathematician would be taken into account. New conceptions challenge the dogma of foundations and try to reexamine the actual research practices of mathematicians and those using mathematics and to avoid the reduction of mathematics to one dimension or aspect only. They want to consider mathematics not only in the context of justification but to take into account also the context of discovery.

One of the first attempts in this direction was the conception of Imre Lakatos (1922-1974). He attempted to apply some of Popper's ideas about the methods of natural science to episodes from the history of mathematics.[19]

Lakatos claims that mathematics is not an indubitable and infallible science – on the contrary, it is fallible. It is being developed by criticising and correcting former theories which never are free of vagueness and ambiguity. One tries to solve a problem by looking simultaneously for a proof and for a counterexample. New proofs explain old counterexamples, new counterexamples undermine old proofs. By proofs Lakatos means here usual non-formalized proofs of actual mathematics. In such a proof one uses explanations, justifications, elaborations which make the conjecture more plausible, more convincing. Lakatos does not analyse the idealized formal mathematics but the informal one actually developed by "normal" mathematicians, hence mathematics in process of growth and discovery. His main work *Proofs and Refutations* (1963-64) is in fact a critical examination of dogmatic theories in the philosophy of mathematics, in particular of logicism and formalism. Main objection raised by Lakatos is that they are not applicable to actual mathematics. Lakatos claims that mathematics is a science in Popper's sense, that it is developed by successive criticism and improvement of theories and by establishing new and rival theories. The role of "basic statements" and "potential falsifiers" is played in the case of formalized mathematical theories by informal theories (cf. Lakatos 1967).

Another attempt to overcome the limitations of the classical theories of philosophy of mathematics is the conception of Raymond L. Wilder (1896-1982). His main thesis says that mathematics is a cultural system.[20] Mathematics can be seen as a subculture, mathematical knowledge belongs to the cultural tradition of a society, mathematical research practice has a social character. Thanks to such an approach the development of mathematics can be better understood and the general laws of changes in a given culture can be applied in historical and philosophical investigations of mathematics. It also enables us to see the interrelations and influences of various elements of the culture and to study their influence on the evolution of mathematics. It makes also possible to discover the mechanisms of the development and evolution of mathematics. Wilder's conception is therefore sometimes called evolutionary epistemology. He has proposed to study mathematics not only from the point of view of logic but also using methods of anthropology, sociology and history. Wilder maintains that mathematical concepts should be located in the Poppers "third world". Mathematics investigates no timeless and spaceless entities. It cannot be understood properly without regarding the culture in the framework of which it is being developed. In this sense mathematics shares many common features with ideology, religion and art. A difference between them is that mathematics is science in which one justifies theorems by providing logical proofs and not on the basis of, say, general acceptance.

Those new anti-foundational trends in the philosophy of mathematics should not be treated as competitive with respect to old theories. They should be rather seen as complements of logicism, intuitionism and formalism. One is looking here not for indubitable, unquestionable and irrefutable foundations of mathematics, one tries not to demonstrate that the actual mathematics can be reconstructed as an infallible and consistent system but one attempts to describe the actual process of building and constructing mathematics (both in individual and historical aspect).

Considering new conceptions in the philosophy of mathematics one must also mention structuralism. It can be characterized as a doctrine claiming that mathematics studies structures and that mathematical objects are featureless positions in these structures. As forerunners of such views one can see R. Dedekind, D. Hilbert, P. Bernays and N. Bourbaki. The latter is in fact a pseudonym of a group of (mainly French) mathematicians who undertook in the thirties the task of a systematization of the whole of mathematics (the result of their investigations was the series of books under the common title *Éléments de mathématique).* Their work refered to Russell's idea of reconstructing mathematics as one system developed on a firm (logical) basis. For bourbakists the mathematical world is the world of structures. The very notion of a structure was explained by them in terms of set theory. They distinguished three principal types of mathematical structures: algebraic, order and topological structures.

The idea of treating mathematics as a science about structures is being developed nowadays by Michael Resnik, Stewart Shapiro and Geoffrey Hellman. Resnik claims (cf. Resnik 1981, 1982) that mathematics can be viewed as a science of patterns with its objects being positions in patterns. The identity of mathematical objects is determined by their relationships to other positions in the given structure to which they belong. He does not postulate a special mental faculty used to acquire knowledge of patterns (they are not seen through a mind's eyes). We go through a series of stages during which we conceptualize our experience in successively more abstract terms. This process do not necessarily yield necessary truths or a priori knowledge. Important is here our tendency to perceive things as structured. The transition from experience to abstract structures depends upon the culture in which it takes place. Add that the transition from simple patterns to more complicated ones and the development of pure theories of patterns rely upon deductive evidence.

S. Shapiro claims (cf. Shapiro 1989, 1991) that there is a strict connection between objects of mathematics and objects of natural sciences. An explanation of it can be provided in his opinion just by structuralism according to which mathematics studies not objects *per se* but structures of objects. Hence objects of mathematics are only "places in a structure". The advantage of such an approach is that it enables us to explain the phenomenon of applicability of abstract mathematical theories in natural sciences as well as the interrelations between various domains of mathematics itself. It enables also a holistic approach to mathematics and science.

G. Hellman argues (cf. Hellman 1989) that one can interpret mathematics (in particular arithmetic and analysis) as nominalistic theories concerned with certain logically possible ways of structuring concrete objects. He uses by such interpretations second-order logic and modal operators (hence his approach is sometimes called modal-structural).

4 CONCLUSIONS

The above presentation of conceptions in the epistemology of mathematics indicates that there were various proposals and attempts to answer the basic questions concerning the epistemological status and the methods of mathematics as a science. There is no unique answer accepted by all philosophers of mathematics (but the role of philosophy is not to give definite answers but rather to indicate problems and show the complexity of considered issues). On the other hand mathematics is dynamic and is being developed rather independently of philosophical settlements of questions. But of course this independence is not complete. The role of the philosophy of mathematics towards mathematics itself is not only descriptive but also normative, i.e., some philosophical conceptions and solutions fix certain norms and rules according to which mathematical knowledge should be (and in fact is) developed and presented.

Those norms and rules are being changed and transformed of course. For example the ideal basic method to develop mathematics since the Greek antiquity was considered to be the axiomatic-deductive method and the basic method to justify a statement was to give a proof. But the very concept of a proof has been changed very deeply from the intuitive one in which a reference to drawings and "self-evident" facts were allowed to the precise notion of a formal proof in a formalized axiomatic system. Also the idea of what is the nature of an axiom has changed very much. On the other hand methods of mathematical logic enabled us to discover several limitations of the axiomatic method and simultaneously to make precise various philosophical concepts and ideas (as for example to distinguish in mathematics between truth and provability).

The development of mathematics leads not only to the formulation of new problems and questions in the philosophy of mathematics but also to the necessity of revising the former conceptions. As an example can serve here the construction of non-Euclidean geometries in the ninteenth century. Nowadays the most spectacular examples are connected with computers and their applications.

Computers are used in mathematics not only to perform complicated (and tedious) numerical calculations but also in automated theorem proving. Studies on the mechanization and automatization of (mathematical) reasonings can be traced back to the seventeenth century, to Leibniz and his idea of the *characteristica* *universalis*. They received an important impulse from the mathematical logic on the turn of the nineteenth century. Recently the possibility of realization of those method on computers brings new contexts. The main question one should ask in the connection with this is: what are the reasons for accepting mathematical results obtained by using a computer. One of the ways to verify such results is to perform the given computer programme several times on various machines and to check whether the results are identical. But note that this procedure is similar to the procedure of verifying experimental data in physics and is in fact quasi-empirical. So can it be used in mathematics?

Recently there appeared some results in mathematics in which computers were essentially applied. The most spectacular and most discussed example is the four-color theorem being a solution of an old problem concerning the coloring of a map. In the proof of this theorem computer calculations are heavily used. What more, computer was applied here not only to perform some computations but some important tricks and ideas used in the proof were improved by certain computer experiments, by "dialogues" with a computer. The validity of the computer programme cannot be checked without a computer. On the other hand no traditional proof (i.e., a proof not referring to computers) of the four-color theorem has been given (there are doubts whether such a proof can be given). Hence the considered theorem is the first example of a mathematical theorem of a new type. Its proof is convincing and can be formalized but is not surveyable, so it has not one of the important features mathematical proofs should (traditionally) have. Consequently one can ask whether the four-color theorem has been proved and whether it can be considered as a mathematical theorem and whether it belongs to the mathematical knowledge? Certainly it is not an a priori statement. There are two possibilities: either extend the scope of methods accepted in mathematics and to allow the usage of computers (hence a type of experiments) or to admit that the four-color theorem has not been proved yet and does not belong to mathematical knowledge. The former possibility implies in particular that mathematics becomes a quasi-empirical (and not an a priori) knowledge. Such solution is represented by Ph.J. Davis, R. Hersh, Ph. Kitcher and E.R. Swart who claim that mathematics always admitted empirical elements and had in fact an empirical character. On the other hand one attempts to defend the a priori character of mathematics by arguing that proofs using computers can be transformed into traditional proofs by adding new axioms or that a computer is in fact a mathematician and it knows the result proved deductively or that procedures similar to applying computer programmes have been used in mathematics for a long time, hence the applications used in the proof of the four-color theorem are in fact nothing essentially new.

Discussing here the problem of the influence of computers on the philosophy of mathematics one should mention also questions connected with the old mind-body problem, in particular with the problem whether machines can act in an intelligent way and the whole scope of problems formulated in the domain called artificial intelligence. They are not directly connected with the philosophy of mathematics – therefore we will not discuss them here. Note only that Gödel's incompleteness theorems are also used in the study of them. In particular it has been argued (cf. Chaitin 1974, 1982) that Gödel's theorems (when interpreted from the point of view of the information theory) show that if one wants to obtain more complex mathematical theorems (i.e., theorems containing more information) then one will have to continually introduce new axioms and new methods. Neither the admissible methods and rules can be fixed and codified nor the concept of a correct mathematical proof can be defined once and for ever. Hence progress in mathematics seems to be much like the progress in the natural sciences than hitherto expected. All such claims provide new arguments for the quasi-empiricism claiming that mathematics is in fact much like natural sciences.

Roman Murawski
Adam Mickiewicz University

NOTES

* The author was supported by the Committee for Scientific Research, grant no. 1 H01A 023 13.

[1] Cf. L. Kolakowski, Zawod blazna jest mi blizszy. Z Leszkiem Kolakowskim rozmawia (korespondencyjnie) Pawel Spiewak, *Res Publica* 9 (1988), 30.

[2] According to Aristotle, Plato distinguished between the arithmetical and geometrical ideas (forms) and the so-called mathematicals, each of which is an instance of some unique form – each form having many such instances.

[3] He wrote: 'And when this comes [i.e., when the idea of universal language is realized – R.M.] then two philosophers wanting to decide something will proceed as two calculators do. It will be enough for them to take pencils, go to their tablets and say: *Calculemus!* (Let us calculate!) (cf. G.W. Leibniz, *Philosophische Schriften,* ed. C.I. Gerhardt, vol. 7, Berlin 1890, pp. 198-201).

[4] Kant distinguished twelve categories divided into four sets of three: (1) of quantity: unity, plurality, totality; (2) of quality: reality, negation, limitation; (3) of relation: substance-and-accident, cause-and-effect, reciprocity; (4) of modality: possibility, existence, necessity.

[5] It is usually called the second crisis. As the first crisis one means the discovery of the incommensurable magnitudes by the Pythagoreans in the ancient Greece. This led to the change of the notion of a number and to replacing arithmetic by a geometrical algebra. Sometimes one adds here also the seventeenth century crisis connected with the development of the differential and integral calculus by W.G. Leibniz and I. Newton (then the crisis of the nineteenth century is called the third crisis). In fact basic notions of this calculus such as, e.g., the notion of a differential (an infinitely small magnitude) in the form introduced by Leibniz, were simply inconsistent. Nevertheless the calculus has been successfully developed and applied. Only in the twentieth century the consistent basis for the Leibniz's calculus has been developed by the nonstandard analysis of A. Robinson. This example indicates that mathematical theories can be (and have been) often successfully developed and applied without a satisfactory consistent basis and that such a basis has been provided many years later.

[6] One should note here that Frege treated geometry in a different way than arithmetic. In fact he claimed that geometry is synthetic (and not analytic as arithmetic) because it says about one particular domain and that it is a priori because its axioms do not need to be proved.

[7] It is today formulated in the following way: Given a set X one can ask whether it is its own element or not. So consider a set Z of all such sets X that X is not its own element, i.e., $Z = \{X : X \in X\}$. What are the properties of the set Z, in particular is Z its own element or not. If one answers YES, i.e., $Z \in Z$, then $Z \notin Z$ because in Z are only sets being not their own elements. On the contrary, if the answer is NO, i.e., $Z \notin Z$, then $Z \in Z$ because Z does not

have the property of the elements of Z. Consequently $Z \in Z$ if and only if $Z \notin Z$ which is a contradiction.

[8] One should note here that Russell's views evoluated. In particular before 1910 he claimed that logic and mathematics are synthetic. The thesis about the analyticity of mathematics has been proclaimed by him since 1910, i.e., since the publication of *Principia Mathematica.*

[9] The restriction of mathematics to algebra and analysis was the consequence of a thesis (Kronecker refered here to C.F. Gauss) that, for example, geometry or mechanics are independent of human mind because they refer to the external reality.

[10] The integer numbers were made by God, everything else is the work of man.

[11] He developed them in his inaugural lecture at the University of Amsterdam *Intuitionisme en formalisme* (1912) and in the paper *Consciousness, Philosophy and Mathematics* (1949).

[12] Later various radical versions of formalism appeared, in particular the so called strict formalism of Haskel B. Curry (cf. Curry 1951). Mathematics was reduced in it to the study of formalized theories and nothing was assumed except the symbols constituing a given system.

[13] The undecidable arithmetical sentence constructed by Gödel in his proof of the incompleteness theorem had a metamathematical contents rather than mathematical (it stated: "I am not a theorem"). Though interesting for logicians it was rather artificial from the mathematical point of view. Hence one could still charish hopes that all sentences which are interesting and reasonable from the mathematical point of view (whatever it means) are decidable and that in the domain of such sentences the attempts to make precise the notion of a mathematical theory and a mathematical proof by using (first order) formal theories are successful. They were shuttered by results of J. Paris, L. Harrington and L. Kirby (1979-1982) indicating examples of undecidable sentences about natural numbers of the directly mathematical (in fact combinatorial or numbertheoretic) contents (cf. Paris-Harrington 1977 and Kirby-Paris 1982; see also Murawski 1994).

[14] It should be noted here that this is only a rough formulation of Gödel's theorem on the unprovability of consistency. One must take here into account also the way in which the metamathematical notion of consistency of a given theory has been formalized.

[15] The ambiguous notion of being effective has been made precise by means of the theory of recursive (computable) functions.

[16] From the philosophical point of view reverse mathematics is a reductionist programme. Its main aim is to study the role of the comprehension axiom (the axiom on the existence of sets) in the mathematics. In particular one considers there a problem which forms of the comprehension axiom are necessary and sufficient to prove various particular theorems of analysis, algebra, topology, etc. Detailed description of the results of the reverse mathematics and of their meaning for the philosophy of mathematics is given in (Murawski 1994) (one can also find there an extensive bibliography).

[17] It does not mean that axioms of mathematical theories were fixed once and for ever. On the contrary, axiomatizations of theories are being developed. We mean here that in proving theorems one can use axioms and only axioms and it is not allowed to apply for example drawings or so called evident facts.

[18] The concept of a consequence was defined by A. Tarski. To the process of formulating a precise definition of the concept of a proof contributed G. Frege, B. Russell, A.N. Whitehead, D. Hilbert, P. Bernays, W. Ackermann, S. Jaskowski and G. Gentzen.

[19] The reference to the history of mathematics is one of the characteristic features of new trends in the philosophy of mathematics.

[20] Wilder presented his ideas in many papers and lectures. A complete version of them can be found in two of his books: *Evolution of Mathematical Concepts. An Elementary Study* (1968) and *Mathematics as a Cultural System* (1981).

REFERENCES

A. Anthologies of texts:

Benacerraf, P. and H. Putnam (eds.): 1964, *Philosophy of Mathematics. Selected Readings*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (2nd edition: Cambridge University Press, Cambridge 1983).

Hart, W. D. (ed.): 1996, *The Philosophy of Mathematics*, Oxford University Press, Oxford.

Hintikka, J. (ed.): 1969, *The Philosophy of Matematics*, Oxford University Press, Oxford.

Murawski, R. (ed.): 1986, *Philosophy of Mathematics. Anthology of Classical Texts* (in Polish), Wydawnictwo Naukowe UAM, Poznań (second revised edition: 1994).

Resnik, M.D. (ed.): 1995, *Mathematical Objects and Mathematical Knowledge*, Dartmouth, Aldershot-Brookfield, USA, Singapore, Sydney.

Thiel, Ch. (ed.): 1982, *Erkenntnistheoretische Grundlagen der Mathematik*, Gerstenbeg Verlag, Hildesheim.

Tymoczko, T. (ed.): 1985, *New Directions in the Philosophy of Mathematics*, Birkhäuser, Boston, Basel, Stuttgart.

B. Original papers:

Aristotle: 1960, *Aristotelis opera ex recensione Immanuelis Bekkeri edidit Academia Regia Borussica.* Editio altera, Hrsg. O. Gigon, Walter de Gruyter, Berlin.

Brouwer, L. E. J.: 1907, *Over de Grondslagen der Wiskunde*, Maas en van Suchtelen, Amsterdam.

Brouwer, L. E. J.: 1912, *Intuitionisme en formalisme*, Noordhoff, Groningen; also in *Wiskundig Tijdschrift* **9** (1912), 180-211; English translation: 'Intuitionism and Formalism', *Bulletin of the American Mathematical Society* **20** (1913), 81-96.

Brouwer, L. E. J.: 1948, 'Consciousness, Philosophy and Mathematics', in E. W. Beth, H. J. Pos, and H. J. A. Hollak (eds.), *Library of the Tenth International Congress in Philosophy*, vol. 1, North-Holland Publ. Comp., Amsterdam, pp. 1235-1249.

Chaitin, G.: 1974, 'Information-Theoretic Computational Complexity', *IEEE Transactions on Information Theory* **IT-20**, 10-15.

Chaitin, G.: 1982, 'Gödel's Theorem and Information', *International Journal of Theoretical Physics* **21**, 941-954.

Chihara, Ch. S.: 1990, *Constructibility and Mathematical Existence*, Clarendon Press, Oxford.

Couturat, L.: 1901, *La logique de Leibniz d'apres des documents inédits*, Alcan, Paris.

Curry, H. B.: 1951, *Outlines of a Formalist Philosophy of Mathematics*, North-Holland Publ. Comp., Amsterdam.

Dedekind, R.: 1872, *Stetigkeit und irrationale Zahlen*, Friedrich Vieweg und Sohn, Braunschweig.

Descartes, R.: 1637, *Discours de la methode. Pour bien conduire sa raison, & chercher la verité dans les sciences. Plus la Dioptrique. Les Meteores. Et la Geometrie. Qui sont des essais de cete Methode*, Ian Marie, Leyden.

Euclid: 1883-1886, *Euclidis Elementa*, vol. I-IV, Edidit I.L. Heibeg, Teubner, Lipsiae.

Euclid: 1956, *The Thirteen Books of Euclidis Elements*, translated from the text of Heiberg, with Introduction and Commentary by Sir Th. L. Heath, Dover Publishers, Inc., New York.

Field, H.: 1980, *Science without Numbers*, Basil Blackwell, Oxford.

Field, H.: 1989, *Realism, Mathematics, and Modality*, Basil Blackwell, Oxford.

Frege, G.: 1879, *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens*, Louis Nebert, Halle.

Frege, G.: 1884, *Die Grundlagen der Arithmetik. Eine logisch mathematische Untersuchung über den Begriff der Zahl*, Wilhelm Koeber, Breslau.

Frege, G.: 1893, *Grundgesetze der Arithmetik. Begriffsschriftlich abgeleitet*, Bd. 1, Hermann Pohle, Jena.

Frege, G.: 1903, *Grundgesetze der Arithmetik. Begriffsschriftlich abgeleitet*, Bd. 2, Hermann Pohle, Jena.

Gödel, K.: 1931, 'Über formal unentscheidbare Sätze der 'Principia Mathematica' und verwandter Systeme'. *I', Monatshefte für Mathematik und Physik* **38**, 173-198; reprinted with English translation 'On Formally Undecidable Propositions of *Principia Mathematica* and Related Systems' in K. Gödel, *Collected Works*, vol. I, S. Feferman et al. (eds.), Oxford University Press, New York and Clarendon Press, Oxford, 1986, pp. 144-195.

Gödel, K.: 1944, 'Russel's Mathematical Logic', in P. A. Schilpp (ed.), *The Philosophy of Bertrand Russell*, Northwestern University, Evanston, pp. 123-153; reprinted in K. Gödel, *Collected Works*, vol. II, S. Feferman et al. (eds.), Oxford University Press, New York and Oxford, 1990, pp. 119-141.

Gödel, K.: 1947, 'What is Cantor's Continuum Problem?', *The American Mathematical Monthly* **54**, 515-525; second revised version in P. Benacerraf and H. Putnam, *Philosophy of Mathematics. Selected Readings*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1964, pp. 258-273; reprinted in K. Gödel, *Collected Works*, vol. II, S. Feferman et al. (eds.), Oxford University Press, New York and Oxford, 1990, pp. 176-187.

Hellman, G.: 1989, *Mathematics without Numbers. Towards a Modal-Structural Interpretation*, Clarendon Press, Oxford.

Hilbert, D.: 1899, *Grundlagen der Geometrie. Festschrift zur Feier der Enthüllung des Gauss-Weber-Denkmals*, B.G. Teubner, Leipzig, pp. 3-92.

Hilbert, D.: 1901, 'Mathematische Probleme', *Archiv der Mathematik und Physik* **1**, 44-63 and 213-237. Reprinted in D. Hilbert, *Gesammelte Abhandlungen*, Verlag von Julius Springer, Berlin, Bd. 3, pp. 290-329; English translation: 'Mathematical Problems', in F. Browder (ed.), *Mathematical Developments Arising from Hilbert's Problems*, Proceedings of the Symposia in Pure Mathematics 28, American Mathematical Society, Providence, RI, 1976, pp. 1-34.

Hilbert, D.: 1926, 'Über das Unendliche', *Mathematische Annalen* **95**, 161-190; English translation: 'On the Infinite', in J. van Heijenoort (ed.), *From Frege to Gödel. A Source Book in Mathematical Logic*, 1879-1931, Harvard University Press, Cambridge, Mass., 1967, pp. 367-392.

Kant, I.: 1781, *Critik der reinen Vernunft*, Johann Friedrich Hartknoch, Riga.

Kirby, L. and J. Paris: 1982, 'Accessible Independence Results for Peano Arithmetic', *Bulletin of London Mathematical Society* **14**, 285-293.

Kitcher, Ph.: 1983, *The Nature of Mathematical Knowledge*, Oxford University Press, New York-Oxford.

Kronecker, L.: 1887, 'Über den Zahlbegriff', *J. Reine Angewandte Mathematik* **101**, 157-177.

Lakatos, I.: 1963-64, 'Proofs and Refutations. The Logic of Mathematical Discovery', *British Journal for the Philosophy of Science* **14**; as a book: Cambridge 1976.

Lakatos, I.: 1967, 'A Renaissance of the Empiricism in the Recent Philosophy of Mathematics?', in I. Lakatos (ed.), *Problems in the Philosophy of Mathematics*, North-Holland Publ. Comp., Amsterdam 1967, pp. 199-202; extended version in I. Lakatos, *Philosophical Papers, vol. 2: Mathematics, Science and Epistemology*, J. Worall and G. Currie (eds.), Cambridge University Press, Cambridge-London-New York-Melbourne, 1978, pp. 24-42.

Leibniz, G. W.: 1875-1890, *Philosophische Schriften*, Hrsg. C. I. Gerhardt, 7 volumes, Weidmannsche Buchhandlung, Berlin.

Maddy, P.: 1980, 'Perception and Mathematical Intuition', *Philosophical Review* **89**, 163-196.

Maddy, P.: 1990a, *Realism in Mathematics*, Clarendon Press, Oxford.

Maddy, P.: 1990b, 'Physicalistic Platonism', in A. D. Irvive (ed.), *Physicalism in Mathematics*, Kluwer Academic Publishers, Dordrecht, pp. 259-289.

Murawski, R.: 1994, 'Hilbert's Program: Incompleteness Theorems vs. Partial Realizations', in J. Woleński (ed.), *Philosophical Logic in Poland*, Kluwer Academic Publishers, Dordrecht, Boston, London, pp. 103-127.

Paris, J. and L. Harrington: 1977, 'A Mathematical Incompleteness in Peano Arithmetic', in J. Barwise (ed.), *Handbook of Mathematical Logic*, North-Holland Publ. Comp., Amsterdam, pp. 1133-1142.

Pasch, M.: 1882, *Vorlesungen über neuere Geometrie*, B.G. Teubner, Leipzig und Berlin.

Poincaré, H.: 1902, *La science et l'hypotèse*, Flammarion, Paris.

Proclus: 1873, *Procli Diadochi in Primum Elementorum Librum Comentarii*, ed. G. Friedlein, B.G. Teubner, Leipzig (reprinted: G. Olms, Hildesheim 1967).

Putnam H.: 1967, 'Mathematics Without Foundations', *Journal of Philosophy* **64**, 5-22. Revised version in H. Putnam, *Mathematics, Matter and Method. Philosophical Papers, vol. I*, Cambridge University Press, Cambridge, London, New York, Melbourne, pp. 43-59.

Putnam H.: 1975, 'What Is Mathematical Truth?'. In: Putnam, H., *Mathematics, Matter and Method. Philosophical Papers, vol. I*, Cambridge University Press, Cambridge, London, New York, Melbourne, pp. 60-78.

Quine, W. V. O.: 1951a, 'Two Dogmas of Empiricism', *Philosophical Review* **60/1**, 20-43; also in W. V. O. Quine, *From a Logical Point of View*, Harvard University Press, Cambridge, Mass., pp. 20-46.

Quine, W. V. O.: 1951b, 'On Carnap's Views on Ontology', *Philosophical Studies* **2**.

Quine, W. V. O.: 1953, 'On What There Is', in W. V. O. Quine, *From a Logical Point of View*, Harvard University Press, Cambridge, Mass., pp. 1-19.

Resnik, M. D.: 1981, 'Mathematics as a Science of Patterns: Ontology and Reference', *Nous* **15**, 529-550.

Resnik, M. D.: 1982, 'Mathematics as a Science of Patterns: Epistemology', *Nous* **16**, 95-105.

Russell, B.: 1903, *The Principles of Mathematics*, The University Press, Cambridge.

Russell, B.: 1919, *Introduction to Mathematical Philosophy*, George Alien & Unwin, Ltd., London, and The Macmillan Co., New York.

Shapiro, S.: 1989, 'Structure and Ontology', *Philosophical Topics* **17**, 145-171. Shapiro, S.: 1991, *Foundations without Foundationalism*, Clarendon Press, Oxford.

Tarski, A.: 1933, *Pojęcie prawdy w językach nauk dedukcyjnych* (The Notion of Truth in Languages of Deductive Sciences), Nakladem Towarzystwa Naukowego Warszawsklego, Warszawa; English translation: 'The Concept of Truth in Formalized Languages', in *Logic, Semantics, Metamathematics. Papers From 1923 To 1938*, Clarendon Press, Oxford 1965, pp. 152-278.

Whitehead, A. N. and B. Russell: 1910-1913, *Principia Mathematica*, 3 volumes, Cambridge University Press, Cambridge.

Wilder R. L.: 1968, *Evolution of Mathematical Concepts. An Elementary Study*, John Wiley & Sons, New York.

Wilder R. L.: 1981, *Mathematics as a Cultural System*, Pergamon Press, Oxford.

Wittgenstein, L.: 1953, *Philosophical Investigations*, Basil Blackwell, Oxford.

Wittgenstein, L.: 1956, *Remarks on the Foundations of Mathematics*, Basil Blackwell, Oxford.

JOSEPH MARGOLIS

KNOWLEDGE IN THE HUMANITIES AND SOCIAL SCIENCES

1. PREAMBLE

The history of the theory of knowledge confronts us with what appears to be a perpetual "frontier" mentality. No matter how exhaustive or ramified its previous philosophical labors may have been, it seems forever bent on testing the need for still another beginning. That is as true today as it ever was during the period of nearly constant innovation running from Descartes to Kant to Hegel. You have only to think of the startling frequency with which theorists continue to believe themselves to be initiating entirely new beginnings or, finally, to be correcting the hopeless conceptual errors and inadequacies of all past canons. Think, for instance, of Edmund Husserl's *Cartesian Meditations* (1960) or W.V. Quine's "Epistemology Naturalized" (1969); or, more adventurously, Michel Foucault's Nietzscheanized genealogies (1977) or Paul Churchland's would-be elimination of the entire "folk" conception of epistemology (1989).

In spite of such scatter, the history of philosophy conveys an almost impregnable impression of orderly advance and canonical assurance. It is probably closer to the truth, however, that every would-be intervention is matched by its own fresh summary of the import of the gathering history of epistemology. If so, then, of course, the comparative assessment of competing theories of knowledge cannot fail to be more complex, more fraught with incommensurabilities, than one might wish.

No area of specialized inquiry into the nature and conditions of knowledge is likely to be more profoundly affected by such vagaries than that of theorizing about the standing and conditions of knowledge in the humanities and social sciences. In fact, it may be fairly argued that the point of such theorizing is, precisely, to test what may be most convincing in the way of the epistemic interrelationship between the world of human culture and the world of physical nature.

Up to the present time, with the single large exception of the post-Kantian tradition and its progeny, the theory of knowledge has, in the modern era, almost always supposed that the conditions of knowledge in matters cultural are essentially the samė as, and therefore rightly guided by, those judged to have proved successful in the exemplary physical sciences. That sense of priority and order and unity still counts as the somewhat uneasy canon of epistemology at the close of the century. It was challenged of course in the Kantian spirit, unsatisfactorily yet promisingly, by Wilhelm Dilthey (1989) and more radically, hence more contentiously, by Karl Marx, Friedrich Nietzsche (see Löwith 1991), and Martin Heidegger (1962).

One may indeed be drawn as a partisan to the seminal intuitions of one or another of these last figures. But it is also possible to construct a reasonably stable and neutral history of modern epistemology that attempts (even if not altogether

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 607-645.
© 2004 *Kluwer Academic Publishers.*

uncontentiously) to explain the conceptual relationship between our knowledge of physical nature and our knowledge of human culture. That may well be the sparest and most useful epistemological inquiry to carry into the next century.

Provisionally, then, the following may be offered as a pointed summary of the best gains of the entire history of epistemology leading up to the turn into the new century: first, the exposé of the insuperable paradoxes or skepticisms belonging to the original "Cartesian" tradition that spans the work of Descartes and Kant right up to the first *Critique*, resolved not without acknowledging the inseparability of ontic and epistemic distinctions; second, the replacement of all the forms of methodological solipsism, notably those of subjective representationalism, through the same interval, flagged, just prior to the publication of the first *Critique*, in Kant's well-known letter to Marcus Herz (February 21, 1772), resolved principally in the post-Kantian period by acknowledging the socially formed and socially shared nature of our cognizing powers; and, third, the dawning discovery that if the first two gains be granted, there cannot be a principled disjunction between the conditions of knowing nature and of knowing the human or cultural world. But though these lessons are grasped in the abstract, the cultural and historical (human) world remains to this day the least explored sector of reality from the time of the Hegelian and post-Hegelian gains down to our own time.

As far as our own age is concerned, the analysis of knowledge in the humanities and the social sciences is probably the most problematic that could be named and, among analytic practitioners, the most neglected. It is now also increasingly contested, since the older canon, which regularly subordinates knowledge of cultural phenomena to knowledge of physical nature, has strengthened its primacy in academic circles, despite having been beleaguered in recent decades. Yet it would be difficult to argue that the "naturalizing" strategy that currently dominates the "canon" has ever demonstrated that it could actually meet the strongest, most interesting challenges that could be drawn from the human studies. In any case, that is surely one of the principal issues that we must inquire into.

In Anglo-American philosophy, for instance, which has taken a leading role in reclaiming the canonical picture, the actual effort at recovery has been marked by a number of odd lapses that confirm that we are still perseverating at one or another of the earlier phases of the history just sketched: for example, in the insistent disjunction of metaphysics and epistemology (Devitt 1991); or the recovery of a regulative norm of objectivity, much as in Kant, linking the "subjective" and "objective" sides of cognition even where they are thought to be inseparable (*see* Putnam 1987); or in the retreat from the contingent social and cultural formation of our cognizing powers, as in restricting Hegel's innovation in Kant's terms, if not in Aristotle's (as in McDowell 1996). It is hard to see the importance of the cultural contribution if we cannot count on surpassing such faults. The old canon's hegemony is bound to appear more problematic wherever the puzzles about the human sciences begin to mount.

These tendencies, hardly isolated, suggest the need to test the gains the earlier history vouchsafes (if indeed it confirms anything). Otherwise, if it challenges the familiar canon, the analysis of the humanities and social sciences may appear more arbitrary or opportunistic than it actually is. If, now, we hold to the history sketched, we may (or must) admit the prominence of a set of conceptual distinctions that are hardly salient in the interval from the beginnings of Greek philosophy to about the time of the French Revolution, or are no longer favored in the same committed spirit that marks the interval from the French Revolution to about the collapse of the Soviet empire.

Three of the most important of these features, whose acceptance would redirect the focus of analysis in the humanities and social sciences, include the following: (i) the appearance of *sui generis* entities in the cultural world exhibiting attributes that cannot be found elsewhere in nature, in particular, attributes that, by a term of art, may be called *Intentional*; second, the recognition that the entities and *denotata* of the cultural world, preeminently *selves* and their "utterances" (nominalized as speech and language, actions and histories, artworks and texts and machines), often called *artifactual* or "culturally emergent" (*see* Margolis 1984), are indissolubly "embodied" in physical, biological, or electronic materiae; and, third, the appreciation that human thought, human perception and experience, is, as enlanguaged and socially formed, *historicized* (*see* Foucault 1979, Gadamer 1975), that is, subject, through formation and use, to the changing history of the enabling society in which they are formed as they are.

In a fair sense, these themes may be abstracted from the work of that important group of thinkers, wedded to historicity, that runs from Hegel through Marx through Nietzsche through Dilthey down to Heidegger and Gadamer and Foucault. From the "canonical" viewpoint, any analysis favoring these themes would be judged to be distinctly heterodox. There is, in fact, almost no influential analytic philosopher in our time who has favored historicity or the implications of doing so. That is already extraordinary. It colors our sense of the most important contributions to the topic before us.

It also serves to mark a philosophical contest for the opening phase of the new century. Certainly, none of the themes mentioned is featured by the "older canon." It is not the concern of this review, however, to establish these doctrines over their rivals. The point is, rather, to gain an appreciation of how far the analysis of knowledge in the humanities and social sciences may diverge from the exemplars drawn from the physical sciences, without risking incoherence or incompatibility with the work of the latter, and of whether what may be learned on the way may even strengthen our account of the physical sciences.

That seems temperate enough as an undertaking: it defines the principal disputes the topic is likely to generate; and it raises the question of reconciling the natural sciences and the human studies – say, somewhere between reductionism or "naturalizing" (following Quine 1969 or, more recently, Davidson 1986) and the recognition that the human studies (and the world they explore) are "natural" though not "naturalizable" in the sense in which the physical sciences count as the paradigms of naturalism.

Naturalism (in the sense suggested) is the leading theme of the contest between those who favor the analytic "canon" and those who construe the human studies (both our cognizing powers and the phenomena they address) in terms of the three doctrines just collected. The contest has actually been before us for more than two hundred years, but it has remained noticeably undeveloped on both sides of the quarrel.

Beyond that and admitting divergent philosophical loyalties, it should be clear that there cannot be any assured unitary line of analysis, though, plainly, there are partisans enough on either side: for example, in recent years, Rudner (1966) on the canonical side, and, in a more sociologized spirit, Shapin (1996) and Kuhn (1970) on the historicist side. But even exemplars of these sorts obscure the sheer complexity of the underlying issues.

Apart, then, from personal loyalties, if reductionism or naturalism succeeded, there would be very little to distinguish the work of the humanities and social sciences. "All" that would remain would be "benign" questions regarding meaning. Intentionality, history, and the like, thereafter judged incapable of generating serious epistemological difficulties. But if naturalizing remained uncertain, the decisive questions would be these: one, whether the resources of language – more particularly, those regarding reference and predication and related competences – significantly color our cognitive claims in the natural sciences, so that *their* objective standing cannot rightly be disjoined from that of the human studies; another, what special problems of objectivity arise in the humanities and social sciences beyond the usual puzzles that arise in the other.

Reasonable answers may justify a moderate, but not insignificant, departure from the naturalizing canon, without yet risking any arbitrarily extreme conception. The truth is, all sides would be well served: the answers would identify what the canon must overcome to reclaim its supposed inclusiveness; and its opponents would begin to understand what might count in their favor by way of the least contestable concessions.

The fact is, the very idea that, for example, the methodology of sociology and history *is not* intrinsically the same as the methodology of the natural sciences still conveys a sense of immense shock to the champions of the "older" canon, though they have heard the charge before. It's hardly more than a single sounding of the larger themes favored by positivism and the unity of science program that, by now, have been completely dismantled for reasons internal to their own undertaking. Their commitment has been replaced, however, by a much leaner and more resilient "naturalism," also committed to the sufficiency of causal explanation and description cast in physicalist terms – or at least in terms that entertain the reduction or elimination of all intentional discourse (in the sense of "intentional" developed from Brentano's (1973) original reinterpretation of the Cartesian distinction between mind and body), which, as already remarked, has been marginalized by theorists like Quine and Davidson. (Brentano's distinction between the "mental" and the "physical," it should be said, does not correspond at all to Descartes' distinction between mind and body.)

You may appreciate the shift in idiom and the change of focus if you compare the programmatic arguments of Carnap (1995), Popper (1983), Hempel (1965), Oppenheim and Putnam (1958) with the more specialized reclamations of the "naturalizing" sort, found nowadays in the philosophy of mind more often than in the philosophy of science, for instance in the work of Fodor (1990) and Dretske (1995). The essential point remains the same, however: the distinctive features of the psychological and cultural world are still neutralized in the service of the same inclusive physicalism as before. Hence, the analytic appraisal of the prospects of knowledge in the humanities and the social sciences remains, through most of the twentieth century, remarkably constant. But we do now see more clearly than before that the issue hangs on the fate of the naturalizing maneuver. There is not likely to be any deeper "canon" of the unity sort than naturalizing.

From the side of the theory of the natural sciences, the most surprising, the single most important, internal development has been the technical challenge to the conceptual standing of nomological necessity – on the assumption of which, of course, the assurance of a single adequate scientific methodology depends (*see*, for instance, van Fraassen 1980, Cartwright 1983). The result has been a decided fragmenting of methodological unity within the natural sciences, a turn to the sociology of science rather than to its supposed rationale (or metaphysics or epistemology), and close studies of the actual ways in which the cognitive powers of the social sciences are inextricably invoked in anything that could be called "scientific method" (*see* Fleck 1979; Latour and Woolgar 1986).

Certainly in the English-language tradition, T.S. Kuhn's *The Structure of Scientific Revolutions* (1970) was, until quite recently, the single most important – ultimately a disappointing – study of the problem of scientific knowledge that refused to give exclusive pride of place to the older canon even in the physical sciences; that is, was not prepared to disallow conceptions of cognition that could not be convincingly naturalized, though that was never its avowed purpose. Kuhn now appears to have been ill-prepared for the debate his own work rightly provoked, though it has its important antecedents in Fleck and Bachelard (1984) and others. By this time, Kuhn's influence has been significantly reduced among the champions of the "older" canon (*not* the sociologists of science), who were prepared to contest the more radical possibilities of Kuhn's theory, for instance against older Marxist, Nietzschean, hermeneutic, and relativistic possibilities. Nevertheless, it would be a mistake to neglect Kuhn's intuitions about artifactuality and historicity on the basis of his own failure to develop a sufficiently robust philosophy beyond the fledging intuitions of *Structure*.

In bringing this preamble to a close, the work of the post-Tractarian Wittgenstein should certainly not be neglected. Kuhn published *Structure* in the International Encyclopedia of Unified Science, and Carnap (one of the editors) expressed genuine respect for Kuhn's book. Wittgenstein's *Tractatus* (1972) was used by the Vienna Circle as a fundamental text presumably providing a foundation for the Circle's own empirical philosophy (which was surely a mistake). But the post-Tractarian Wittgenstein, notably in *Philosophical Investigations* (1953), completely outflanked the conceptual boundaries of the earlier *Tractatus* by introducing the twin notions of "language games" and *Lebensform*, which, for reasons not entirely unrelated to phenomenological and Hegelian concerns (neither of which are at all explicitly present in Wittgenstein), signaled the most dampened concession possible to the significance of cultural diversity, artifactuality, and historicity.

Wittgenstein was obviously temperamentally disinclined to pursue the historicist theme, but it is difficult to make sense of his own conjectures in any way that does not mention these themes at all. Wittgenstein may be characterized as the most abstracted Hegelian and/or phenomenologist imaginable – neither committed to *geistlich* history nor to phenomenological apodicticity – who did, nevertheless, incipiently address the issue of self-understanding (*a fortiori*, the issue of a science extended to the human studies) within a minimal reading of the three notions mentioned a moment ago.

Some current analytic thinking regarding the humanities and the social sciences is attracted to this side of Wittgenstein, precisely because it is informed (on the evidence of the *Tractatus* and more) about the "older" canon and yet ingeniously obliges the canon's champions to consider a possible compromise between naturalism and the more disjunctive intuitions of the human studies. There is very little in the analytic literature that goes beyond Kuhn and Wittgenstein in this regard. The dawning problematic is clear enough, however, even if not yet entirely determinate: namely, whether, though real enough – if the human world is real – history and cultural formation can or cannot be described or explained in terms restricted to what (on familiar arguments) is deemed sufficient for the analysis of physical nature. In fact, the very idea of a principled conceptual distinction between (physical) "nature" and (human) "culture" is a prospect that hardly exists before Hegel. It does not rightly appear in Aristotle for instance, and it is hardly more than incipient in Vico and Herder (*see* Berlin 1976).

2. THE PARADIGM ARGUMENT

If we accept the phased lessons of the history of modern epistemology offered in the preamble, it becomes quite feasible to reclaim what it is about Kuhn's and Wittgenstein's accounts that makes them so arresting, without committing us to whatever are their actual philosophies. As a matter of fact, neither is particularly explicit (obviously, for different reasons) in the way of formulating specific philosophical doctrines. That may be a decided advantage, for there is a sense in which both bequeath us a series of compelling vignettess – Kuhn's, featuring certain exceptional moments in the history of science; Wittgenstein's, certain ordinary puzzles that are easily ignored but take on exceptional meaning in the telling. There is a legible common thread in the two accounts (that may be drawn as well from other sources) that dominates the local puzzles each has chosen to examine. Strenuous efforts have indeed been made to penetrate Kuhn's notion of a paradigm and Wittgenstein's language games correctly. But surely the master theme in both is much less elusive and less difficult to specify, namely, that, taken in the widest sense, thinking and perception are manifestations (or "utterances") of the culturally regularized practices of one or another historical society.

That is a disarmingly simple finding, certainly less quarrelsome than that of the proper analysis of "paradigms" and "language games." But it is also incontestable, and it is for that reason that the most fateful clues regarding a possible rapprochement between an adequate analysis of the conditions of knowledge in the natural sciences and the human studies may be found in Kuhn and Wittgenstein. If you step back from the local features of their specific arguments, you will find that the master theme is also implied in that large company of discussants that runs from Hegel to Foucault: including Kant, *if* (against Kant's intention) we paraphrase his transcendentalism along the lines of the collective and historical formation of cultural competences that, gropingly, leads from Kant to Hegel.

The uncontested theme that emerges runs more or less as follows: thinking and perceiving, *in whatever sense counts as confirmable knowledge*, appear, paradigmatically, in the form of what may be affirmed or denied, avowed, conjectured, or declared as what one believes. Left for the moment unanalyzed, the thesis is simply that thinking and perceiving, however originally "potentiated" (but not fully formed) as biological gifts, cannot properly be characterized in biological terms alone – or primarily; and that, now, at the evolved level at which *we* undertake the effort, *they* cannot be characterized except dependently, that is, *from* the vantage of reflexive "utterances": explicitly languaged (speech) or, by a plausible extension, modeled or interpreted in accord with linguistic utterance (behavior).

Put in the slimmest terms, what this means is that mental states and cognition cannot even be identified, unless either *paradigmatically* in terms of exercising "languaged thought" or "languaged perception"; or, by extension (in human cases or among languageless animals or machines), by invoking those frankly "*anthropomorphized*" forms in which ("uttered") behavior *is*, for descriptive purposes, linguistically modeled. By both maneuvers, they are cast in propositional form (even where language is absent: as in describing lions as seeing *that* the antelope have gathered near the water). This is the same model that has proved so helpful – possibly even unavoidable – in assimilating the apparent structure of thought and belief-states to explicit speech (*see* Geach 1957 and Kenny 1963).

We are at the point of deciding the baffling, enormously important question of *how* thinking and perceiving, or phenomenal or sensory experience of any kind, can be rightly identified and characterized. Of course, trivially, *to* identify and characterize is effectively to treat matters in a linguistic way. But, beyond that, the exemplars of what it means *to be* perceiving and thinking, to be conscious, to be undergoing experience, to believe, fear, hope, remember, plan, intend, and the rest must be *modeled* in the linguistic way if, discursively, we claim to be able to identify any non-linguistically "uttered" mental states. Here, if the force of Kuhn's and Wittgenstein's intuitions may be invoked again, without explicating their particular views, we surely find ourselves forced to acknowledge a set of wide-ranging conditions on the very intelligibility of speaking of mental life – which, of course, enters into every complexity regarding the cognitive standing of the humanities and social sciences.

There are at least two fundamental theorems that are notably congruent with Kuhn's and Wittgenstein's intuitions and that relevantly yield an unexpected benefit at very little conceptual cost. One holds that the paradigm of consciousness is self-consciousness, that is, the ability to state what one is conscious of. The other holds that all known societies competent in the first sense are also bilingual and that, effectively, bilingualism is tantamount to biculturalism. (Empirically, every linguistically apt society is bilingually competent; there is no principled difference between inter- and intra-societal divergence among human societies; *ergo*, all human societies are apt for bicultural understanding.)

The prospects of the human studies justifying claims of objective knowledge *of any sort* would be zero if these two theorems were refused. The first cannot be meaningfully denied, since denial itself implicates the intended model; and the denial of the second runs afoul of the tripartite lesson regarding the history of epistemology already adduced.

It needs to be said emphatically that these two truisms (and others like them) may be acknowledged without returning to the shoals of Cartesian dualism. To put the point in the best light: that is just the gain made possible by Brentano's reinterpretation of the Cartesian puzzle; the "substantive" disjunction Descartes is said to have favored is now rendered predicatively, where attributional difference need not be ontic disjunction. Brentano's maneuver permits us to reclaim intentionality (which is critical for psychological and cultural studies) that (for instance in Quine 1960) was canonically thought to entail dualism or to disallow reductionism. Those worries may now be put to rest, without yet entrenching reductionism or naturalism.

It is but a step from there to conclude that every science, every practice involving truth-claims, entails epistemic parity between the humanities and social sciences and the natural sciences! That is a windfall gained at very little cost.

For the moment, we need to dwell a little more closely on the paradigm of the mental and the cognitive. For one thing, the paradigm must be the same for consciousness and cognition, if indeed the paradigm of consciousness is self-consciousness: where, that is, self-consciousness is directly manifested in affirming what one perceives or thinks. Read in the most straightforward way, this means very simply that the paradigm data of conscious life *cannot* be initially ordered as (say) phenomenal experience *and* beliefs *applied for the first time to* such experience, or as biologically determinate experience *plus* linguistically or culturally formed reflections on same. *Paradigmatically*, mental life is indissolubly enlanguaged, apt for being discursively reported *or* at least described, top-down, on the model of reflexive consciousness, wherever experience can be characterized at all.

All attributions of mental states and content among prelinguistic infants and sublinguistic animals and supposed biological contributions to conscious life are, as we may say, *anthropomorphized* in accord with the governing paradigm. There is no other plausible access to mental life. Behaviorism has proved defective (*see* Chomsky 1965; Taylor 1964); supervenience is questionbegging, since it presupposes cognitive access to the mental (*see*, for arguments in favor of supervenience, Kim 1993); and the most prominent recent forms of naturalism bearing on the theory of mind, notably Fodor's (1998) and Dretske's (1995), attempt to outflank the primacy of the cognitive question altogether by proceeding, bottom-up, from would-be "necessities" alleged to govern perception and phenomenal experience. The paradigm argument outflanks such maneuvers by demonstrating that, if cognitively eligible, they already implicate the paradigm's primacy. *Tertium non datur.*

On the strength of such findings, we arrive at the pivot of all inquiries into the conditions of knowledge in the human studies: that is, the pivot for all those views that oppose, or at least find inconclusive, the older canon's insistence that whatever objectivity is assigned the humanities and social sciences is never more than what already belongs in an exemplary way to the physical sciences – the epistemic resources "known" to be adequate to the work of the latter.

Here, now, is the argument in its most abbreviated form: The paradigm of mental life and cognition is self-consciousness (*Premise*); Hence, the cognitive resources judged adequate for the physical sciences are themselves selected applications of the paradigmatic competence – self-understanding – without which linguistic behavior, cultural life in general, makes no sense (*Conclusion*).

Once you agree to any more ramified version of this enthymeme, you realize that you must concede two further findings: one, that there is and can be no principled disjunction, or priority of epistemic order, between the natural sciences and the human studies – they are ultimately one and the same, however conveniently sorted for particular descriptive and explanatory purposes; the second, that it is entirely possible that the special purposes favored in the humanities and social sciences may indeed feature distinctive epistemic powers – possibly implicated in the work of the natural sciences but often not noticed there, perhaps ignored, even denied – that are pointedly needed in the distinctive work of the human studies. Both conclusions are convincing, as we shall see. What needs to be noticed, however, is how difficult and strenuous a labor it is to formulate a countertheory to that of the naturalizing canon, at least in circumstances in which enlisting the support of analytic philosophy might (on the gathering evidence) be thought possible.

Once matters are put this way, a promising strategy for fleshing out the abbreviated argument just given might look like this: first, consider the principal epistemic resources that *all* inquiry requires but cannot be meaningfully admitted in any way that would reinstate the bifurcation of the physical sciences and the human studies or would confirm the priority of the first over the second; and, second, consider the special resources that the human studies positively require, that may be either ignored or denied application in the physical sciences.

Correspondingly, the naturalizing strategy takes two somewhat different forms: the more extreme (Kim 1993, Churchland 1989, Dennett 1991) tends to deny (more often than not, explicitly denies) that concessions of either sort need be made at all; the more agnostic or concessive (Davidson 1980a, Chalmers 1996, Searle 1992) tends to believe that whatever allowances are made need not depart in any irreconcilable way from what the naturalizing canon can allow. On the argument being mounted here, the evidence goes contrary to both claims.

The argument on the first count is surprisingly robust. In a way, it concerns the implications of denying any disjunction between the "subjective" and the "objective" along lines that oppose both the mind/body dualisms of the early stages of Cartesianism (rationalism and empiricism) and the nature/culture dualisms of its later stages (culminating – *not* overcome or resolved or even addressed – in Kant). So the elaboration of the epistemology of the humanities and social sciences belongs in an entirely uncontroversial way to the continuation of the standard history of epistemology already sketched.

There are at least three very powerful general constraints that strengthen the hand of all those who favor the second form of anti-dualism – which, of course, affects the fortunes of the first. For one thing, the admission of the paradigm already adduced signifies that, in principle, there cannot be *any* determinate characterization of physical nature that is not inseparably encumbered by whatever conceptual resources (historically contingent, remember, rather than assuredly transcendental or apodictic) prevail in any inquiry. That is, the very idea of a paradigmatic model of the mind (in both a phenomenal and a cognitive sense), cast in terms of the enculturing processes of language and social practice – *and* featuring the empirical likelihood that our epistemic resources (at the paradigmatic level) diverge significantly from society to society and from age to age – establishes, as already remarked, that the objectivity of the natural sciences cannot fail to be *part* of whatever objectivity belongs to the human studies (that is, belongs to the ramified forms of self-understanding). Hence, in a sense that may be said to be broadly Hegelian (favoring the *Phenomenology* 1977), the cognition of *nature* is, we may say (unguardedly), a part of the self-understanding of *Geist*. Of course, the idea must be freed from Hegel's purple prose. But the fact remains that the elaboration of this first constraint is noticeably central to the tradition that spans Hegel and Foucault and is largely absent from analytic theories.

A second constraint (also suggested) is entailed by the fact that, whatever the conditions of their objectivity may be, actually fixing reference, predication, context, reidentification, and allied notions determinately – which are, after all, applicable in all inquiries equally – *are*, on the argument drawn from the history of early modern epistemology, impossible to process in any way that would disjoin the "subjective" and the "objective" or would attempt to assign separate contributions to the one and the other. This is also the master lesson pursued in the period spanning Kant and Hegel, the theme still vestigially debated at this late date in Devitt and Putnam (already remarked), and the essential clue to the failed contests between what (on the ontological side) are canonically termed realism and idealism.

Once you grant the impossibility of disjoining, *in epistemic terms*, the subjective and the objective with specific respect to the issues raised on the second count, you see that the physical and human sciences cannot be disjoined in principle, cannot be ordered in terms of any cognitive priorities; hence, cannot give aid and comfort to any of the familiar programs of physicalism, extensionalism, naturalism, eliminativism, reductionism, supervenientism, behaviorism, or functionalism.

Seen that way, you cannot fail to grasp the remarkably strategic importance of getting clear about the epistemology of the humanities and the social sciences. For, on the argument, there is in *this* respect no principled disjunction between the natural and human sciences: the physical sciences are themselves reasonably characterized as abstractions made within an encompassing inquiry of "self-understanding" (to speak in Hegel's way) sorted for reasons bearing on their distinctive predictive power, technological control and precision, explanatory systematicity, nomologicality, and the like, which, to be frank, cannot be matched in the specifically human studies. Still, conceding this conventional contrast goes no distance at all toward supporting any of the naturalizing or analytic extremes mentioned just above. To argue otherwise would be a *non sequitur* and a violation of the lesson of the second constraint.

A third constraint makes itself felt at once. For, on the argument, *if* physical nature may be usefully abstracted from (say) the benign holism that makes objective predication possible, then the same reasoning requires admitting the distinction of the human or cultural world in which the very practice of predication implicates the paradigm of languaged thought.

The point is simply this: *if* predication (or reference or the like) cannot fail to be modeled or directly expressed in accord with the paradigm of enlanguaged thought, then whatever its conditions of objectivity are, merely acknowledging the constraint would oblige us to acknowledge as well the *sui generis* features of cultural life that, epistemically, provide the principal – possibly the only – descriptive model we have for objective inquiry: viz., what may be called the "Intentional" features of human life (still to be distinguished from what Brentano marks as the "intentional").

This means that the very objectivity of the natural sciences implicates the objectivity of the human studies: where, that is, predication is common to both (in effect, everywhere) and where Intentionality is not reducible (or eliminable in physicalist terms.

Seen from the vantage of the "older canon," this will be viewed as a heterodox proposal. But it is entirely straightforward when viewed in terms of the post-Kantian tradition stretching, say, from Hegel to Foucault. Strange as it may seem, the relative strengths of these two readings has never been satisfactorily settled within the terms of the now-dominant modes of analytic philosophy. Reopening the question of the conditions of knowledge in the humanities and social sciences permits us to reconsider the matter in a cool hour. The question is at least two hundred years old in its modern form and, admitting Vico's inventiveness, might even be thought to be incipient in Renaissance and post-Renaissance Europe.

3. THE BASIC EPISTEMIC COMPETENCES

We need to isolate the epistemological implications of reference and predication and related competences; but the foregoing discussion is bound to color our strategy, for it explains why we cannot satisfactorily disjoin the ontic and epistemic aspects of any inquiry. The short lesson is simply that the world we posit as real is never separable from what we claim to know about it, and what we claim to know about it is meant to correspond to the way the world actually is. Put thus, the lesson is logically trivial – though not unimportant; read in accord with the "canonical" history, the supposed nature of physical reality is very often thought to set prior substantive constraints on what we should admit as the true nature of the cognitive process itself. But that cannot be shown: in the limit, we should arrive again at the fatal disjunction Descartes originally favored. The gains philosophy has produced in the interval from Kant to Hegel should by rights have disallowed the Cartesian temptation; but it obviously has not yet done so – two hundred years later! What would the consequence be of banning it now? You cannot find the answer easily in recent analytic philosophy (*see* McDowell 1996; Putnam 1994).

The entire matter of the epistemology of the humanities and social sciences is entangled in that question. So too is the epistemology of the natural sciences. The trick is to isolate the conceptual clues we already have, in a way that simply outflanks the Cartesian option at the start, a way that never permits us to patch over Descartes' paradox by invoking ad hoc inventions that cannot fail to be too late. The decisive clue is this: we cannot save the objectivity of the natural sciences without also saving the objectivity of the human studies, even though there are important differences between the two. We cannot separate the epistemology of the one from that of the other; we cannot prioritize the resources of the one over those of the other; and, of course, we cannot disjoin epistemology and metaphysics.

This seems so sensible a concession that one may even be prepared to believe that allowing it cannot possibly defeat the canonical assumptions of the strongest analytic strategies, for instance, those of naturalism and reductionism. In fact, those (analytic) strategies claim to rest on our ability to state what the real world is like independently of any entangling conditions of inquiry; and *that* now looks suspiciously like reclaiming the Cartesian option. At the very least, we must come to terms with the following question: Are there any epistemic resources, or cognizing competences, that, if admitted, unconditionally disallow the Cartesian disjunction – and, as a consequence, the viability (perhaps the very admissibility) of programs like those of naturalism and reductionism *if* tendered in the Cartesian way? The surprisng answer is: indeed there are!

To make a long story short, the familiar competences of fixing reference, predication, reidentification, as well as fixing the contexts in which they are successfully deployed, are impossible to specify in epistemically operative ways except on the condition of disallowing the Cartesian option. But if that is so, then there is no viable realism fitted to the work of the physical sciences that is not already a *constructive realism* (a constructivism): by which is meant a realism abstracted, projected, constructed *from epistemic* data that (*a*) eschew all forms of cognitive privilege, indubitability, foundations, or the like (for instance, *contra* Chisholm 1977; (*b*) admit that there is no principled disjunction between the subjective and the objective (for instance, *contra* Devitt 1991 and in agreement with Putnam 1987, though Putnam 1994 no longer favors that earlier formulation); and (*c*) disallow the supposedly exhaustive disjunction between the classic forms of realism and idealism premised on first opposing (*a*) and (*b*) (*see*, for instance, Moore 1959, Collingwood 1978).

If the supporting argument proves compelling, then a very large family of analytic programs – naturalism, reductionism, eliminativism, supervenientism, functionalism, and the forms of physicalism and extensionalism favored by the positivists and the unity of science program – will be put in mortal jeopardy. Unless, of course, they can be recovered (benignly or dependently in the constructivist's sense) under conditions known not to yield to naturalism or reductionism directly. The larger question is not our concern here, except in the sense that, if the argument holds, no principled disjunction between the natural sciences and the human studies (the humanities and social sciences) could ever be legitimated; also, in the sense that the first cannot then provide an independent or antecedent model for the realist and objective standing of the second (for instance, *contra* Carnap 1959, Hempel 1965, Popper 1950). As a result, the difference between the two sorts of discipline becomes itself "constructive."

The first clue is a dialectical one: viz., that the defeat of Cartesianism entails the unavoidability of a constructive realism and the inseparability of the natural sciences from the human sciences. If that held, it would be a gain of immense importance – conditional, of course, on the premise that reference and predication and allied powers *were* "constructive" competences. You must bear in mind that reference and predication are, paradigmatically, linguistic powers; that is, ineliminable cognitional competences acknowledged on the strength of the assumption that mental and cognitive states (even the mental states of nonlinguistic animals) are paradigmatically modeled on *our* ability to report, veridically, perceptual and other experiences.

This is not to deny that mental and cognitional states may be ascribed to creatures lacking language. It is only to insist that animals and prelinguistic children are said to have the phenomenal experiences they have and to have knowledge of same – and, because of that, knowledge of the world – because their states are describable only as modeled on the human paradigm. (The point has already been remarked: there seems to be no other way of accessing the phenomenal and cognitive content of the mental states of nonlanguaged creatures.) But if all that is admitted, then the second clue concerns evidence that referential and predicative success in natural-language contexts does indeed depend on conditions that confirm the inseparability of the subjective and the objective and the inseparability of the resources of the natural sciences and human studies.

The slimmest argument in favor of this last finding holds that if we are to avoid skepticism and the Cartesian *aporia* and abandon, at the same time, all the forms of cognitive privilege, we have no recourse but to admit that success in these regards must be *consensual* (without yet being *criterial* for that reason); that: (*a*) success is itself constructive (or constructionist), and (*b*) success accords with the saliencies and limits of tolerance spontaneously favored in the fluent practices of the natural-language discourse of this or that particular society. In our own time, the barest, the most minimal sketch of what the thesis requires may be found in Wittgenstein's notions of language games and the human *Lebensform,* as well as in Wittgenstein's analysis, on *lebensformlich* grounds, of Moore's apparently privileging account of such locutions as "I know" (*see* Wittgenstein 1953; 1958).

On the reading of the history of philosophy favored here, Wittgenstein is, as already remarked, the thinnest, most marginal proponent of the essential epistemological lessons captured by the tradition that stems from Hegel. That is, Wittgenstein's fundamental argument may be paraphrased thus – if indeed (contrary to his own protestations) in the way of an argument: first, that, if we are to avoid an infinite regress or an appeal to cognitive privilege, truth-claims must be grounded in consensual practices rather than by way of foundational propositions (the thesis of the *Lebensform*); and, second, that grounding them thus precludes any criterial or determinately rule-governed certitude in determining their truth. (There is some evidence, in fact, that Moore's common sense philosophy is, despite Moore's own repudiation of Hegelian themes, not very for from the same finding.)

All that is needed to bring Wittgenstein's thesis into line with the Hegelian notion of historicity (which Wittgenstein never advances but which is implicit, at least incipiently, in what he does say – admitting cultural diversity and the gradual change in the details of any language game in actual use) is to admit that the human competence to discern the right use of language (in reference and predication and in fixing context and meaning) depends, "constructively," on the continual evolution of collective practices that lack (in the sense already suggested) any strictly rule-governed structure.

The textual issue is, however, not the central one. The point is, rather, that we cannot find any similar thesis (with the single exception of Kuhn's failed alternative) among the most important figures of the analytic tradition. For, if you admit the near-total opposition to historicity among the analytically-minded and their overwhelming inclination to recover some form of Cartesian certitude (at least up to naturalism or reductive physicalism), it becomes clear at once that the barest attempt to recover the basic competences of reference and predication along lines that span (say) Hegel and Wittgenstein will be viewed as fundamentally opposed to the most entrenched philosophical convictions of the analytic movement. That is an extraordinary finding.

All the more reason, therefore, to insist that there may be no more convincing way of managing reference and predication epistemically than along the lines so sparely sketched in Wittgenstein. The analytic canon – certainly the canon thought to be adequately represented in the first-order predicate calculus, as characterized in Russell (1905) and Quine (1960) – is clearly convinced that, on purely formal grounds, the natural-language devices of reference and denotation can be effectively replaced by the resources of quantification and predication (purely formally, in Quine; backed by "knowledge by acquaintance," in Russell).

Later figures, notably Davidson (1984), have actually recommended abandoning reference as an ineliminable explanatory ingredient among the semantic resources by which we account for our grasp of reality. Quine seems to have violated Leibniz's (1956) compelling dictum to the effect that successful reference cannot be captured (hence, cannot be explained) by any predicative means; Russell (1912) is troublesome on the issue because of his reliance on knowledge by acquaintance with regard to predicables; and Davidson nowhere attempts to demonstrate that we can indeed function in natural-language contexts without well-formed referential facilities – or, better, without being able to explain how we do so.

The matter is of decisive importance in spite of the fact that there is almost no sustained discussion, in specifically epistemic terms, of reference and denotation in the whole of Western philosophy, certainly not in the strictest analytic literature. For example, Kripke's (1980) influential account offers no epistemic considerations at all. Yet, surely, the theory of reference has no point if disjoined from the operative conditions of actual discourse. It may also be said that a related, though entirely distinct, difficulty confronts the presumption of predicative success.

There is a narrow and a broad lesson to be drawn here. The narrow lesson confirms the impossibility of explaining referential and predicative success in any way that disjoins the subjective and the objective along Cartesian lines: that was offered earlier as the thin ("Hegelian") promise of Wittgenstein's *lebensformlich* conception. The broad lesson confirms that, on the basis of the narrow one, *every viable realism must be a constructive realism* (a constructivism). If both lessons hold, then we should arrive quite quickly at the heterodox thesis already favored in this review: namely, that there is and can be no principled disjunction between the epistemic resources of the natural sciences and the human studies and no privilege or priority or higher normative standing assigned the methodology of the first over that of the second. If you see all this, you see at once that all the usual forms of naturalism and reductionism are put at risk; *and* that their prospects will be defeated if indeed referential and predicative success behaves in the manner sketched – that is, consensually (even where criteria are provided).

The argument proceeds by spiraling through the evolving topics. The reason is simply that language and thought and cultural life *are holist* – benignly holist – in the plain sense that the structure of meanings and the meaningful structure of enlanguaged thought and social practice make no sense except in terms of a part/whole relationship that is itself heuristically constructed to explain such structures from one or another point of view. This is the core intuition of the notorious "hermeneutic circle": that is, the notion that the "parts" (of meaning) cannot be understood except in terms of the "whole," or the whole except in terms of the parts. You have only to think of a dictionary to grant the idea's validity. But what is more important is that the admission of the hermeneutic circle (*see* Hirsch 1967; Gadamer 1975) is *not* tantamount to the infamous doctrine of "internal relations" (attributed, perhaps falsely, to F.H. Bradley and in a more allusive way to Hegel (*see* Rorty 1967).

On the strength of the latter doctrine, the meaning of any "part" (or proposition) cannot ever be assigned without reference to the meaning of the single, "whole," inclusive proposition about what is real; but it is also true (on the "internal relations" argument) that *that* "whole" is beyond the recuperative epistemic powers of any and all human agents. Hence, on the doctrine of internal relations, nothing can ever be said that *is* either meaningful or true! The second doctrine is a form of skepticism, but not because of its holist features. The first is not skeptical at all *and*, more to the point, not (despite its holist features) in the least inimical to the discursive resources of reference and predication and truth.

A holism of a more doubtful sort (defying factual analysis) has indeed been championed by Quine (1992) and attacked by Fodor and Lepore (1992). But holism need not take either of these forms, for the simple reason that the part/whole relationship it invokes is: (*a*) confined to meanings or Intentional structures (where the term "Intentional" has still to be explicitly characterized); (*b*) constructed as part of the process of interpreting cultural life (or what may be modeled, in the anthropomorphized way, on languaged thought); (*c*) real, in the cultural sense, in the way of being determinable in accord with (*b*); (*d*) open to plural such constructions under the condition of historicity; and (*e*) fully compatible with admitting, for epistemic purposes, an operative (constructed) disjunction between cognizing subjects and cognized objects.

Davidson links holism with the model of rationality (1980b), but he also views that model as unacceptable for the explanatory work of a proper science (though it remains adequate enough for quotidian use). In fact, analytic philosophy generally takes an extreme, quite unnecessary, even indefensible view of "meaning holism" (Fodor's term), possibly because it relies on Russell's attack on Bradley's putative doctrine favoring "internal relations." The notion of the hermeneutic circle, though it has its own penchant for fixity and privilege (*see* Hirsch 1967, Schleiermacher 1977), has shown the way to a more manageable practice as far as human powers are concerned. The fact remains that *if* cultural phenomena are not reducible to physical phenomena, then *some* form of cultural holism is unavoidable; all that is needed – and that is clearly in place – is that cultural holism (or "meaning holism") need not preclude the discursive functioning of truth-claims, linking and distinguishing cognizers and cognized objects within a holist space.

It is true that the solution to the problems of reference, predication, context, reidentification, rationality, and meaning *are* hermeneutic issues, issues that invoke a benign form of the part/whole relationship that analytic philosophy wishes to avoid. But it is also true that the essential Wittgenstenian clue – regarding language games and the human *Lebensform* – offers a form of holism that is clearly compatible with the familiar discursive distinctions between stably identified subjects and objects. Nothing more is needed but the assurance that to admit the *sui generis* features of the cultural world (which effectively implicate the holism of

Intentional phenomena) does not disable in any way the methodological rigor of the sciences. On the contrary, whatever rigor the sciences claim need be no more than the constructive work of culturally apt inquirers functioning within the benign holism of their own *ethos*. If so, then, on the gathering argument, the rigor of the natural sciences presupposes rather than escapes the holism that is directly admitted in the human studies. That's to say, the realism of the natural sciences is as much a constructive realism as the realism of the social sciences.

There are two lines of inquiry that intersect here. We actually need to develop them together. But we may, for convenience, postpone treating the one that is now dawning until we have settled the one that has been before us for a longer inning. The first returns us to reference and predication; the other concerns the bearing of constructivism and Intentionality on our choice of models of objectivity. It has already been suggested that the answer to the first would help to fashion a satisfactory answer to the second. What has just been aired about the part/whole problem confirms the good sense of that conjecture. (That is enough for the moment.)

Obviously, we need a measure of closure on the first question. Here at least is the thread of the argument intended: *if* reference and predication may be epistemically legitimated only on the condition that there is no principled disjunction between the subjective and the objective, and *if* selves are culturally constituted so that *they* function cognitionally only in conformity with the practices of their enabling society, then all forms of objectivity are constructed, never more than consensual, subject to historical divergence and change, and corrected in accord with interpreted societal interests. Furthermore, if *that* is so, then there can be no single, neutral, true, Cartesian description of the world. That is, what is now called objectivism is demonstrably indefensible (*see* Bernstein 1983).

Here, the two clues are compressed into a single finding. About the first, we may say with assurance that reference *is*, logically, insurmountably informal; hence, that success *is* consensual, *not* algorithmic or criterial, in a sense that cannot be very different from Wittgenstein's *lebensformlich* theme. The required argument would feature the following obvious truths about natural-language discourse: (*a*) general referential success is indispensable and rightly taken to be exemplary; (*b*) reference cannot be reduced to predication, and predicative success is inseparable from referential success; (*c*) referential success presupposes a benign form of "meaning holism" relative to which the subjective and the objective may be constructively disjoined in epistemic terms but only in a way that disallows any ontically prior separation of subjects and objects; hence (*d*) referential success cannot be accounted for in terms that fail to implicate the epistemic powers appropriate to the human studies, and the epistemic competence assigned the natural sciences must be a constructivist abstraction from the encompassing reflexive competence of human self-description and the work of the human studies. Add only this: that what holds for reference holds for predication as well – and beyond that, for reidentification and context – and you see at once why (regarding the second clue) models of objectivity suited to any and every inquiry cannot disjoin the natural and the human sciences.

This hardly signifies that there are no important epistemic differences between the two sorts of inquiry, only that the decisive differences cannot be based on general differences regarding reference (or predication) but rest instead on the different interests that may separately inform the natural sciences and the human studies. In general, such interests concern the description and explanation of phenomena that possess "Intentional" properties (human and cultural phenomena) and those that do not (natural phenomena), particularly with respect to explanation under covering laws, normative explanation, prediction, and technological control.

We cannot rightly anticipate what such differences may entail, except to say that they presuppose the generic constraints called into play by the need for referential and predicative success and that, granting the first point, it follows that even where Intentional distinctions (which must still be properly delineated) are absent (as among natural or physical objects and properties) the corresponding disciplines (the natural or physical sciences) are *still* constrained by whatever generically affects the human studies in the way of reference and predication.

It is a curious fact that these two sorts of constraint (and the link between them) are almost entirely ignored by positivism, the unity of science program, Popperianism, and even by more recent analytic models critical of these (*see* Feyerabend 1975; van Fraassen 1989, Fine 1986), as well as by the post-Kantian bifurcation between the *Naturwissenschaften* and the *Geisteswissenschaften* (as in Dilthey 1989 and Gadamer 1975).

The argument remains conditional of course, both because more must be said about predication and related competences and because the meaning of the Intentional has not yet been pointedly supplied. But the dialectical gains are palpable enough, and the inseparability of predicative and referential success may surely be taken to confirm the essential lesson even without a fuller account of predicative practices.

We may be reasonably certain that reference is insuperably consensual, that predication has no epistemic function wherever disjoined from the conditions of referential success, that reidentification and the fixing of the contexts in which referential and predicative success obtain cannot be epistemically more secure than that of reference itself, and that all further epistemic questions of meaning, rationality, Intentionality, history and historicity, and interpretation – which distinctly bear on resolving the problematic objectivity of the human studies – logically *presuppose* some form of referential and predicative success that, on the gathering argument, *already implicates our having adopted a viable form of objectivity suited to the human studies.*

In short, the objectivity of reference and predication is, however explained (if denied cognitive privilege or Cartesian realism), suited only to inquiries that cannot be disjoined in principle from whatever limitations beset the human studies. That is a finding of the greatest importance, denied as a matter of course by Cartesian convictions that are still very much in force at our end-of-century.

4. PREDICATION AND INTENTIONALITY

There are two key findings about referential practice that are plainly matched by predicative practice, though the two are very different in ontic and epistemic respects. It is true that rigor and precision in the way of objectivity are insurmountably informal in both. There is no viable principle, criterion, algorithm, or rule of application by which epistemic or communicative success can be ensured in either case. That means of course that whatever precision may be accorded the strongest sciences rests, finally, on the informality of natural-language reference and predication: *no* subsequent precising of their criterial or evidentiary grounds can escape this insuperable informality. But if that is so, then, as has been said, the precision of both is *lebensformlich* (or akin to the *lebensformlich*, as Wittgenstein very lightly hints, speaking of natural-language discourse in general).

That now leads to a decisive (second) finding, namely, that the objectivity of reference and predication cannot be separated from one another and cannot be disjunctively assigned to either the work of cognizing subjects or to the independent properties of cognized objects, but belongs indissolubly to what (regarding the intelligible or intelligibilized world) can be identified only as *that* world (that is, "symbiotized" between subjects and objects). Only on that assumption can a viable disjunction be defended between cognizing subjects and cognized objects in virtue of which objective claims may be processed in our inquiries. There you have the single most important theorem by which the epistemic and ontic fortunes of the natural and human sciences prove to be inexorably linked; hence, also, linked benignly enough within whatever holist space we assign the human studies.

On the argument, predication is even more decisive than reference. For reference is "consensual" in an informal (non-criterial) sense chiefly as a result of *faute de mieux* considerations – in particular, because reference cannot be retired, epistemically, in favor of predicative powers and because there are no other evidentiary grounds to invoke. But predication, which depends on the very *generality* of predicates, succeeds by consensual means, because, although there must be evidence of a predicable sort, Platonism, which is the only conceivable alternative to *lebensformlich* consensus, is known to be inaccessible to human cognizers.

We do succeed, in predication, *in* the *lebensformlich* way, but *not* by default, as with reference. We succeed by discerning general predicables but it must be in a way that precludes the *a priori* assignment of cognizing powers to apt subjects adequate for discerning Platonized properties inhering in the cognizable world.(There is no parallel in referential contexts.) *Lebensformlich* success is always consensual, constructed, non-criterial, and symbiotized between subjects and objects. That is the reason realism in our time can no longer be characterized in the familiar (Cartesian) way: by disjoining and opposing realism and idealism or by opposing realism and anti-realism (*see* Devitt 1991). The substantive conditions of predicative success make that impossible.

You must bear in mind that there simply *is* no perceptual or similar resource for fixing identity in referential or denotative ways. We may of course rightly claim epistemic grounds for fixing reference; but there always remains a conceptual risk in claiming to have reidentified one and the same *denotatum* under the conditions of natural-world change. There is no successful predicative means by which reference can be fixed. All the well-known analytic efforts to decide the question are known to fail, though reference and reidentification need not fail, *cannot* fail massively: that is, all the solutions offered by Frege (1960), Russell (1905), Strawson (1959), Quine (1960), Searle (1958), and Kripke (1980) must be *epistemically* inadequate. Reference obliges us to rely on the *sui generis* (*lebensformlich*) resources of the human studies. There you have part of the sense in which the natural sciences are already human studies.

By contrast, predication makes no sense if disjoined from perceptual and similar powers. Furthermore, predicative success is inseparable from referential success: we can rightly claim only that *this* or *that* particular *denotatum* is determinately thus and so. The grammar of predication plainly affects its epistemic prospects. Notice, for instance, that what counts as "red" in the context of individuatable wines is not the same as what counts as "red" in the context of individuatable oil paints, tomatoes, lipsticks, or roses: the extensional tolerance of the predicate is clearly affected by our conception of the collection of individuatable *denotata* that are accepted as being of these or those "natural kinds" or simply of acknowledged kinds – to which, that is, the predicate in question applies in different ways.

The problem with predication is this: it makes no difference whether the predicates we apply are non-Intentional or Intentional ("red," say, or "baroque"); *the conditions of consensual success are Intentional in either case*. But if so, then there can be no principled disjunction between the natural and the human sciences; our bare reliance on predication (and reference and contextual relevance and the like) signifies once again that the physical sciences are, *qua* sciences, human sciences! For, of course, they rest on some evidentiary ground and, on the argument, there are no grounds that escape the symbiosis of reference and predication.

The objective standing of reference, remember, depends on construing story-relative reference as objectively valid about the "independent" world. But the objective standing of predication depends on a substantive attribution of properties. That is the whole point of the ancient quarrel regarding the realist or nominalist standing of universals. Both accounts fail; for the generality of predicates (mere words) does not entail independent universals or properties (realism) and does not ensure our epistemic power to discern universals or general properties by their use (nominalism). The entire classical/medieval debate collapses.

Here again, Wittgenstein grasps the right solution (which he does not fully explicate) in speaking of "family resemblances" and "strands of similarity." But it would be a mistake to think that reference to family resemblances is, or is the key to, Wittgenstein's solution (*see* Bambrough, 1960-61): it's not the shift from natural-kind terms to strands of similarity that counts; it's only that both have objective standing only on *lebensformlich* grounds.

If you grasp the point, you have in effect caught the nerve of all the arguments that oppose objectivism. For you see that, admitting the symbiotized nature of reference and predication, it makes no sense to think of truth as the correspondence between what we affirm and what the state of the independent world is. (*See* Wittgenstein 1972). Not only is such correspondence inaccessible, there is no intelligible sense in which the doctrine could possibly inform our objective inquiries. The imputed objectivity with which we fix the truths of science must be as *lebensformlich* as reference and predication are and for the same reasons. (Correspondence may be treated artifactually, of course, but then it cannot claim neutrality.)

This now begins to strengthen the sense in which explicating the relationship between the objectivity of the natural sciences and of the human studies is tantamount to dismissing the entire Cartesian spirit that spans mid-17th and late 18th-century philosophy (Cartesian realism). In fact, the objectivism that runs from Descartes to Kant is, in spite of enormous differences in local doctrine, not very different from the objectivism that spans the work of the positivists and unity of science theorists down to the minimal naturalisms of Quine (1969) and Davidson (1986). Devitt's frank Cartesianism (1991) confirms the continuity.

This helps to explain why it is so difficult to mount a straightforward analysis of the epistemology of the human studies. Inevitably, the counterargument must reverse the canonical charge that the human studies are a diminished form of the exemplary competence of the natural sciences. That cannot be true if (*a*) the domain of the human studies is different *in kind* from that claimed for the natural sciences, (*b*) objectivity in the natural sciences presupposes and entails the objectivity of the other, and (*c*) what distinguishes the phenomena of the human or cultural world cannot be eliminated or reductively replaced by "natural" or physicalist means.

There is a collateral line of reasoning that may be usefully inserted here. *If* the processing of objective claims in physics or sociology (say) is *lebensformlich* in the same sense, then there cannot be a principled disjunction between theoretical and practical reason: for theoretical truths – for instance, explanations under covering laws (if possible) – must be a species of truths grounded in our consensual practice. But if so, then even second-order views of what to count as objectivity are as much open to confirmation (by *lebensformlich* means) as are the familiar first-order claims of the natural sciences – once again for the same reasons. It's our intellectual history that dissuades us from that finding, but there remain no compelling grounds on which to demur.

The decisive discovery – the one already broached – is simply that, now, even in the strongest inquiries, we cannot count on any antecedently independent state of affairs *that we can consult* in the correspondence sense. All judgments are "practical" (even if "theoretical") because they are grounded in the consensual practices of our home society. The conventional disjunction is now seen to be transparently paradoxical, because it presupposes the Cartesian option. The option is still very much before us (*see* for instance Putnam 1987; McDowell 1996): we must note the fact in assessing the most salient contemporary theories of objectivity and objective truth.

Two theorems may be drawn from all this – already bruited – that confirm the inseparability of the natural sciences and the human studies: one, that the solution of the Cartesian *aporia* regarding realism (which runs from Descartes to Kant) leads inexorably to the symbiosis of subjects and objects, hence to constructive forms of realism regarding both the natural sciences and the human studies; the other, that, on constructivist grounds, there cannot be any evidentiary basis on which, *a priori*, to ensure our knowledge of the independent world as it is independent of our inquiries. Hence, all forms of the correspondence theory of truth are epistemically inoperative. If these two theorems hold, then it follows at once that objectivity must, in being an artifact of inquiry, be consensual; and, in being consensual, it must oppose any epistemic disjunction between the theoretical and the practical.

That is a considerable gain. For, on the argument, *if* the physical sciences are committed to some form of constructive realism, then their objectivity depends on our being able to support objective claims among the human studies; for the objectivity of the first presupposes the objectivity of reference and predication and contextual relevance and meaning (and much more), and all such processes are precisely what the would-be objectivity of the human studies must (and means to) secure.

Predicates, as already remarked, may be Intentional or non-Intentional. "Red" and "round" and "possessing a positive electrical charge" are reasonably treated as non-Intentional (or, as designating non-Intentional properties). "Baroque" and "just" and "remembered" and "feared" are surely Intentional predicates. Nevertheless, on the argument being advanced, to discern the objective presence of *non*-Intentional properties now implicates a companion competence regarding Intentional properties. For, for one thing, predicative similarity among discernible differences is inherently consensual – *not* neutral or correspondentist in the objectivist's sense; and, for another, there is in principle no other ground, independent of *lebensformlich* constraints, on which predicative success can be confirmed. But that means that the objectivity of the natural sciences presupposes and entails a comparable objectivity with regard to Intentional properties – which, on the usual physicalist and naturalist arguments, either have no place in nature or are eliminable or reductively replaceable by purely physicalist means (*see*, for instance Churchland 1989; Fodor 1990; 1998). The burden of proof clearly belongs to the reductionist and naturalist.

It would take us too far afield to demonstrate that eliminativism and reductionism (and even non-reductive supervenientism – *see* Davidson 1980b) cannot succeed. It is enough for our present purpose to accept the argument in its conditional form. In any case, unless some very strong form of physicalism holds, all the foregoing arguments confirm the strategic importance of vindicating the objectivity of the human studies.

It takes but a moment to realize that we have in effect returned to the paradigm argument of section 2. For it turns out now that any objective claim to the effect that *k is F* (for both Intentional and non-Intentional predicates) implicates the objectivity of our being able to reconcile that claim with our *lebensformlich* practices.

Here it may be usefully explained that "Intentional" = "cultural," in the straightforward sense of designating whatever is significative or meaningful in a linguistic or linguistically informed way: in a symbolic, semiotic, representational, expressive, rhetorical, conventional, verbal, gestural, institutional, rulelike or similarly structured way. The paradigm of the cultural is whatever may be linguistically affirmed; all else in the way of perception and mentation and behavior and activity is "anthropomorphized" in accord with the paradigm. The entire range – both the proper range of the paradigm and its anthropomorphized extension – counts as Intentional: the term is meant primarily as a convenience for identifying what belongs to the cultural and what is absent from the merely physical. If that disjunction runs true (assuming naturalism to be false), then we have indeed reached an important finding in having demonstrated that the objectivity of the natural sciences presupposes and entails the objectivity of the human studies (the social sciences and humanities). For, now, it appears (conditionally) that *the objectivity of the natural sciences is itself Intentional.*

We need to have the general features of the Intentional before us in order to proceed effectively. Here are the most important ones for epistemic purposes. The Intentional is: (i) ontically emergent in a *sui generis* way, not reducible to the physical or biological though inseparable from same; (ii) paradigmatically described in terms of our linguistic competence (our ability to affirm what we take to be true) or modeled in the anthropomorphizing way on that same paradigm; (iii) regularized in terms of the collective practices belonging to the paradigm, for instance in terms of referential and predicative practices, practices of fixing meaning, context, rationality, and the like; (iv) subject to constructivist constraints on truth and objectivity by consensual (but not *a priori* criterial) means, openendedly in accord with the changing drift and tolerance of our *lebensformlich* practices; (v) capable of reconciling past and present epistemic claims in alternative, divergent, even potentially incompatible ways viewed from the vantage of current consensual saliencies; and (vi) intrinsically apt for interpretation and reinterpretation along all the lines of significative structure. The Intentional is confined to the predicative, accommodates but is not equivalent to the "intentional" as defined by Brentano and Husserl (*see* Mohanty 1989); and marks the ontic distinction between the natural and the cultural (*see* Margolis 1995).

On the argument, there are no purely Intentional entities (meanings, say, *or* universals, or thoughts). There are indissolubly complex, culturally emergent entities (selves and artworks, preeminently) that are, we may say, *embodied* in physical, biological, or electronic *denotata.* On this reading, selves – *not* the members of *Homo sapiens* in which selves are (emergently) "embodied" – are "second-natured," transformed in infancy by internalizing the language and *lebensformlich* practices of their enabling home society, so that they (the emergent new generations) become the apt continuators of the specifically Intentional competences of their society. All other cultural phenomena (artworks, speech, actions, traditions, institutions) are, at different levels of abstraction and idealization, the "utterances" of aggregated selves, which, for the special purposes of particular social sciences, humanities, and related human studies, may be conveniently nominalized as culturally distinct *denotata* – artworks, for instance, or historical "deeds" like the events of the Second World War.

Predicatively, the Intentional is indissolubly *incarnate* in natural properties, matching *embodied* entities. The important point is that the entire cultural world is *sui generis*, emergent, second-natured, significative, artifactual, and intrinsically interpretable. On the usual theories, nothing of the kind occurs in physical nature, except where anthropomorphized (as in speaking of the perceptions or intentions of lions stalking eland) or by the courtesy of intervening scientific theory (as in "interpreting" the Olduvai Gorge or the "meaning" of the increased presence of iridium in geological formations relative to the disappearance of the dinosaurs). These accommodations depend ultimately on the symbiotized nature of predicables; for the objective attribution of relevant properties, even where confined to physical nature, presupposes and entails our competence to discern the "meaningful" structures in question. To admit the point is to acknowledge a strong distinction between (physical) nature and (human) culture, a distinction that could hardly have been meaningfully made before the French Revolution or, indeed, before Hegel. (And yet, of course, it is just that constraint that is absent in Brentano's and Husserl's speculations about the "intentional.") All this may be collected in a word: selves and other cultural *denotata* are, we may say, ontically *hybrid.*

5. INTERPRETABILITY

Once we cross over from the non-Intentional to the Intentional, the entire question of the unity of science is put at risk. There is no convincing sense in which natural language can be denied reality or in which human selves are denied a mental life that, in virtue of being "enlanguaged," is utterly unlike the mental lives of sublinguistic creatures. There are admittedly tantalizing analogues of proto-cultural and proto-historical life among the monkeys and the apes and even birds, but it is surely no more than a very small start. Storks have apparently changed their eating habits with the rise of modern cities. A community of Japanese monkeys has learned to wash sweet potatoes and rice and convey the discovery to the next generation. Konrad Lorenz (1970-71) claims that a much-diminished flock of jackdaws of his acquaintance has preserved a traditional flight pattern over several generations that has nothing to do with purely biological patterns. Chimpanzees and gorillas appear to be capable of mastering something approaching grammatical competence and the command of a vocabulary (*see* Premack 1976). But only human societies exhibit the ability to invent and control any number of rapidly changing, thoroughly artifactual, and profoundly different cultures; and there seems to be no explanation of this capacity that does not feature man's mastery of true linguistic fluency. There is also no known analysis of linguistic and cultural competence in purely physicalist terms, though there are memorable efforts to account for language biologically (Chomsky, 1986 most notably), even thought (as in Fodor 1975; 1983; *see*, also, Churchland 1995).

It's not merely language of course that counts: it's also the fact that the appearance of *selves* as the "second-natured" sites of the emergent competence of the members of *Homo sapiens* (in virtue of internalizing the linguistic and cultural powers they gain as infants living among the apt selves of an environing society) counts as the irreducible, *sui generis*, reflexive ability of human beings to reinvent themselves without end that makes the study of the human world a discipline radically different from that of the natural sciences. Although, of course, the achievement of a science is itself a cultural feat, completely beyond the cognitive competence of sublinguistic creatures and inexplicable in terms of sublinguistic abilities.

The saving connection between the two sorts of discipline rests with the fact that whatever belongs exclusively to the Intentional dimension of the human world – even the enabling conditions for the emergence of selves – is indissolubly incarnate in the potentiating powers of the pre-cultural natural world. That alone permits us to make full sense of human perception and thought and of the intelligibility of Intentional process and artifact. For it explains the sense in which the mental life of humans – *a fortiori*, their cognitive powers, their science – cannot be explained, bottom-up, in terms of any merely biologically evolved competence. For, as was already been argued, the very paradigm of consciousness is self-consciousness; that is, the paradigm of the mental is the cultural.

So it is that the novelty of the human world cannot be acknowledged without also admitting the differences between the conditions of objectivity of the natural sciences and those of the human studies. Let it be granted that whatever counts as the objectivity of the natural sciences does so only under the encompassing constraints of whatever functions as objectivity in the human studies. Nothing need be disturbed as far as the powers of science are concerned: except for false conceptions of the disjunction between the two sorts of discipline or the exemplary standing of the natural sciences vis-à-vis the human studies (and what that might imply). These are now entirely outmoded conjectures.

If you turn, then, to collect the distinguishing marks of the objectivity of the human studies, you soon realize that you must allow for the following at least: (*a*) that the cognizing subjects and cognized objects of the human world are, ultimately, one and the same, or the difference is mediated, interpretively, by overlapping pairs (as the lesson of the Rosetta Stone makes clear); (*b*) that objectivity joins whatever holds among the non-Intentional sciences and what may be rightly assigned the Intentional "utterances" of the human world (as in speech, deed, artifact, institution, tradition, and history); (*c*) that objectivity in the human world is centered on determining the significative or meaningful structure (or import) of whatever transpires in a culture, without disturbing any causal regularities that hold at the purely physical level but also without precluding the causally effective force of the (incarnate) powers of (embodied) selves; and (*d*) that the cognizing powers of interpreting selves are formed and transformed in a historicized way as a result of the drift of the *lebensformlich* practices of their encompassing society, as those powers are actually exercised.

Alternatively, we may say that selves are historied artifacts of a peculiarly gifted sort; that nomological and natural-kind uniformities must accommodate the causal efficacy of cultural events; and that we cannot assume *a priori* any "closure of the physical" (*see* Kim 1998) that would force us without argument to admit the cultural and the mental only in reductionist terms. Seen that way, the human sciences require a reconciliation of the natural and the hermeneutic. (For a sample of some inadequate versions of the adjustment needed, *see* Ricoeur 1981; Gadamer 1976; and Habermas 1988; also, Cassirer 1953-57).

Given an earlier argument to the effect that the ontic and the epistemic cannot be disjoined, it is impossible to avoid altogether tendering some remarks about the ontic distinction of human selves; but it would deflect us from our chief purpose to attempt to resolve the obvious ontic puzzles that the admission of the *sui generis* emergence of the Intentional (the culturally significative) entails.

It may be enough to say that what makes the cultural both emergent and *sui generis* rests with the conceptual irreducibility of language in physical or biological terms. This of course marks the essential difference between biological emergence within the natural world (*see* Margolis 1987) and the emergence of the Intentional. The entire question of the objectivity of the social sciences and the humanities rests on that condition: if it collapses, then, as has already been conceded, the human studies would be little more than a specialty within the natural sciences.

The other ontic question that bears in an important way on the epistemic concerns of the human studies rests with the fact that the emergence of the cultural and (paradigmatically) the culturally informed mental life of humans counts as a "third" option (largely ignored in analytic philosophies of science) between dualism and reductionism. For, of course, if the cultural is indissolubly incarnate in the natural (call that doctrine *incarnatism*), then the bare description – *a fortiori*, the explanation and, in particular, the causal explanation – of human events ("utterances," in the nominalized sense suggested for scanning art, texts, speech, history, institutions, and activities) will be fundamentally affected in ways that cannot be extrapolated from the description and explanation of mere natural events. (It has already been noted that the emergence of the cultural adversely affects supervenientism and the "closure of the physical"; hence, also, the prospects of reductionism.)

The emergence of the Intentional makes no sense – ontically – apart from a realism that admits "symbiosis" and "incarnation." But, epistemically, all truth-claims in the human studies center on the interpretability of the Intentional features of cultural "utterances" – whether in speech, art, deed, or history. Broadly speaking, *interpretation* is description (or an analogue of description) directed at objectively discerning the linguistic or semiotic or similarly significative import of Intentionally qualified phenomena. Put in the simplest terms: we are, in cultural contexts, normally interested in just these sorts of inquiry. Thus, we may (1) try to fathom the bare sense in which, say, a particular gesture or remark constitutes an insult; and, fathoming that, we may (2) try to determine what the effect of uttering that insult is or was; or, additionally, we may (3) try to construe the utterance in a narrative way that fits (without precluding the causal process) our norms of rational behavior.

There is no antecedent reason why such undertakings should not admit of a *sui generis* rigor that deserves to be called "objective," even if the conditions of objectivity cannot be the same as those thought to obtain in the natural sciences. For, clearly, Intentional properties may be – in fact are – very different from the properties admitted in the natural sciences. It would be very odd if, admitting Intentional properties to be real, we insisted (without argument) that objectivity in the human studies must abide by precisely the same logic, the same norms of rigor, the same explanatory forms, the same methodology, that are reasonably invoked in the natural sciences.

If the matter is empirical at all, a proper comparison of the characteristics of each sort of discipline should be required. For example, Davidson holds (1980b) that "explanations by reasons" (rationalization or narrative explanation) must be a species of causal explanation, because appropriately attributed "reasons" for acting thus and so may also be the causes of one's acting thus and so. But that is plainly a *non sequitur* and very likely false (or false, if physicalism or supervenientism is false). Davidson's conjecture cannot rightly rely on the validity of such doctrines (on antecedent grounds): that is just what is being resisted in this account – that is, naturalism.

The admission of the three sorts of inquiry just mentioned is not meant to be inclusive or systematic, merely a selection of salient concerns that draw attention to changes in our conception of what a science might be taken to be, *if* their realist standing were admitted. For example, the second sort of inquiry precludes the "closure of the physical" as well as nomological necessities in any sector of the world in which culturally incarnate forces are acknowledged. And both the second and third sorts of inquiry preclude any obvious form of objectivist neutrality – though not any merely reasonable construction of what to regard as the operative conditions of objectivity. The fact that admitting Intentional phenomena utterly baffles the familiar strictures positivism and the unity of science program once seriously proposed testifies to the improbability of ever invoking their would-be canons in cultural studies. The picture that is now emerging appears to be coherent, certainly not unreasonable.

It is the first of the three sorts of inquiry mentioned, however, that is the nerve of the social sciences and humanities. For the second and third presuppose the would-be rigor of the first, which is often dubbed "hermeneutic": that is, the objective interpretation of texts, speech, artworks, actions, histories, institutional and traditional practices. Questions of objectivity in the human studies are bound to be as vexed and as contested as any that belong to the natural sciences – for instance, regarding the validity of understanding the meaning of an Intentional *denotatum* (a text, say) by way of fixing the original agent's (or utterer's) productive intention (what, in literary analysis, is known as Romantic hermeneutics); or, alternatively, whether it is possible to understand a text produced in the past in terms that capture the conditions of objective meaning in the past, without being affected by the consensual conditions of understanding that first shaped us as the apt interpreters that we are (*see*, for instance, Gadamer 1975; Beardsley 1970; Hirsch 1967).

Without resolving such questions here, we may remark that, on any defensible theory, the objectivity of interpretation will be constrained by the salient features of the Intentional world: in particular, that, where meanings and significative structure are concerned, there must be some benign form of "meaning holism" that obtains (to revert once again to Fodor's term, though not to Fodor's assessment); that Intentional properties are *determinable* in a *sui generis* way, that is, are *not* characterizable in the same way in which the *determinate/determinable* idiom is applied in the physical sciences; and that Intentional properties are historicized and alterable as a result of being interpreted – or interpretable. (Physical properties are *never* altered as a result of merely being described.) There are other peculiarities that infect Intentional properties. But these are surely among the most noteworthy.

Certainly we see that *if* such peculiarities run true, then the "logic" of the human studies must be very different from the familiar canon assigned the natural sciences. What is worth mentioning about these differences is at least: (*a*) that they do not (need not) produce chaos or paradox or self-contradiction; (*b*) that we ourselves are distinctly interested in developing the possibilities they promise in interpretive contexts; and (*c*) that the rigor they invite (if conceded at all) is demonstrably reconcilable with the more standard views that hold in the natural sciences – always provided, of course, that we cannot, *a priori*, impose on either discipline any modally necessary constraints (*de re* or *de cogitatione*) with regard to reality, logic, conceivability, rationality, or the like.

These last three features deserve a closer look. For if they do not automatically produce incoherence – they assuredly do not – they go a long way toward altering in the profoundest sense the very idea of a rigorous and rational discipline that might be called a human or social science or might affect the actual rigor of the natural sciences. The most strategic of the three features mentioned is, of course, the *sui generis* "determinability" of Intentional properties – which, as has already been suggested, implicates some form of meaning holism.

The point is that we must construe the various holisms of cultural life as benign enough to match the actual fluency of language and linguistically informed activity. Once again, the *lebensformlich* effectiveness of reference and predication shows the way. But the *sui generis* "determinability" of Intentional phenomena remains the decisive puzzle – both ontically and epistemically. Certainly, meanings and intrinsically significative structures ("symbolic forms," for instance, in Susanne Langer's useful but ill-defended gloss on Cassirer {1953-57}) cannot be regarded as determinate "properties" of a kind at all like the kinds assigned physical objects. It may be disputed whether physical properties *are* or must be crisply determinate, or whether it is even clear what it means to say that they are: the matter is complicated by what has already been said about the *lebensformlich* standing of predication. Is the red, for instance, of particular red objects a "determinate" property? Are vague properties not real properties or, in spite of appearances, determinate though vague? Is a straight edge determinately straight?

These are not troublesome questions in the present context. We obviously have a working sense of the determinacy of physical properties, even where our theories are found wanting. But there is a decisive contrast that remains, when all such worries have been met, that continues to segregate "meanings" and physical properties and therefore affects the epistemology of the disciplines that are concerned with each.

Consider that physical properties are *both* determinate *and* determinable in a perfectly straightforward sense: any familiar predicate of the physical sort is, we may say, "determinable" even when it is "determinate," in that it can be made *more* determinate through greater specificity and that, for any degree of such determinacy, there is no infimate or final determinacy such that no further determination is possible. That is more or less what we mean when we speak of the determinacy of physical properties. The notion obviously affects what we are willing to regard as admissible knowledge of the physical world.

The picture is altogether different with meanings and significative structures. Nothing, certainly no physical object, *has* determinate "meanings" in whatever sense physical objects *have* determinate physical properties. Meanings (and Intentional structures) are rightly predicated of suitable Intentional *denotata*; and, in that purely formal sense, texts, artworks, actions, histories, and the like do *have* "meanings," do possess significative import of some kind. Nevertheless, in individuating Intentional *denotata*, it remains a matter of profound dispute – in fact, there *is* no settled general theory of the (predicated) "possession" of meanings that compares with the accepted sense in which physical objects "have" (predicated) "natures" (*see* Beardsley 1970) – whether the "natures" of Intentional entities are, must be, can be, comparably determinate in the sense in which physical properties are said to be both determinate and determinable.

The coherence of the question has already been assured by what was said earlier about the *lebensformlich* standing of reference and predication *and* the asymmetry between fixing "number" and "nature" in general and, in particular, as between physical objects and Intentional (that is, embodied, Intentionally qualified) objects. The idea has already been broached that texts and artworks and actions can be as determinately denoted as are natural objects. (That is: bearing in mind that referential fixity cannot be gained by predicative means.)

The decisive difference between the determinability of physical properties and that of Intentional properties is simply that increasingly determinate specifications of the latter *cannot usually be arrayed as an ordered set of increasing determinations of a determinably constant property.* The determinate properties of Intentional *denotata* are *not*, *qua* determinable, determined in the same sense accepted in ordinary discourse about physical objects.

In Intentional contexts, interpretations of meaning are compared as alternative, possibly competing, *determinate* ascriptions; their *denotata* remain openendedly *determinable*, chiefly by being open to plural – often incompatible – interpretations. That is never true of physical objects. Our theory of physical properties commits us to holding that, *once* objectively determined, physical properties prove to be additionally *determinable*, without risking the objective standing of the valid (determinate) properties that are thus far precised. But, conceding the "meaning holism" of the cultural world and the *sui generis* nature of meanings, no similar practice can be counted on. For example, literary *genres* are not sufficiently like natural-kind kinds that we can ever apply to *genres* anything like the precising practice that holds in the physical sciences (*see* Margolis 1992). In specifying the meaning of "baroque" within the extension of "baroque painting," it is often remarked that the sense in which it is determinable changes with new attributions. The "determinate"/"determinable" idiom is read in two entirely different ways in the natural sciences and the human studies.

The interpretation of a text or artwork or history is as *determinate* as we can defend. If we add more detail to a particular account – think, for instance, of the accumulating detail Roland Barthes has collected in his extraordinary reading of Balzac's *Sarrasine* (Barthes 1974) – where the *whole* interpretation has been severely contested by more conventional readers – the *serially added* detail cannot rightly be construed as simply progressing in the way of an increasingly determinate account of the further determinability of some first determinate interpretation. No. The new provisions, read in accord with the hermeneutic holism of the text's putative meaning, is simply *another* (determinate) interpretation, however plausibly linked or opposed to prior readings. Think of the proliferating histories of World War I: there is no rulelike way of ordering the increasing or added detail of later histories relative to earlier ones, except Intentionally. (*See* Aron 1961). Furthermore, if, say, Wordworth's Lucy poem may be objectively interpreted in divergent ways that cannot be reconciled within a single interpretation (*see* Beardsley 1970), we begin to see the sense in which the objectivity of interpretation favors the open determinability of pertinent *denotata*, where every valid reading is as determinate as it can be.

6. OBJECTIVISM AND RELATIVISM

There's no doubt that the ascription of meanings and significative structure trades on the entrenched practices of interpretation. But the metaphysics of Intentional properties is such that what will count as knowledge in the human studies will have to be very different from what holds in the physical sciences; although of course the two practices remain formally compatible and although, on the incarnatist doctrine, interpretation, much like reference and predication, implicates within its own compass the methodological rigors proper to the physical world. Symptomatically, Davidson (1980b), in his well-known analysis of what an "action" is, offers a definite prescription for reidentifying physical movements but none for Intentionally qualified actions as such; for, clearly, despite his adherence to nonreductive materialism, Davidson is persuaded that some form of supervenience (construed in modally necessary terms) can always be counted on. Yet, *if*, as Davidson also admits, the Intentional (or intentional in the psychological sense) is subject to some form of meaning holism (say, the model of rationality), then, on his own view, supervenience cannot ever be confirmed.

It is in fact the failure of reductionism and supervenientism and eliminativism and strict functionalism, on both empirical and logical grounds, that leads us in the direction of conceding the radical difference between the two readings of the determinate/determinable idiom examined just above. But if so, then no physicalism or naturalism of Davidson's sort – or of anything akin to it – can be expected to inform our understanding of what should count as a reasonable picture of the objectivity of such disciplines as those of history or interpretive criticism in the arts – or, indeed, of sociology or economics.

That is an astonishing defect. For, of course, as far as the human studies are concerned, there is little point to a reductive or supervenientist theory of action that does not – and perhaps cannot – provide an operative sense of the conditions of validity on which alternative (Intentional) descriptions of action objectively apply. One cannot help noticing, for instance, that, within the terms of historical narrative, the physicalist's criterion for individuating actions (by individuating physical movements) has almost no relevance: how the killer's fingers were crooked around which gun in the assassination at Sarajevo has almost no bearing at all, except to refute the denial that the event could ever have happened. (*Contra* Davidson, a mere physical *movement* could never establish *that* a murder had occurred.)

You see, from these considerations, how difficult it would be to claim either that the objective findings of the human studies rightly conform, or must conform, to the exemplary rigor assigned the physical sciences; or that interpretations of the Intentional world (or the causal or rationalizing explanations proper to that world, which depend on would-be objective interpretations and which are themselves interpretively informed) must, or even can, in principle, yield any single, neutral, context-free characterization that matched the familiar objectivist conception of the natural sciences favored by positivism, the unity of science program, and the naturalisms drawn from Quine's or Davidson's or cognate conceptions.

The inseparability of ontic and epistemic questions and the symbiotized realism implicated in reference and predication utterly defeat objectivism in the natural sciences. Hence, given the metaphysics of Intentional properties and the *sui generis* sense in which they may be said to be both objective and "determinate," it is quite impossible that anything like objectivism could be defended among the human sciences.

What we had earlier concluded about the epistemic standing of any empirical inquiry was that objectivity was, ultimately, *lebensformlich*, consensual without yielding any prior or necessary or privileged criteria of objectivity, *and* that, if that were so, then (abandoning neutrality and presumptions of context-free inquiry) objectivity could never be more than a reflexive, critical, artifactual, historically alterable, insuperably provisional conception of what best to posit as the operative marks of realist standing. In short, we find ourselves drawn (as we have been drawn before) to a constructivist reading of realism (opposed to objectivism). That now turns out to be as hospitable to the distinctive work of the human studies as it is to the natural sciences.

There is, therefore, no reason to think that developing a conception of objectivity specifically fitted to the ontic and epistemic peculiarities of the human sciences must lead "somewhere" to incoherence or paradox. All that needs to be granted is that *if* the cultural world is admitted to be real (hence, that *we* ourselves are real – that we are encultured selves), then the real world will be seen to include *denotata* that cannot exist independently of our mental life and that such *denotata* have "natures" that are *not* determinable in the way reckoned to hold among merely physical objects (*contra* Devitt 1991 and Searle 1995, for instance).

The lesson has been drawn before. But, now, it is not so much that we see that *if* we wish to secure the objectivity of inquiries of the kind favored in the social sciences and humanities we will have to depart from the canon of objectivity that holds in the physical sciences; it's rather that there is no way to save objectivism even among the natural sciences and that what must be yielded there proves to be straightforwardly reconcilable with what appears to be needed to make sense of the distinctive objectivity of the interpretive and explanatory work of the human studies.

That is something of a windfall – and a welcome benefit. You see, for instance, that, on the constructivist thesis, there *cannot* be any prior regulative function assigned to truth – say, in the correspondentist way (or, by any "deflationary" or "disquotational" reading of truth or correspondence): *see* Ramsey 1931; Leeds 1978). But if there is no such prior function to save, then objectivism has no epistemic grounds at all – unless it can count on one or another form of foundationalism or cognitive privilege. All that is surely irretrievably gone by now; our conceptions of truth, like our conceptions of objectivity and rationality, must be artifacts of our own history (*contra* Putnam 1994), but not for that reason indefensible at all. It is only the inertia of vestigial forms of objectivism, transcendentalism, apodicticity, and the like that cause us to wonder about the concessions being recommended. On the argument, those concessions have been long overdue.

What is often not perceived is that the executive interests that govern our inquiries in the natural sciences are relatively assured, though their criteria are not; and that there is nothing in the human studies that is as stable or as explicit as the interests that drive the other disciplines. There's no secret there. Whatever else is true, the physical sciences are committed to effective prediction and technological control and invention (*see* Hacking 1983); hence, whatever of systematic explanation (involving non-Intentional *denotata* and non-Intentional causality) facilitates these interests (as we may call them) contributes to our intuitions regarding what to count as objectivity among those sciences.

But *if* the work of the human studies is inherently Intentional, and if the Intentional is intrinsically interpretable, and if the interpretable is never more than "determinable" in the *sui generis* sense sketched a moment ago, then it is quite impossible to apply the predictive and technological marks of the physical sciences in any simple way to the phenomena of the human studies – because *whether* the Intentional yields comparable predictive and technological power itself depends on how, precisely, Intentional regularities can be discerned! *That* cannot now be strictly governed by prior constraints of predictive or technological effectiveness cast in *non*-Intentional terms, although it is true enough that we should expect these different concerns to be ultimately compatible. Interpretation will always be subject to the peculiar determinability that belongs to what is constrained by the various forms of meaning holism and the ontic and epistemic pecularities of meaning. That is what is neglected by the physicalists and the naturalists.

How then should those peculiarities be finally characterized? The most important constraint insists that the Intentional is *lebensformlich*; hence, that the objectivity of meanings and significative structures is essentially a form of reflexive understanding, a function of our native capacity as apt speakers of a common language and as apt agents sharing common practices. This means that there cannot be any radically mistaken continuation of the Intentional habits and practices of the aggregated members of a viable society; *and* that, in principle, there need be no unquestioned criteria of correct interpretation and no necessary truths about the conditions of determinate meaning.

That is precisely the point of genius of Wittgenstein's notion of a form of life and of language games – which renders utterly otiose the "third realm" conjectures of Frege (*see* Dummett 1991) and Popper (1972) and similar-minded theorists who (fearing the flux of history) invent an entire realm of abstract meanings and allied "entities" in order to vouchsafe an objectivism regarding the Intentional (that they could not otherwise secure).

The point of the entire foregoing argument is that all such "Fregean" speculations are for nothing (like the physicalist and naturalizing speculations that have followed them): the "realm" intended remains as inaccessible as any world of Platonic Forms – and is as extravagant – if *lebensformlich* regularities prove viable at all. Broadly speaking, the interpretable world is the world of human history; and the world of human history is the world of collective practices changing at a pace slow enough to permit an aggregated community of selves, apt as speakers and as cultural agents, to recover spontaneously and objectively the meaning of the Intentional world spanning their own society's past.

That is, their being encultured ("second-natured") signifies their competence to recover their own history and traditions in the only way they can – consensually. They continually reconcile the would-be recovery of their own past with their own aptness for extending those same traditional practices acceptably. They understand and perceive the past in terms of their present aptitudes, and their present aptitudes are formed as they are by their own enculturing past. This is very close to what Gadamer (1975) intends by his twin notions of the "fusion of horizons" (*Horizontverschmelzung*) and "effective-historical understanding" (*wirkungsgeschichtliches Bewusstsein*). Gadamer's formula is not unhelpful, but it hardly probes beyond the generic epistemic aptness of selves for discerning Intentional structures.

What is particularly clear now, at the end of our century, is that there are indefinitely many competing, even irreconcilable, "strands" of (our) enabling history and culture that can claim as much objective standing as any other strand – *if* indeed any *can* claim objective standing. Gadamer himself pretends that that there is a "classic," even timeless, convergence of all historical traditions, something close to unchanging human values (despite the flux of history) by which historied scatter is effectively constrained – in fact, made "Hellenic." But that can hardly be more than a pretty story.

It is closer to the truth to admit that the convergence and divergence of traditions and sub-traditions are themselves artifacts of our own variable consensual tolerance, and that the boundaries and horizons of a well-formed practice are continually and diversely redrawn by evolving practice. We interpret our world in accord with our tradition, but *there is* no determinate tradition that we apply in doing that. Traditions are determinable but not determinate (Intentionally) in the *sui generis* sense already adduced. That too may be put in Wittgenstein's terms: for, to speak a language, Wittgenstein says, is "to follow a rule" – but there is no rule that we follow. Rules and traditions are predicative idealizations, not *denotata* in the relevant sense.

Viewed thus, the only objectivity that can be accorded the work of the human studies requires identifying all the viable interpretive strands of our common tradition – without foreclosing prematurely or too narrowly on what to include or exclude – for instance, hurrying to exclude too quickly new interpretive possibilities that could be so effectively grafted onto the trunk of our tradition that they themselves could become, in time, part of the most reliable sources of our further practice. Freudian psychoanalysis and Marxist political economy are among the best-known instances of such innovations. But, even locally, one may concede the objective standing of Barthes's analysis of *Sarrasine* or Ernest Jones's (1949) Freudian reading of *Hamlet* or George Thomson's (1941) Marxist reading of Aeschylus's *Oresteia*. (It's not their quality that is at stake, it's only their power to be accepted and absorbed.)

If you consider events as complex as the Vietnam War, it should be clear that much that is Intentionally freighted – for instance, actual skirmishes, troop movements, and the like – is not likely to require interminable dispute among opposed interpretations. Dispute usually arises at a higher level. The same holds for massive parts of ordinary discourse. Inevitably, in featuring controversial Intentional elements, we set such elements against a backdrop of relatively uncontested elements (equally Intentional) that interpretation, must accommodate.

It is also true that since Intentional things are embodied in physical things, certain culturally entrenched linkages – for instance, how the usual mode of greeting is embedded in particular physical movements – is bound to set convenient and additional constraints on interpretive invention. But conceding all that, and conceding as well that "meanings" cannot be discerned (in the paradigmatic sense) except in language and linguistically informed "utterance," whatever determinacy objective interpretation may claim cannot but be consensual (in the sense that first provides for criterial cooperation).

Hence, if, as is plainly true, the center of gravity of consensual life, the saliencies of communal life, is (are) bound to change with whatever changes obtain in historical experience, then even the fixity of the Intentional past (the "past" of human history) will change, however conservatively, under the conditions of coherent reflection (*contra* Danto 1985; *see* Margolis 1993). (Historical time and the historical past will, of course, be viewed as incarnate in physical time and the physical past.)

This is not to claim that the physical past can be altered by mere interpretation. It is also not to deny that the coherence of human history requires a compensating assurance to the effect that any large change in understanding our past (not because of factual mistake but because of a shift in our interpretive horizon) need not produce paradox. Think of the reinterpretation of ancient history under the advent of Christianity or the spread of Marxist or Freudian or even Nazi conceptions (*see* Hitler 1943), or even such minor inventions as Spengler's (1926) or Toynbee's (1935) conceptual armatures. (Plainly, Spengler's and Toynbee's histories have only a marginal grip on the entrenched views of our tradition – enough perhaps to make them plausible, possibly even compelling in piecemeal applications. But it is worth remarking that the Nazi conception, hateful though it is, surely had a firmer grip on our tradition than did either of their particular interpretive visions. Many will be disposed to reject the prima facie "plausibility" of the Nazi interpretation of history as a result of condemning its values, but that would be to misunderstand the issues at stake.)

Under the circumstances, a fair suggestion is this: let a thousand flowers bloom! The objectivity of competing interpretations of history, or art, or human commitment and activity, or linguistic exchange, or reported dreams for that matter, is, as a general policy, inclusive more than exclusive; plurally reasonable more than uniquely correct; and it tolerates diverging, even incompatible, constructions more than it insists on forcing an exclusionary choice between alternatives that would otherwise appear to be separately valid. Surely the objective "meaning" of the French Revolution must accommodate both the competing ideologies of the strong participants in the original conflict (grasped in the moving present) and a wide sample of the competing ideologies of the changing cohort of contemporary interpreters capable of reconciling the first sort of diversity with that of their own. In accommodating all that, they need not of course abandon their partisan standing in disputes about "objective" meanings and values.

The solution is startlingly straightforward: restrict the scope of a bivalent logic, admit a "relativistic logic" – ad hoc if necessary – and reconcile the resources of the two in order to allow, as objectively confirmable (not jointly true, to be sure, but valid nevertheless), *some* set of otherwise incompatible interpretations, in accord with suitably weakened truth-values or truth-like values. We are entirely free to choose our "logic" as we wish! We are, after all, simply making sense of the conditions under which we understand ourselves and our world. An objective history, let us say, is a history or a set of histories that attempts to exhaust, under conditions of consensual tolerance and openended practice, all the interpretations that can be reasonably mustered in spite of being incompatible or irreconcilable with similarly (not necessarily equally) defensible interpretations. (Here, we are occupied with alternative "logics," not yet with alternative "criteria" – as of objective meaning or objective value – within the scope of the logic we favor.)

Whatever strictures may be placed on the dialectical play of competing accounts will be artifactual and provisional in the sense already sketched. Doubtless there will be different strictures favored by different groups, bearing on different *denotata*, at different times, in contest with different opposing views. But *some* reasonably generous selection within that space will be the most that can be achieved – and will be sufficient for interpretive purposes and for rational commitment so informed. Here, it is a foregone conclusion that "objectivity," "truth," "rationality," "reasonableness," "validity," "confirmation," "norms," and related notions will be "constructed" in the sense already remarked.

The important point to remember is that these concessions do not amount to endorsing the anarchic policy of "anything goes" (*see* Feyerabend 1975). It will only seem to do so, because of the vestigial effects of the objectivisms that have for so long dominated Anglo-American and much of continental European thinking through the twentieth century. (Even Feyerabend's anarchism, of course, is a blast at the prevailing objectivisms, not the advocacy of complete unreason.)

The older canon is gone now, or at least stalemated so effectively by its own internal difficulties that a new constructivism – even a historicism and a relativism – begin to seem both viable and worth defending. At the very least, the close analysis of the human studies no longer permits the following heterodox themes to be discarded as outlaw doctrines: (i) relativism; (ii) the historicity of thought; and (iii) the continual reinterpretability of the historical past perceived from the vantage of an evolving present.

Of these three themes, relativism is the most strategically placed. All that it requires from the canon is that bivalence not be viewed as exceptionless or necessary or universally binding in the modal sense, wherever substantive inquiries are involved. The other two doctrines are more narrow-guaged, more explicit about the cultural world. Nevertheless, relativism is not worth defending if it is no more than a formal thesis about the coherence of certain uninterpreted versions of a many-valued logic (*see* Margolis 1999a; 1999b). The truth is: in the context of items (ii)-(iii), a relativistic logic yields a formidable departure from the theories of Hempel and Popper regarding history and the social sciences and from those of Beardsley and Hirsch regarding interpretation in the arts.

Not much is needed to ensure the coherence of a relativistic "logic" or its compatibility with a limited bivalence. The general requirements have already been hinted at in pursuing other issues, but it would be helpful to collect them in one place. For one thing, the would-be formal constraints of any "logic" are never rightly defended except in the context of the substantive inquiries they are thought to constrain. Logic, semantics, the rules of reason cannot claim any methodological priority over the perceived needs of a particular inquiry: "Fregean" claims, for instance, as in Dummett's (1991) prioritizing semantics over metaphysics, fail to grasp that "semantics" *is* metaphysics "by other means." No questions of coherence or consistency are risked by this concession and everything bearing on relevance may be reasonably secured.

Secondly, relativism need not be an all-or-nothing affair: it may be championed piecemeal in one or another discipline at a time; it may be reconciled with bivalence so long as the two "logics" are suitably segregated (even ad hoc) on grounds of relevance; and relativism itself may be defended (without paradox) on non-relativistic grounds.

Thirdly, relativism need not be construed as the incoherent doctrine sometimes known as "relationalism" (*see* Margolis 1991), which holds that "true" (for relativists) means "true-for-*x*" (some individual *x*), which would signify that no two speakers could ever share the same criterial view of truth, or even the same person at different times, and which would make it impossible to deny that the same claim might be both true and false. That is the ancient charge, of course, Socrates' charge in the *Theaetetus*. But it is a preposterous thesis.

All that is needed to offset such difficulties are the following constraints: (1) "True" and "False" are treated asymmetrically and, in the context in which relativistic truth-values are invoked, "True" is replaced by some member of a set of appropriately selected many-valued values, whereas "False" is simply retained; (2) propositions may be shown to be false, but they may also be shown to be not true without yet being false; so that in the space in which relativistic truth-values are admitted, "False" is opposed not to "True" but to any of the array of the many-valued values admitted (but not the familiar value "Indeterminate" found in three-valued logics); (3) a many-valued logic is a relativistic logic if and only if it admits as valid (as entitled to one or another of the replacing many-valued values) propositions that on a bivalent logic but not now would be incompatible or contradictory (let us say, "incongruent"); (4) the resultant "logic" may at any point in an argument and in however ad hoc a way we please, be reconciled with evidentiary claims that are themselves drawn from a practice constrained bivalently, so long as the assignment of particular values is suitably segregated in accord with relevance constraints; and (5) all further questions of consistency, coherence, contradiction, and the like obtain much as they do in bivalent logics, that is, only where interpreted and context-bound. On the argument already given, contradiction is not a purely formal principle that can be algorithmically applied *from* uninterpreted formulas *to* the meaningful sentences in question. (*see* Wittgenstein 1972.) These considerations surely obviate the familiar *aporiai.*

You must remember that the advocacy of relativism is supported by the substantive analysis of a given domain of inquiry. On the foregoing argument, the entire range of the human studies – the social sciences and humanities in particular – appear to be especially suited to the use of a relativistic logic. That is the upshot of admitting the distinctive sense in which interpretive claims, as well as explanations that depend on prior interpretive findings or that are informed by them, admit the realist standing of the Intentional world. That is the linchpin of the entire matter.

Beyond that, historicism – in the sense in which thought is "historicized" – is itself a form of relativism, perhaps the most radical that belongs to our age, the one that best fits (however disputatiously) the leading doctrines of the Hegelian tradition running, say, from Marx through Nietzsche, through Dilthey, through Heidegger, through Gadamer, through Foucault, and (conceivably) through Kuhn. (Although most of these theorists were opposed to relativism proper.) Here, historicism is definitely not the incoherent relationalism that Ranke (1983) advocated in opposing Hegel.

It takes but a step to conclude that the historical past may be interpretively altered – without paradox. For there can be no privileging of the Intentional import of the past over that of the present. On the contrary, as Gadamer (1975) effectively argues, the "meaning" of the past must be fixed interpretively by agents whose own competence is first encultured in a historicized way.

Once you have all this before you, you see that the argument stands or falls as a single doctrine. There is only one way to defeat it (assuming it is coherent) and that is by way of attacking the ineliminability and irreducibility of the Intentional. There are good reasons for believing that all the familiar counterstrategies fail or are questionbegging, but there is no prospect of ever coming to the end of the quarrel. It is enough for present purposes that the supporting argument be admitted to depend on the continued failure of the countervailing views. To demonstrate the likely futility of all reductionisms and supervenientisms and functionalisms and eliminativisms would, however, demand a much more strenuous effort than what our present question requires.

Joseph Margolis
Temple University

REFERENCES

Aron, R.: 1961, *Introduction to the Philosophy of History*, trans. G. J. Irwin, Beacon Press, Boston.

Bachelard, G.: 1984, *The New Scientific Spirit*, trans. A. Goldhammer, Beacon Press, Boston.

Bambrough, R.: 1960-61, 'Universals and Family Resemblances', *Proceedings of the Aristotelian Society* **60**, 207-222.

Barthes, R.: 1974, *S/Z*, trans. R. Miller, Farrer, Straus, and Giroux, New York.

Beardsley, M. C.: 1970, *The Possibility of Criticism*, Wayne State University Press, Detroit.

Berlin, I.: 1976, *Vico and Herder*, Viking, New York.

Bernstein, R. J.: 1983, *Beyond Objectivism and Relativism*, University of Pennsylvania Press, Philadelphia,

Brentano, F.: 1973, *Psychology from an Empirical Standpoint*; English ed. L. L. McAllister, Humanities Press, New York.

Carnap, R.: 1959, 'Psychology in Physical Language', trans. G. Shick, in A.J. Ayer (ed.), *Logical Positivism*, Free Press, Glencoe.

Carnap, R.: 1995, *The Unity of Science*, trans. M. Black, Thoemmes Press, Bristol.

Cartwright, N.: 1983, *How the Laws of Physics Lie*, Clarendon, Oxford.

Cassirer, E.: 1953-57, *The Philosophy of Symbolic Forms*, 3 vols., trans. R. Manheim, Yale University Press, New Haven.

Chalmers, D. J.: 1996, *The Conscious Mind*, Harvard University Press, Cambridge.

Chisholm, R. M.: 1977, *Theory of Knowledge*, 2nd ed., Prentice-Hall, Englewood Cliffs.

Chomsky, N.: 1965, 'Review of B.F. Skinner, *Verbal Behavior*', *Language* **35**, 56-58.

Chomsky, N.: 1986, *Knowledge of Language: Its Nature, Origin, and Use*, Praeger, New York.

Churchland, P. M.: 1989, *A Neurocomputational Perspective*, MIT Press, Cambridge.

Churchland, P. M.: 1995, *The Engine of Reason, The Seat of the Soul*, MIT Press, Cambridge.

Collingwood, C. G.: 1978, *An Autobiography*, Clarendon, Oxford.

Danto, A. C.: 1985, *Narrative and Knowledge*, Columbia University Press, New York.

Davidson, D.: 1980a, *Essays on Actions and Events*, Clarendon, Oxford.

Davidson, D.: 1980b, 'Mental Events', in *Essays on Actions an Events*, Clarendon, Oxford.

Davidson, D.: 1984, 'Reality without Reference', in *Inquiries into Truth and Interpretation*, Clarendon, Oxford.

Davidson, D.: 1986, 'A Coherence Theory of Truth and Knowledge', in E. Lepore (ed.), *Truth and Interpretation*, Basil Blackwell, Oxford.

Dennett, D. C.: 1991, *Consciousness Explained*, Little Brown, Boston.

Devitt, M.: 1991, *Reason and Truth*, 2nd ed., Princeton University Press, Princeton.

Dilthey, W.: 1989, *Selected Works*, vol. 1, R. Makkreel and F. Rodi (eds.), Princeton University Press, Princeton.

Dretske, F.: 1995, *Naturalizing the Mind*, MIT Press, Cambridge.

Dummett, M.: 1991a, *The Logical Basis of Metaphysics*, Harvard University Press, Cambridge.

Dummett, M.: 1991b, *Frege and Other Philosophers*, Clarendon, Oxford.

Feyerabend, P. K.: 1975, *Against Method*, Verso, London.

Fine, A.: 1986, *The Shaky Game: Einstein, Realism, and the Quantum Theory*, University of Chicago Press, Chicago.

Fleck, L.: 1979, *Genesis and Development of a Scientific Fact*, T. J. Trenn and R. K. Merton (eds.), trans. F. Bradley and T. J. Trenn, University of Chicago Press, Chicago.

Fodor, J. A.: 1975, *The Language of Thought*, Thomas Y. Crowell, New York.

Fodor, J. A.: 1983, *The Modularity of Mind*, MIT Press, Cambridge.

Fodor, J. A.: 1990, *A Theory of Content*, MIT Press, Cambridge.

Fodor, J. A.: 1998, *Concepts*, Clarendon, Oxford.

Foucault, M.: 1977, 'Nietzsche, Genealogy, History', in D. F. Bouchard (ed.), *Language, Counter-Memory, Practice*, trans. D. F. Bouchard and S. Simon, Cornell University Press, Ithaca.

Foucault, M.: 1979, *Discipline and Punish*, trans. A. Sheridan, Vintage, New York.

Frege, G.: 1960, 'On Sense and Reference', in *The Philosophical Writings of Gottlob Frege*, P. Geach and M. Black (eds.), Basil Blackwell, Oxford.

Gadamer, H.-G.: 1975, *Truth and Method*, trans. G. Barden and J. Cumming, Seabury Press, New York.

Gadamer, H.-G.: 1976, *Philosophical Hermeneutics*, ed. and trans. D. E. Linge, University of California Press, Berkeley.

Geach, P.: 1957, *Mental Acts*, Routledge and Kegan Paul, London.

Habermas, J.: 1988, *On the Logic of the Social Sciences*, trans. S. Weber Nicholsen and J. A. Stark, MIT Press, Cambridge.

Hacking, I.: 1983, *Representing and Intervening*, Cambridge University Press, Cambridge.

Hegel, G. W. F.: 1977, *Phenomenology of Spirit*, trans. A. V. Miller, Oxford University Press, Oxford.

Heidegger, M.: 1962, *Being and Time*, trans. J. Macquarrie and E. Robinson, Harper and Row, New York.

Hempel, C. G.: 1965, *Aspects of Scientific Explanation*, Free Press, New York.

Hirsch, Jr., E.D.: 1967, *Validity in Interpretation*, Yale University Press, New Haven.

Hitler, A.: 1943, *Mein Kampf*, trans. R. Manheim, Houghton Mifflin, Boston.

Husserl, E.: 1960, *Cartesian Meditations*, trans. D. Cairns, Martinus Nijhoff, The Hague.

Jones, E.: *Hamlet and Oedipus*, W.W. Norton, New York.

Kenny, A.: 1963, *Action, Emotion and Will*, Routledge and Kegan Paul, London.

Kim, J.: 1993, *Supervenience and Mind*, Cambridge University Press, Cambridge.

Kim, J.: 1998, *Philosophy of Mind*, Westview, Boulder.

Kripke, S. A.: 1980, *Naming and Necessity*, Harvard University Press, Cambridge.

Kuhn, T. S.: 1970, *The Structure of Scientific Revolutions*, 2nd ed., University of Chicago Press, Chicago.

Latour, B. and S. Woolgar: 1986, *Laboratory Life*, 2nd ed., Princeton University Press, Princeton.

Leeds, S.: 1978, 'Theories of Reference and Truth', *Erkenntnis* **13**, 111-129.

Leibniz, G.W.: 1956, *The Leibniz-Clarke Correspondence*, H. G. Alexander (ed.), Manchester University Press, Manchester.

Lorenz, K.: 1970-71, *Studies in Animal and Human Behaviour*, 2 vols, trans. R. Martin, Harvard University Press, Cambridge.

Löwith, K.: 1991, *From Hegel to Nietzsche*, trans. D. E. Green, Columbia University Press, New York.

Margolis, J.: 1984, *Culture and Cultural Entities*, D. Reidel, Dordrecht.

Margolis, J.: 1987, *Science without Unity*, Basil Blackwell, Oxford.

Margolis, J.: 1991, The Truth about Relativism, Basil Blackwell, Oxford.

Margolis, J.: 1992, 'Genres, Laws, Canons, Principles', in M. Hjort (ed.), *Rules and Conventions*, Johns Hopkins University Press, Baltimore.

Margolis, J.: 1993, *The Flux of History and the Flux of Science*, University of California Press, Berkeley.

Margolis, J.: 1995, *Historied Thought, Constructed World*, University of California Press, Berkeley.

Margolis, J.: 1997a, 'Relativism and Cultural Relativity', *JTLA* (Journal of the Faculty of Letters, The University of Tokyo, Aesthetics) **22**, 19-34.

Margolis, J.: 1997b, 'Reconciling Relativism and Cultural Realism', *JTLA* (Journal of the Faculty of Letters, The University of Tokyo, Aesthetics) **22**, 53-68.

McDowell, J.: 1996, *Mind and World*, Harvard University Press, Cambridge.

Moore, G.E.: 1978, 'Proof of the External World', in *Philosophical Papers*, George Allen & Unwin, London.

Oppenheim, P. and H. Putnam: 1958, 'Unity of Science as a Working Hypothesis', in H. Feigl, M. Scriven, and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science*, vol. 2, University of Minnesota Press, Minneapolis.

Popper, K. R.: 1959, *The Poverty of Historicism*, 2nd ed., Routledge and Kegan Paul, London.

Popper, K. R.: 1983, *Realism and the Aims of Science*, W.W. Bartley III, (ed.), Rowman and Littlefield, Totowa.

Premack, D.: 1978, *Intelligence in Ape and Man*, Lawrence Erlbaum, Hillsdale.

Putnam, H.: 1987, *The Many Faces of Realism*, Open Court, LaSalle.

Putnam, H.: 1994, 'Sense, Nonsense, and the Senses: An Inquiry into the Powers of the Human Mind', *Journal of Philosophy* **91**, 445-517.

Quine, W. V.: 1960, *Word and Object*, MIT Press, Cambridge.

Quine, W. V.: 1969, 'Epistemology Naturalized', in *Ontological Relativity*, Columbia University Press, New York.

Quine, W. V.: 1992, *Pursuit of Truth*, 2nd ed., Harvard University Press, Cambridge.

Ramsey, F. P.: 1931, *The Foundations of Mathematics*, Routledge and Kegan Paul, London.

Ranke, L. von: 1983, *The Theory and Practice of History*, trans. Wilma A. Iggers, G. G. Iggers and K. von Moltke (eds.), Irvington, New York.

Ricoeur, P.: 1981, *Hermeneutics and the Human Sciences*, ed. and trans. J. R. Thompson, Cambridge University Press, Cambridge.

Rorty, R.: 1967, 'Relations, Internal and External', in P. Edwards (ed.), *The Encyclopedia of Philosophy*, vol. 7, Macmillan, New York.

Rudner, R.: 1966, *Philosophy of Social Science*, Prentice-Hall, Englewood Cliffs.

Russell, B.: 1905, 'On Denoting', *Mind* **14**, 479-493.

Russell, B.: 1912, *The Problems of Philosophy*, Oxford University Press, London.

Schleiermacher, F. D. E: 1977, *Hermeneutics: The Handwritten Manuscripts*, H. Kimmerle (ed.), trans. J. Duke and J. Frostman, Scholars Press, Missoula.

Searle, J. R.: 1958, 'Proper Names', *Mind* **67**, 166-173.

Searle, J. R.: 1992, *The Rediscovery of the Mind*, MIT Press, Cambridge.

Searle, J. R.: 1995, *The Construction of Social Reality*, Free Press, New York.

Shapin, S.: 1996, *The Scientific Revolution*, University of Chicago Press, Chicago.

Spengler, O.: 1926, *The Decline of the West*, 2 vols., trans. C. F. Atkinson, Alfred A. Knoff, New York.

Strawson, P. F.: 1958, 'On Referring', *Mind* **59**, 320-344.

Taylor, C.: 1964, *The Explanation of Behaviour*, Routledge and Kegan Paul, London.

Thomson, G.: 1941, *Aeschylus and Athens*, Lawrence and Wishart, London.

Toynbee, A.: 1935, *A Study of History*, vols. 1-3, Oxford University Press, Oxford.

van Fraassen, B. C.: 1980, *The Scientific Image*, Clarendon, Oxford.

Wittgenstein, L.: 1953, *Philosophical Investigations*, trans. G. E. M. Anscombe, Basil Blackwell, Oxford.

Wittgenstein, L.: 1958, *The Blue and Brown Books*, Basil Blackwell, Oxford.

Wittgenstein, L.: 1972, *Tractatus Logico-Philosophicus*, trans. D. F. Pears and B. F. McGuinness, Routledge and Kegan Paul, London.

TOM STONEHAM

SELF-KNOWLEDGE

1. MINDS AND EPISTEMOLOGY

A certain conception of epistemology is often seen, by historians of philosophy, as definitive of the modern period in philosophy. This conception structures the epistemological task by a contrast between our privileged or certain knowledge of our own minds and our problematic knowledge of the external world. With this contrast in mind, our knowledge of the external world seems either impossible or inadequate. Even epistemologies which try to take our knowledge of our minds as a foundation for knowledge of the world fail to bestow upon the latter the certainties of the former, because the bridging principles are tentative or probabilistic.

Noticing this weakness in the resultant epistemologies, some philosophers have tried to reject the contrast between our knowledge of our minds and of the world (e.g. Sellars 1956, McDowell 1986). As a general strategy in epistemology, this has great virtue, for it is undoubtedly bad philosophy to bemoan a lack of certainty in our knowledge of the external world, when it is clear that that sort of certainty is inappropriate and unnecessary. We can and should learn to live with defeasibility in our dealings with the world around us (Williamson 1996).

However, it does not follow directly from this lesson about general epistemology that Self-Knowledge is just another form of empirical knowledge. Rather, we can and should ask what are the limits and nature of Self-Knowledge. In so doing, we may well discover that it has a different epistemological character to knowledge of the external world, that it has certain epistemic privileges. More strongly we may even discover that these privileges are essential to thinking, or perhaps a certain type of thinking. And no such conclusion need reflect badly on our empirical knowledge.

In the next section I use Descartes' *Meditations* to distinguish three questions to do with our knowledge of our minds. Then I present and discuss an account of Self-Knowledge called Cartesianism, though we find a reason to doubt it was Descartes' own view in section 6. In section 4 I consider what are the key components of the Cartesian view and look at how various alternatives reject different key Cartesian claims. In section 5 I consider the question of whether it is necessary that minded creatures like us possess privileged Self-Knowledge. Following that, I consider and reject the contemporary dogma that present tense self-ascriptions of belief and desire can be mistaken. The argument here provides a foundation for a positive epistemology of Self-Knowledge, the distinctive feature of which is that the justification of present tense attributions of belief to oneself is independent of how those judgements are brought about, it is, to coin a phrase, a non-aetiological justification. Finally, I consider whether the thesis of Anti-Individualism in the philosophy of mind presents any special problems for Self-Knowledge. Here we turn

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 647-672.
© 2004 *Kluwer Academic Publishers.*

full circle and consider again the relations between knowledge of our minds and knowledge of the world.

2. SELF-IDENTIFICATION, SELF-CATEGORIZATION AND SELF-ATTRIBUTION

A good place to begin an investigation of self-knowledge is with Descartes' *Second Meditation*. The reason for beginning here is not only that Descartes' own view on self-knowledge is a good starting point for critical discussion, but that there are three distinct phases in his account of his knowledge of himself, and these three phases correspond to three questions of self-knowledge which we need to distinguish. He begins the Meditation looking for a certainty which is immune to the doubts he had raised in the *First Meditation*, and famously he hits upon his own existence:

> Does it not follow that I too do not exist? No: if I convinced myself of something, then I certainly existed. But there is a deceiver of supreme power and cunning who is deliberately and constantly deceiving me. In that case I too undoubtedly exist, if he is deceiving me; and let him deceive me as much as he can, he will never bring it about that I am nothing so long as I think that I am something. So after considering everything very thoroughly, I must finally conclude that this proposition, *I am, I exist,* is necessarily true whenever it is put forward by me or conceived in my mind. (AT VII 25)

The core of this argument is that the first person pronoun cannot fail to refer, so whenever I consider the proposition that I exist, the 'I' refers and thus the proposition is true. Furthermore, however many or few things there are in existence, the first person pronoun gives me a way of picking out the one that is me, in a way which is immune to error through misidentification. We can call the problem Descartes is addressing here that of Self-Identification.

He then goes on to consider what kind of thing he has identified in the first argument. He considers various traditional conceptions of human nature and settles on thought:

> At last I have discovered it – thought; this alone is inseparable from me. I am, I exist – that is certain. ... At present I am not admitting anything except what is necessarily true. I am, then, in the strict sense only a thing that thinks; that is, I am a mind, or intelligence, or intellect, or reason ... But for all that I am a thing which is real and truly exists. But what kind of thing? As I have just said – a thinking thing. (AT VII 27)

This passage addresses a question we might describe as Self-Categorization, and the main argument is that I cannot separate in thought my existence from my thinking; that is, I cannot imagine existing and not thinking. Whatever the merits of this argument, and ignoring the use to which Descartes puts it in arguing for the Real Distinction between mind and body, it would seem that the conclusion that I am a thing which thinks is fairly unobjectionable. What is striking, however, is that Descartes reaches this conclusion before establishing *what* he is thinking, before attributing any particular thoughts to himself. He can do this because his argument that he is a thinking thing is loosely logical: it follows that I have thoughts from the first-person premise that I exist.

In contrast there is no attempt at a logical argument for the third claim of self-knowledge that he makes:

> But what then am I? ... A thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions. (AT VII 28)

When Descartes rhetorically challenges this claim, it is to challenge whether all these different activities are really inseparable from me. But we should note the difference between this and the Self-Categorization argument, for in that argument it was claimed that thinking was inseparable from my existence, yet it would hardly be plausible to claim that all the mental qualities listed above are inseparable from my existence. Rather the argument must be conditional, with the truth of the antecedent assumed: if I am aware of doubting, then it is not possible that it is someone else who is doing the doubting:

> The fact that it is I who am doubting and understanding and willing is so evident that I see no way of making it any clearer. (AT VII 29)

Somewhere in the argument Descartes has assumed that he knows about the doubt, that he is not, and cannot, be mistaken that doubting, or willing, or whatever, is going on. Why does he assume this? Well, he must have an implicit theory of Self-Attribution which entails that attributions of mental states are infallible. The questions of how we attribute specific mental states to ourselves, and how those attributions are warranted are the central epistemological questions an account of self-knowledge needs to address. I shall from now on use the phrase 'Self-Attribution' to refer to the ascriptions of mental properties one makes to oneself *directly and without recourse to the evidence one uses to make similar ascriptions to others*. Not all our knowledge of our own minds is Self-Attribution, since sometimes we come to realize that we have some characteristic such as jealousy, or some intentional state such as an aversion to getting wet, by noticing patterns in our behaviour.

3. CARTESIANISM

The clue to Descartes' account of Self-Attribution comes in a discussion of sense-perception:

> For example, I am now seeing light, hearing a noise, feeling heat. But I am asleep so all this is false. Yet I certainly *seem* to see, to hear, and to be warmed. This cannot be false. (AT VII 29)

To see why Descartes thinks we cannot be mistaken about what we seem to see or hear, we need to make clear that seemings are often relativized: it may seem to me that p, but not to you, and it may even seem that q to us or to people in general. Further there is an apparently unrelativized use of verbs such as 'seems' and 'appears' and 'looks' which allows dispute and disagreement. For example, I might say that it looks as if it will rain and you might dispute this by pointing out that the darkness is caused by a solar eclipse. Now Descartes' claim is that I cannot be mistaken about how things *seem to me*. This is because he equates how things seem to me with my conscious experience and the objects of conscious experience seem as they are and are as they seem.

This is especially plausible if we consider such mental states as being in pain or having colour experiences. If I seem to be in pain, that is if I am having the conscious experience of pain, then surely I am in pain, and if I feel no pain, then I am not in pain. Similarly, if I seem to be having an experience of a particular colour, such as red, then I am experiencing red, and if I do not have any consciousness of the experience of red, then I do not have it.

The view was elegantly expressed by Thomas Reid more than a century after Descartes, and taken by him to be simple commonsense:

> When a man is conscious of pain, he is certain of its existence; when he is conscious that he doubts or believes, he is certain of the existence of those operations.
>
> But the irresistible conviction he has of the reality of those operations is not the effect of reasoning; it is immediate and intuitive. The existence therefore of those passions and operations of our minds, of which we are conscious, is a first principle, which nature requires us to believe upon her authority.
>
> If I am asked to prove that I cannot be deceived by consciousness – to prove that it is not a fallacious sense – I can find no proof. I cannot find any antecedent truth from which it is deduced, or upon which its evidence depends. It seems to disdain any such derived authority, and to claim my assent in its own right. (1785, Essay VI, Ch. V)

The view that consciousness gives us infallible access to our mental states, that x is conscious of mental state m if, and only if, x has mental state m, can reasonably be called Cartesianism. It has three main areas of difficulty: (i) it leads to dualism; (ii) some mental states have no conscious character; (iii) we can and do make mistakes.

(i) The argument that Cartesianism leads to dualism makes the modest assumption that any non-dualist metaphysics of mind will have to accept that some claims about the physical world *entail* claims about the mental. At its very weakest, this might simply be a claim such as: because of its lack of suitable internal structure and limited behavioural repertoire, a rock cannot feel pain or experience red. A slightly stronger claim would be that if you have a head full of sawdust, then you are not conscious. Stronger still would be the claim that a creature which displays complex physical behaviour and uses language must be conscious (this is a variant of the claim underlying the Turing Test). Now, given that Self-Attributions on the basis of consciousness are made completely independently of any knowledge of the physical world, and, further, according to Cartesianism, cannot be contradicted by any physical facts, it follows that there is always the possibility of a conflict between Self-Attributions and attributions of mental properties made on physical grounds. Thus, for example, I attribute various thoughts and experiences to myself, but it may turn out upon inspection that my head is full of sawdust, entailing, according to physicalism, that I have no conscious thoughts. Since Self-Attributions are infallible, any such conflict would show the non-dualist to be mistaken.

This connection between Cartesianism and dualism has lead many philosophers convinced of physicalism in one form or another to reject Cartesianism about Self-Attribution. Another response would be to deny that the conflicts are in fact possible. The argument moves from the fact that I do not know what is inside my head when I seem to be in pain to the conclusion that it is possible my head is full of sawdust. However, we might argue that this confuses imaginability with genuine possibility. The physicalist can claim that, given I have all these conscious thoughts and experiences, it is not in fact possible that my head is full of sawdust, though it is imaginable. Ignorance sometimes enables us to imagine the impossible.

(ii) Much more problematic for Cartesianism is the over-emphasis on consciousness. One objection raised is that much of our mental life is not conscious. This objection points out that Cartesianism is most plausible in a restricted range of cases like pain and colour experience, but is very implausible for beliefs and desires: I am certainly not conscious of everything I desire. But there is also a difficulty with even the best cases for the Cartesian. If I am driving and talking to a passenger I might suddenly brake sharply as a reflex response to the car in front braking. Suppose that car's brake lights have briefly illuminated and caused the reflex action, even though I was not conscious of seeing them. The Cartesian must implausibly claim that I did not in fact have a visual experience of red lights. There are also difficulties with pains. Imagine that you have a painful insect bite. As you sit at your desk absorbed in your work, you cease to notice it, you do not have a conscious experience of pain, but as soon as you relax and look out the window, you feel the pain again. The Cartesian must say that the pain ceases while you are concentrating on your work or reject the two-way connection between consciousness and our mental life. If he did the latter, we would remain incorrigible about our mental states whenever we did Self-Attribute, but be susceptible to mistakes of ignorance.

Even more pressing for the Cartesian than these implausible consequences of missing experiences is the fact that most mental states *do not have a distinctive conscious character*. There is something particular that it is like to be in pain, but there is no particular conscious experience associated with, say, the belief that the dog is hungry or the desire to take a holiday. It is not merely that beliefs and desires may or may not figure in your conscious experience, but that even when they do, their conscious character is not consistent across time and circumstance but varies greatly and may be no different from the conscious character of a quite different mental state, therefore it does not serve to identify them. The phrase 'what it seems like to believe the dog is hungry' does not pick out a conscious experience uniquely associated with that belief and thus how things seem to us cannot give us infallible knowledge of all our mental states. It is the dependence upon conscious seemings which is the real problem for Cartesianism as an account of our Self-Attributions.

(iii) The third objection raised against Cartesianism is that we do in fact make mistakes in our Self-Attributions. This is as much a dogma of contemporary philosophy as its contrary used to be, and as such deserves careful consideration, which I will give in the section on Fallibility below.

4. ALTERNATIVES TO CARTESIANISM

Cartesianism about Self-Attribution combined five theses:

1. Objectivism
Self-Attributions are truth-evaluable judgements of distinctly existing mental states.

2. Asymmetry
The epistemic ground or warrant of Self-Attributions is independent of anything else we may know in some other manner.

3. Necessity
The Asymmetry claim is not due to a contingent feature of our minds.

4. Infallibility
We cannot make mistakes, either of misjudgement or ignorance, in our Self-Attributions.

5. Consciousness
Consciousness is the faculty which enables us to make Self-Attributions.

The fourth and fifth theses are the most distinctive of Cartesianism, so if we are to reject Cartesianism we should consider whether also to reject any of the first three theses.

One option would be to deny Necessity. One might do this if one thought that Self-Attributions were based on interior perception analogous to sense perception. Locke probably held this view, and Berkeley was certainly at pains to deny it. More recently it appears in the work of functionalists and physicalists such as David Armstrong (1963) who hold that our minds are contingently equipped with a self-scanning mechanism. This view differs from Cartesianism in holding that (a) it is possible not to have this faculty of introspection, whereas for the Cartesian minds are necessarily conscious, and (b) there is no a priori reason why introspection must be reliable, let alone infallible, though natural selection is likely to have dealt pretty swiftly with creatures whose introspection was grossly unreliable.

Another alternative would be to deny not just the Necessity of the Asymmetry, but its very existence. Such a view would hold that we know about our own minds in very much the same way we know about other people's minds, namely by observing our behaviour. The appearance of an Asymmetry between Self-Attribution and knowledge of other minds is an illusion brought about by our greater intimacy with ourselves. This view naturally goes with behaviourism and was explicitly endorsed by Gilbert Ryle (1949, ch.6). It may also be forced upon a philosopher who takes our attributions of thoughts to others to be the exercise of a sophisticated theory (sometimes called folk-psychology), for if attributing thoughts to others involves inference from behaviour via a psychological theory, then it is hard to see how we could attribute the same thoughts to ourselves without the theoretical inferences from behaviour (Lyons 1986; Gopnik 1993).

The denial of Asymmetry can take a more or a less plausible version. The less plausible version says that we only come to form beliefs about our own minds on the basis of observing our behaviour and listening to what we say, in exactly the way we form beliefs about other minds. This is implausible simply because we can often answer questions about what we think or want without having noticed our behaviour, and sometimes in advance of that behaviour. If you picked a random person working in the library and asked me whether they wanted a chocolate ice-cream right now, I would not be able to answer without more information. But if you asked me the question about myself, even when I had spent the whole morning at a desk reading philosophy, I could answer straight away. It would seem hard to deny that we can form opinions about our own minds directly and without referring to the evidence we use for other minds. So the more plausible version of the denial of Asymmetry will say that, however we come to make Self-Attributions, whatever the mechanism might be, their *justification* is no different from judgements about other minds. In other words, our ability to make judgements about ourselves directly and without recourse to evidence does not signify an *epistemologically* privileged access to our own minds. It is rather like the ability some people have to tell the time directly and without looking at a clock: that it seems to them to be 3.30pm or two hours since lunch, does not justify their claim, but merely explains why they have made it.

In the case of instinctive time-telling, the most likely justification will be: it seems that way and how it seems to me has proved to be reliable. One option for the denier of Asymmetry is to say that the justification of Self-Attributions is of the same form: it strikes me that I want an ice-cream (or whatever) and in the past these instincts or intuitions have proved to be reliable. Though we rarely use a similar justification when talking of other minds (though we might with people we know very well, such as a close sibling or a spouse), this does not re-introduce an Asymmetry because the reliability of one's intuitions must be determined empirically by comparing them to the behavioural evidence.

A final option would be to deny Objectivity. Assuming that judgements about the physical world and about other minds are Objective, this view would maintain Asymmetry and Necessity. This view is inspired by some comments of Wittgenstein's in *The Blue Book*:

> The difference between the propositions 'I have a pain' and 'he has a pain' is not that of 'LW has a pain' and 'Smith has a pain'. Rather it corresponds to the difference between moaning and saying someone moans. (1958a, 68)

Wittgenstein's point is that saying 'I am in pain' is a way of expressing your pain, like moaning or wincing, and as such is neither true nor false. A wince can be insincere or deceptive, but not false. But if 'I am in pain' is not a truth-apt statement, then the claim to know that one is in pain is not really a claim to genuine knowledge. The difference between saying 'I am in pain' and 'I know that I am in pain', if there is one, is akin to the difference between moaning and screaming. Of course the analogy between linguistic expressions of pain, which we can call avowals, and non-linguistic ones such as groans and winces, should not be over stretched. In virtue of using language, the avowal has a logical status, has connections to other parts of language. If the dentist tells you that the drill will not hurt you, and you then scream in pain, you have shown him mistaken but you have not contradicted him. But if you say 'That hurts', you have contradicted him.

The idea of Self-Attributions as avowals can be extended beyond the case of sensation (see Hacker 1990, essays II and V). Wittgenstein hints at this in the following passage about the contrast between our Self-Attributions and our knowledge of other minds:

> "I can only *believe* that someone else is in pain, but I *know* it if I am." – Yes: one can make the decision to say "I believe he is in pain" instead of 'He is in pain". But that is all. – What looks like an explanation here, or like a statement about a mental process, is in truth an exchange of one expression for another which, while we are doing philosophy, seems the more appropriate one. (1958b, 303)

The thought here is that in saying 'I believe p', as opposed to 'p', it looks as if I am saying something about myself, my beliefs. However, the difference between saying 'I believe p' and 'I know q' is one of emphasis. In effect, to say 'I know q' rather than just 'q' is an attempt to stop the audience questioning whether q is true or not, and to say 'I believe p' is almost inviting the audience to disagree. An avowal of a thought, emotion or sensation, then, can come in different forms to serve different conversational purposes, and few, if any, of those purposes require me to describe my mind in the way that you might.

This Wittgensteinian view of Self-Attributions can be summed up as the claim that a syntactic similarity between avowals and knowledge of other minds obscures an important grammatical difference. A variation of the non-Objectivist account of Self-Attributions holds that the difference between avowals and third person attributions is not grammatical but semantic: both are truth-apt judgements or descriptions, but the truth of the third person attributions depends upon the avowals, upon what the subject says about her own mind (Wright 1989a, 1989b; Heal, 1994). The view takes a provisoed bi-conditional claim about the relation between avowals and the subject's mind to be explained by the fact that the state of mind partly consists in the disposition to avow that state of mind in certain circumstances. Thus:

(A) In circumstances C, x avows he* is m if and only if x is m,

is true, but not because in those circumstances avowals are based upon an infallible cognitive access to one's mind, but because (A) is a necessary truth about the mental state m.

If we call Self-Attributions with privileged epistemic warrant 'Self-Knowledge', then both non-Objectivist positions can be seen to be denying that we have Self-Knowledge. According to the Wittgensteinian, one does not normally have *knowledge* of one's own mind at all, one simply has thoughts, emotions and sensations and the ability to avow them. Since one can no more express someone else's pain by avowing than by groaning, there is a necessary Asymmetry, but not an epistemological one.

On the second version of non-Objectivism it is possible to talk of our knowledge of our own minds, but this has no epistemological connotations. This knowledge is not something we achieve but something we cannot avoid having, given the nature of the mental:

> knowing of one's own beliefs, desires, and intentions is not really a matter of "access to" – being in cognitive touch with – a state of affairs at all.
>
> ... the authority standardly granted to a subject's own beliefs, or expressed avowals, about his intentional states is ... not a by-product of the nature of those states, and an associated epistemologically privileged relation in which the subject stands to them, but enters primitively into the conditions of identification of what a subject believes, hopes, and intends. (Wright 1989a, 632)

5. THE NECESSITY OF SELF-KNOWLEDGE

Those who deny Asymmetry or Objectivity also deny that there is any distinctive epistemology of Self-Knowledge. This is often described as denying privileged access. The main argument for these views is the fundamental inadequacy of any theory of privileged access. And yet that we have privileged access to our own minds is a very natural and commonsensical claim. So it is reasonable to proceed by looking more closely at the epistemology of privileged access to see if it can be made to work.

The big divide in theories of privileged access is over Necessity. We should first distinguish five claims that the proponent of Necessity may be making:

1] Necessarily: We are able to make (some) Self-Attributions directly and without recourse to evidence.
2] Necessarily: We have Self-Knowledge.
3] Necessarily: If we are able to make Self-Attributions directly, then when we do so, we usually do so knowledgeably.
4] Necessarily: All Self-Attributions are true.
5] Necessarily: If we have some mental feature, we are able to make a Self-Attribution of that.

As was argued in the discussion of Cartesianism above, 5] is simply false: though we can often Self-Attribute love or jealousy or anger or conviction or desire on the basis of some introspectible symptom, there is no guarantee that we can. For example, the whole plot of Jane Austen's *Mansfield Park* turns upon the different ways in which people come to recognize the dawning of love. Nor can one rescue 5] by restricting it to a subclass of mental states, that is by simply ruling out the emotions, for our convictions may dawn on us equally indirectly. For example, someone interviewing candidates for a job may notice that she has been favouring candidates with a certain qualification. This may reveal, to her as well as to us, that she believes that qualification to be relevant to the job.

Rejecting 5] should help clarify 1]. The claim here is simply that, of necessity, minded, rational beings like us can sometimes and in some circumstances, make judgements about our own minds directly and without recourse to any publicly accessible evidence. The truth of this claim depends upon what counts as being sufficiently like us. It is clearly possible for there to be creatures who have a mental life and yet lack the psychological concepts involved in Self-Attribution. For example, a dog can clearly think, can make decisions (which rabbit to chase, whether or not to obey a command), but does not appear to have any conception of the mind. Thus a dog may react to your anger or pleasure, but as can be seen from the limited range of its reactions, does not think of you *as* angry. Rather, he reacts to you as a natural phenomenon, not as a being acting on the basis of beliefs, desires, emotions and intentions. [If you doubt any of these claims about actual dogs, the point only needs the possibility of such creatures.]

So someone who wants to defend 1] will have to say that dogs (or whatever), if they have mental lives at all, have minds very different from ours. One could either say that while dogs have some mental life, they are not rational beings like us, and it is our rationality which requires Self-Attribution, or one could say that having a mind at all like ours requires Self-Attribution, so our psychological concepts can only be applied metaphorically to dogs.

Since 2] entails 1], any proponent of 2] will be committed to 1]. There are three major lines of argument for 2]. One is the Cartesian argument that minds are essentially conscious and consciousness brings with it Self-Knowledge. We saw above that not all our mental lives are essentially conscious, and even for that which is conscious, consciousness itself does not fully explain Self-Knowledge.

The second argument has been put forward by Donald Davidson (1991). He holds that knowledge of the objective world, knowledge of other minds and knowledge of oneself are all interdependent. The crucial stage in the argument, for our purposes, is that thought of items in the external world requires a triangulation, with oneself and another person forming the base and the object of thought at the apex. This image is meant to express the thought that one only has a determinate thought about a particular thing in so far as someone else, an interpreter, can attribute one a thought about that object. Since that requires the interpreter to be also thinking of the object, and the same conditions on successful thought apply to him, determinate thought about an object requires (at least) two thinkers mutually interpreting each other as thinking about that object. Furthermore, to interpret someone, one has to grant them Self-Knowledge (Davidson 1984). The point is simply that if one does not grant a thinker authority over what they mean, one has no way to even begin interpreting them. Of course, the Principle of Charity also grants the thinker knowledge of her environment, but there is a difference. Where one finds exceptions to perceptual knowledge, that is, where one attributes a perceptual error, one does not need to attribute some different piece of perceptual knowledge, but where one attributes an error about meaning, as in the case of someone who misuses words, one must still allow that they know what they intend to mean, and this is just another piece of Self-Knowledge. So the necessary Asymmetry is that in order to interpret someone, one must always grant them authority over what they mean, even when one is also finding them mistaken on some matter.

On Davidson's account of Self-Knowledge, then, our privileged knowledge of our own minds exists only 'by courtesy of an interpreter' (the phrase comes from Barry Smith). This is only plausible if one has already accepted that one only has thoughts in so far as one is interpretable as having those thoughts. But this claim is in danger of conflicting with the very idea that we have Self-Knowledge, since I know what I think prior to and independently of how I might be interpreted. Of course, Davidson will insist that the interpreter must respect my authority over my own mind, but all this requires is that he *interpret* my Self-Attributions so as to make them correspond with *his* ascriptions of thoughts to me. So while Davidson appears to be defending a strong Self-Knowledge claim, he has lost the idea that Self-Knowledge is a genuine cognitive achievement. In fact, its status as knowledge is not earned by the subject at all but bestowed graciously by the interpreter.

The third line of argument is found in Tyler Burge (1996). Burge does not try to establish that all minded creatures must have Self-Knowledge, but the weaker thesis that all creatures capable of critical reasoning must. A creature which made transitions in thought which conformed to standards of good reasoning, but was not aware of this, would not be a critical reasoner. Often our thought processes are not critical in this sense, but we also have this ability to make inferences while aware of their correctness and because of their correctness, which requires Self-Knowledge:

> To reason critically – to consider reasons bearing on the truth of some matter, to suspend belief or desire, to weigh values under a conception of the good – one must treat one's own commitments as matters to be considered and evaluated. ...
>
> So critical reasoning requires thinking about one's thought. But it further requires that that thinking be normally knowledgeable. ... which [requirement] is shared with the other cognitive faculties, such as perception. (1996, 100)

However, as the last sentence makes clear, this does not establish the Necessity of an Asymmetry, since perceptual knowledge is equally necessary (for creatures which can perceive). Burge takes the argument one step further to show that Self-Knowledge has an epistemic characteristic lacking in other forms of knowledge and that critical reasoning requires it to have this characteristic. The crucial characteristic is not infallibility but immunity to a certain sort of error which Burge calls 'brute error'. Brute errors occur when one makes a mistake, but that mistake is through no fault of one's own, nor is it a result of the malfunctioning of a cognitive faculty. Thus, for example, one might mistake a Chiffchaff for a Willow Warbler, simply because they look very similar. Burge's thesis is that if we are to be critical reasoners, we must have Self-Knowledge and that Self-Knowledge must be immune from brute error. This provides a clearly necessary asymmetry between Self-Knowledge and other knowledge.

The argument is that to be critical reasoners our first-order beliefs and values have to be responsive to second-order considerations. Thus, if I know I believe p and that if not-q then not-p, and I recognize the validity of contraposition and modus ponens, then I have good grounds to believe q. In contrast, if I know Mary believes p and if not-q then not-p, and I recognize the validity of contraposition and modus ponens, then it follows that Mary ought to believe q, but not that I ought to. The point is that my Self-Knowledge does more than merely allow me to *predict* what I will come to believe, rather it gives me a reason to believe those things.

Now it is Burge's main contention that were brute error in Self-Attribution possible, this would not be the case. The argument is that were brute error possible, it would always be possible that a Self-Attribution was mistaken through no fault of one's own, and consequently there could not be 'an immediate rationally necessary' connection between the second-order premises and the first-order belief: knowing that I believe that p and that if not-q then not-p, and recognizing contraposition and modus ponens, would not rationally require me to believe that q. The immediate rational necessity follows from the fact that the first-order belief and the second-order belief are both parts of the same 'point of view', and thus bear rational connections to each other, in the same way that two first-order beliefs, both being parts of the same point of view, are required to be consistent. Whereas, when something is known in such a way that brute error is possible, for example when we look at ourselves as others do in order better to understand our motives, or we read an old diary unsure whether it is our own, there is always the possibility that the judgement and its subject matter are parts of different points of view. When this possibility is in play, the second-order judgements do not have immediate and rationally necessary consequences for the first-order beliefs and evaluations. Any such connection would have to be mediated by an independently justified belief that the points of view are the same.

There are two problems with Burge's account of Self-Knowledge. One is that he gives us no indication of what it is about a Self-Attribution which makes it from the same point of view as its subject matter. Having the form: I now have attitude A towards content p, is a necessary condition, but it is not sufficient because one can make a judgement of that form on the basis of considering the evidence of one's behaviour, as in the case of an interviewer discovering her bias. Burge's claim must be that the special epistemic entitlement attaches to judgements of a certain form *when they are arrived at in a certain way*. When we ask what way judgements must be formed if they are to have the special entitlement, the answer is they must be formed directly and not on the basis of consideration of the evidence. But here it looks like Burge has just shifted the problem. The original puzzle was how judgments about our own minds made directly and without recourse to the evidence could constitute knowledge. Burge's answer to that question is that they are necessarily made from the same point of view as their subject matter, which makes them immune to brute error and thus gives them a special epistemic status. However, we now face the puzzle of why these direct judgements are necessarily from the same point of view as their subject matter. This is a puzzle because, on the one hand, their form is not sufficient, since there can be judgements of the same form which are not so epistemically privileged, and on the other their directness is not sufficient either, since there can be direct judgements of other matters such as elapsed time. Either they have some other feature, or it is the combination of these two features, form and directness, which explains how come the judgement is necessarily from the same point of view as its subject matter. Until we have such an explanation, the account is incomplete.

The second problem with Burge's account is the existence of an embedded assumption. Burge is arguing that second-order judgements can produce immediate first-order rational requirements, that this is so because the judgements are made from the same point of view as their subject matter, and further, that this identity of points of view is not contingent. It is this very last move which is questionable. What is necessary for there to be an immediate rational connection between the second-order judgement and its subject matter, is that we are entitled by default, that is we do not normally need a justification, to accept the identity of the points of view. Burge goes one step further and requires the connection to be necessary. There are plenty of immediate rational connections which are defeasible, such as that between perceptual appearance and judgement, but Burge assumes that in the case of Self-Knowledge, the connection must be indefeasible and thus necessary. One reason for this would be that were the connection defeasible due to the possibility of brute error, this would undermine the rational *requirement* imposed by the second-order judgements upon the first-order beliefs and desires. There is an analogy here with Kant's conception of moral requirements, since Kant thought that the prescriptive nature of morality required there to be a necessary connection between moral judgements and reasons for action. The Humean, in contrast, holds that it is a contingent fact about us that our moral judgements give us reason for action, but given that it is a fact about us, contingent or not, we ought to do what morality prescribes. It would seem that there is the possibility of a parallel Humean move against Burge on Self-Knowledge: it is a contingent fact about us that there is a connection between our Self-Attributions and what we have reason to think at the first-order, but it is a fact about us all the same and thus critical reasoning is possible.

Arguments for 2], the necessity of Self-Knowledge, are of great philosophical suggestiveness but fail to be conclusive. The conditional claim 3] looks more promising, but I know of no direct argument for it. There is, however, an argument for 4], immunity to error, and this entails a slightly weaker version of 3]. Burge argued that Self-Attributions must be immune from brute errors, but faced the problem of having no explanation of how they achieve this. In the next section I shall argue for immunity from all errors, not just brute errors, in a way which explains how this comes about.

6. FALLIBILITY

As I mentioned above, the claim that Self-Attributions can be mistaken is as much a dogma of current philosophy as the opposite was of an earlier age. In fact, even Descartes did not hold the view that each of us should be taken as the unchangeable authority on what we think:

> I thought that in order to discover what opinions they really held I had to attend to what they did rather than what they said. For with our declining standards of behaviour, few people are willing to say everything that they believe; and besides, many people do not know what they believe, since believing something and knowing that one believes it are different acts of thinking, and the one often occurs without the other. *Discourse on Method* iii (AT VI 23)

Let us call the claim that Self-Attributions are always true 'Incorrigibility', following the usage of Williams (1978, Appendix 1) and Dennett (1978, 226), to distinguish it from Infallibility, which is the conjunction of 4] and 5]. Schematically Incorrigibility is the claim that necessarily BAp → Ap (if someone believes that he has a particular attitude towards p then he does), and Infallibility the claim that necessarily BAp ↔ Ap. Descartes' observations may present a problem for Infallibility, a claim which is normally associated with Cartesianism, but not for Incorrigibility.

We can go a long way towards reconciling doubters by noting that five common situations are not in fact counter-examples to Incorrigibility.

(1) Incorrigibility does not preclude ignorance. I simply may not know whether I believe that the trains run on time until someone points out to me that I always get nervous if I do not arrive early at the station. Similar sorts of situations arise for desires, for we often do not know what we want until we realize that we are setting about obtaining it. The commonest case of ignorance is when one has simply not thought about the matter, but it may also arise when one has (see (5) below).

(2) A counterexample in which BBp and B¬p would need to show that the thinker did not also Bp, that is did not have contradictory first-order beliefs. Contradictory beliefs are clearly not impossible, but they only occur in far from ideal cognitive circumstances. The problem with this is that the behaviour of anyone who had a false self-ascriptive belief would have to be sufficiently far from the ideal of reasonable or rational behaviour to make the ascription of contradictory beliefs equally as plausible as the ascription of an error in Self-Attribution.

(3) There are some mental states which are intentional, in that they have an object, but which one *ought* not to self-ascribe directly. Examples often cited are emotions such as love, jealousy and pride. The 'ought' here is epistemic: if one formed the second-order belief that one was in love with a certain person, and one formed that 'directly and without recourse to evidence or inference', one would be epistemically irresponsible. What Incorrigibility is concerned with are the Self-Attributions we would carry on making *directly* in an epistemically perfect world. It is clearly not a counterexample to Incorrigibility if some people attribute physical properties such as posture or location to themselves directly and not on the basis of evidence and thereby make mistakes. Incorrigibility is only intended to apply to a sub-class of our properties, namely the propositional attitudes. The existence of borderline cases should not affect the claim to Incorrigibility so long as there are clear cases on either side of the boundary.

(4) It is sometimes noted in this context that my sincere account of my reasons for doing something may be mistaken. This could mean either of two things, only one of which is incompatible with Incorrigibility. Suppose I claim to have volunteered for something because I wanted the job to be done and I realized that no one else was going to do it. Someone might challenge whether my motives were quite so selfless. The correctness of their challenge does not entail that I did not have the belief and desire cited in my rationalization, merely that, even if I had them, it was not *because of* them that I volunteered. I could have been right about those particular beliefs and desires while being wrong in my further claim that they provided my motive for volunteering. (The locus classicus for the view that one could have more than one sufficient reason for an action only one of which in fact explains the action is Davidson (1963).) The Incorrigibility of Self-Attributions of beliefs and desires does not entail the Incorrigibility of our judgements as to which of our beliefs and desires lead us to perform a given action. Combined with the possibility of ignorance (see (1) above), it would seem likely that we do often mistake our motives.

(5) There is some complication about negative self-ascriptions. An effort at Self-Attribution of an attitude towards the proposition that p can have four results with respect to belief (and equally four with respect to desire):

(i) x believes that he believes that p,
(ii) x believes that he believes that not-p,
(iii) x believes that he is indifferent whether p,
(iv) x does not know whether he believes that p or not.

Indifference is here being used to describe an attitude towards a proposition which is incompatible with both belief and disbelief. Thus I am indifferent to the proposition that there are an even number of books in this room right now. There are some propositions, one of which we have never considered, which we do not believe, nor disbelieve, nor are we indifferent to. We can distinguish (iii) and (iv) by saying that in (iii) x is indifferent to p, but in (iv) he is indifferent to Bp.

The Incorrigibility claim has it that (i) entails that x believes that p, (ii) entails that x believes that not-p, (iii) entails that x does not believe that p nor that not-p, and (iv) has no consequences for what x believes. This fourth situation is very important to the Incorrigibility claim, for while one can choose to say what one likes about one's state of mind, one cannot so choose one's second-order beliefs. Even if, in truth, one had no opinion, one might still voice an opinion, and equally one might decline to tell what one knows about one's beliefs. The false statement that one believed that p would not be a counterexample to Incorrigibility unless we could also show one to *believe* that the statement is true. I suspect that most of the common cases in which we retract what we have said about our beliefs and desires fall into this category and thus do not threaten Incorrigibility. Social pressures, scarcity of time and sheer indolence often lead us to state things we do not in fact believe, about ourselves as much as about the rest of the world (v. Smullyan 1981 for an interesting illustration of this point).

Even taking into account these points, there are still thought to be counterexamples to Incorrigibility. An illustration is given by Hugh Mellor (1977, 91):

> A husband, we suppose, can (subconsciously) believe his wife to be unfaithful, while (consciously) believing that he believes nothing of the sort. We see these two beliefs in different aspects of his behaviour. Typically we see the latter in his sincere rationalizations of those actions that to us reveal the former.

The trouble with such examples, reference to the subconscious aside, is that they only preach to the converted. For the defender of Incorrigibility it is insufficiently clear what the described situation involves. A man who believed himself to *trust* his wife but clearly did not would not be a counterexample (see (3) above). What we really need is a case from real life, if all the relevant belief attributions are to be suitably secure. In the absence of this, let us resort to literature and imagine that Othello makes the apparently sincere assertion that he does not believe Desdemona to be unfaithful, while at the same time displaying behaviour, such as spying on her, which we would generally take as evidence that he believes her to be unfaithful. In such a case there are three options in the explanation of Othello's behaviour. The first is to admit a false self-ascriptive belief, the second to deny that Othello really has the second-order belief, and the third to deny that he has the first-order belief that she is unfaithful. Dialectically, to refute the counterexample we need only to show that the first option is not *obligatory*.

I have emphasized the possibility of the second course under (5) above, so now let us suppose that the context rules it out and pursue the third. Othello behaves in a way which would normally count as evidence that he believes Desdemona to be unfaithful; but this is not a normal situation, for it is a situation in which he believes that he believes that she is faithful. Behaviour that may count as evidence for an ascription in one situation will not be evidence for a similar ascription in countless other situations. For example, that I stepped smartly back onto the kerb does not always prove that I believed the oncoming motorist would not stop at the lights. I may have believed that he would stop, but realized I was setting a dangerous example to a child. Of course, we must find some explanation for Othello's jealous behaviour and it is clear that his own rationalizations will not be our best guide, but why insist that the only explanation is that he believes that Desdemona is unfaithful? To assume that is the only explanation is to oversimplify the philosophical psychology of jealousy. Surely jealousy does not presuppose belief, for might one not be jealous despite knowing that there is no cause for jealousy? Mere recognition that a state of affairs is possible can awake 'the green-eyed monster which doth mock The meat it feeds on'.

Othello's jealousy begins when he realizes how other people perceive his wife. He believes that he believes that Desdemona is faithful, he believes that she is faithful, and yet he still behaves jealously, being overprotective, listening to rumours and trying to find independent evidence of what she has been doing. Such behaviour need not be explained by the *belief* that she is faithless. The jealous behaviour arises from the intersection of a certain pessimistic view of the world, encouraged by Iago, and his intense love for Desdemona. This is as coherent an explanation of the situation as the postulation of some sort of self-deceit.

It is not possible here to respond to every alleged counterexample, and even if one could, this would at best show that the denial of Incorrigibility is unfounded. So we need an argument for Incorrigibility. This begins with a now familiar point about the conceptual resources required for Self-Attribution (Davidson 1987, Burge 1988). One cannot believe that anyone, let alone oneself, believes that p without grasping the proposition that p. Further, if there are any necessary conceptual preconditions for having a propositional attitude towards a content p, such as causal conditions on the grasp of the constituent concepts, these will also be conditions upon believing that one has an attitude towards p. A more interesting question is over the consequences of believing (or desiring) that p. Assuming that the belief (or desire) occurs in the context of other mental states, these will be of two sorts: inferences and actions. Inferences here include any transition from one state to another, thus if I believe that p, and I believe that p entails q, then I should believe that q. And if I desire that p, and believe that p entails q, then I am committed to desiring q (even though I may wish that not-q).

There are two important features to note about such consequences or commitments of a mental state. First, they are normative, they specify what one *ought* to do or think. Thus it is always possible for an individual in a given state not to conform to the requirements on her thought and action, however, some degree of non-conformity is sufficient for not being in the state in question. For example, I may fail to draw some of the consequences from my belief that it is raining, but if I refuse to draw any, or draw only the wrong consequences, then this counts against the hypothesis that I had that belief in the first place. Secondly, talk of the commitments of specific beliefs does not commit one to the thought that these commitments are specifiable in advance, that they are codifiable. However, we may still be able to make certain general claims.

The argument now has two stages. The first is to argue that a Self-Attribution has *at least* all the consequences and commitments of the mental state attributed. The second stage is to argue that having the self-ascriptive belief is sufficient for having the ascribed state of mind. We can call this the Containment Claim, because the idea is that the belief that one, say, believes that p, is a state of mind which includes the state of believing that p. (The argument is sketched but not endorsed in Shoemaker (1996) and endorsed in Stoneham (1998), from which much of this section is taken.)

The intuitive argument for the first stage is that the commitments of the belief that p can be summed up as 'think and act as if p is true'. If one believes p, then one may infer anything that follows from p, and one may not infer anything inconsistent with p, and one may act in any way that would help achieve one's goals were p true, but one may not act in any way which would hinder one's goals were p true. These are the normative commitments of a rational believer that p. Now suppose one believed that one believed that p, surely a rational believer would be equally prohibited from inferring something inconsistent with p, or acting in a way which will only achieve her goals if not-p? Similarly for all the other constraints. The argument can be run, mutatis mutandis, for desires and other propositional attitudes.

From the premise that if one believes that p one ought to act as if p, it follows that if one believes that one believes that p, one ought to act as if one ought to act as if p. But if one ought to act as if one ought to act as if p, does it follow that one ought to act as if p? One can imagine a situation, say a drama school, in which the instruction to act as if one is acting as if p will produce different behaviour from the instruction to act as if p. This is because the first instruction will usually be interpreted as requiring one to act as if one was acting that p less than perfectly. If, however, we gave the instruction to a drama student to act as if they were a perfect actor acting as if p, then the only way they could fulfil this would be to act as if p to the best of their ability. So this instruction would produce the same result as the instruction to act as if one was, to the best of one's ability, acting as if p. If we were to assume that everyone always tries to act to the best of their ability, then the instruction to act as if one was oneself acting as if p, would produce the same result as the instruction to act as if p.

Stage two of the argument, the Containment Claim, can be achieved by a mild 'functionalism' about belief and desire. If a thinker is in a state which meets all the conditions of conceptual grasp and involvement or activation for believing that p, and also has all the consequences and commitments of believing that p, what more could be required for it to be the case that she believes that p? The state of believing that one believes that p is just such a state. So it would seem that the state of believing that one believes that p involves or contains the state of believing that p, and mutatis mutandis for the other attitudes. And if the Containment Claim is true, then so is Incorrigibility, for according to Containment BAp *entails* Ap.

One source of resistance to Containment is the conception of beliefs as parts of people rather than states of people. Thus it is odd to say that having a part of one type is sufficient for having a distinct part of another type, without some auxiliary hypothesis about the whole meeting some specification. However, what constitutes being square also constitutes being rectangular, though something can be rectangular without being square. One way of being rectangular is being square, and we might want to express this by saying that the conditions for being square contain the conditions for being rectangular in the obvious sense that a square is an equilateral, right-angled parallelogram and a rectangle is a right-angled parallelogram. A person's being in one state (believing that they believe that p) might be sufficient for them to be in another state (believing that p), though not *vice versa*, since it is arguable that one can have the belief that snow is white without having the concept of belief, and even, perhaps, without being able to refer to oneself indexically. But according to the Containment Claim, one way of being in the state of believing that snow is white is by believing that one believes that snow is white, because what it takes to have the self-ascriptive belief includes what it takes to have the first-order belief.

7. SELF-KNOWLEDGE

If this argument is correct, it establishes Incorrigibility, but that is just a claim about the truth of Self-Attributions, not their epistemic status. Of course, one might think that if we cannot make mistakes, there is not a lot more to be said on the side of epistemology, but that is not quite right. This can be illustrated with an example. Suppose I somehow and without any good reason, come to form the belief that anyone wearing a hat desires chocolate ice-cream, and also, equally without reason, that I am currently wearing a hat. I then deduce, and believe, that I desire chocolate ice-cream. According to Incorrigibility, this is true. The argument I presented above allows us to see that forming the belief that I desire chocolate ice-cream actually gives me the desire. However, the belief is completely unjustified, therefore not knowledge. It would be a very extreme and unreasonable reliabilism which denied this.

Which is not to say that Incorrigibility establishes nothing about the epistemology of Self-Knowledge. Rather we should distinguish two epistemological tasks, that of accounting for our entitlement to knowledge and that of accounting for the justification of particular things known. A sceptic is someone who challenges our entitlement to knowledge, thus, for example, the standard charge against the representative theory of perception is that if it were true we could never have knowledge of the world, deals with entitlements. Suppose that the theory of perception is such that we are entitled to knowledge of the world, then it is still possible for someone to completely lack perceptual knowledge, perhaps because they are (mistakenly) certain that there is an evil demon deceiving them. So epistemology also needs to specify the conditions in which someone entitled to a certain type of knowledge actually acquires that knowledge. That is the task of a theory of justification. In the case of perception, the theory of justification may be totally negative: given the entitling nature of perception, all one needs to do is to use one's eyes and be free from doubts about their trustworthiness.

It should be noted that this distinction between justifications and entitlements is not the same as that found in Burge (1996). For Burge, justifications and entitlements are different types of warrant which can apply to a particular belief. Thus the thinker who judges 'in accord with norms of reason' is entitled to the belief, but one who can articulate and defend those norms has a justification. Epistemically speaking, the latter is no better off than the former. On my distinction, in contrast, one can have an entitlement but lack knowledge.

Incorrigibility establishes our entitlement to Self-Knowledge. The simplest accompanying theory of justification would simply be that to acquire Self-Knowledge one must make Self-Attributions directly and without recourse to evidence, which would rule out the unjustified inference example given above. But this faces a difficulty with the person who attributes to themselves the beliefs and desires of another. One way this might happen is by an inference such as 'If X thinks that then so do I', which is ruled out as being indirect. However, it may be that our credulous subject is such that his recognition that the other believes such-and-such plays no evidential role in his Self-Attribution. It is simply that whenever X asserts something, it strikes him that he believes it too.

To rule out this sort of case from counting as knowledge, we need a stronger condition. There are two ways to go here, corresponding roughly to epistemological externalism and internalism. The externalist will look at the causal origin of the self-ascriptive belief and require not only that the belief be direct at the personal level, but also that it be caused by whatever sub-personal mechanism is normally involved in our Self-Attributions. The internalist, however, will look for something at the personal level to justify the self-ascriptive belief. Typically internalist accounts of justification appeal to some feature of the aetiology of the belief which increase its chance of being true, but the point of saying Self-Attributions are direct is that, at the personal level, they have no aetiology. We find examples of non-aetiological justifications when we consider logical beliefs and inference rules. Typically an internalist will say that we are justified in inferring q from (p and (p → q)) if we are in fact disposed to make the inference *because it has that form*. Having the form of modus ponens increases the chance of an inference being truth-preserving (to 1), and thus if our disposition to make the inference depends upon its having that form, then the inference is justified. The internalist can make a similar move with respect to Self-Knowledge. Having a content of the form 'I now have attitude A towards content p' raises the chance of a belief being true (to 1, according to Incorrigibility), so if the existence and persistence of the belief depends upon its having that form, it is justified. A mark that the Self-Attribution does have that form would be a certain sort of response to a challenge: suppose I assert that I want a chocolate ice-cream and you challenge the truth of that assertion, if I respond 'Look, I am sure I want one, for after all it is my current attitude we are talking about', then I am emphasizing the form of my belief in its justification.

Alston once held a view of Self-Knowledge similar to this (Alston 1976), but later retracted it. His view was that Self-Attributions are *self-warranting*, but his explanation for this was based on a contingent truth, unlike Incorrigibility. His retraction (Alston 1989, 314-5) was based on the sort of case which above motivated the distinction between justification and entitlement, though he does not make such a distinction. A successful non-aetiological account of Self-Knowledge needs both that the connection between the form of a Self-Attribution and its truth is non-contingent, and a distinction between justification and entitlement.

Peacocke (1996) argues that a causal condition is necessary in many important and common cases of Self-Knowledge. However, his argument is premised upon Self-Attributions in such cases not being 'self-verifying'. The Containment Claim is not quite the same as self-verification, but it has a similar consequence for epistemology, in that it makes aetiology redundant.

Both the externalist and internalist accounts of justification here sketched entail that when we make Self-Attributions we normally do so knowledgeably. But neither, on its own, entails the necessitation of this claim. But they do entail a weaker necessitated conditional:

Necessarily, if a creature has the capacity to make Self-Attributions directly, it has an epistemic entitlement to Self-Knowledge.

In contrast, the philosopher who insists that Asymmetry is contingent must allow for the biologically unlikely possibility of a creature which makes direct Self-Attributions which bear no relation to what it in fact believes and desires. It would then be necessary to show the sceptic that we are not such creatures.

8. ANTI-INDIVIDUALISM AND SELF-KNOWLEDGE

We turn now to an issue about Self-Knowledge which has generated intense interest recently, namely the apparent conflict between accounts of Self-Knowledge which preserve Necessity, like the one just given, and a very plausible thesis in the philosophy of mind called Anti-Individualism.

Anti-Individualism in the philosophy of mind is more often called Externalism though unrelated to externalism in epistemology. It has many forms, but all have in common the denial of local supervenience, which is the thesis that all our mental properties, and our contentful thoughts in particular, supervene upon the individual subject's non-mental constitution. This captures the vaguer intuition that one's thoughts, though often caused by encounters with one's environment, have a nature which is independent of the external world. It follows from local supervenience that if there were two thinkers identical in all respects other than their thoughts, they would necessarily be having the same thoughts. In denying this, the Anti-Individualist is asserting that differences in one's social and physical environment can entail differences in the identity of the thoughts one is having *without affecting one's physical constitution*.

This striking claim is usually motivated by thought-experiments about a different planet known as Twin Earth. Twin Earth is superficially very similar to Earth but contains some subtle differences. The most common example is to do with the chemical constitution of water, but others are more plausible. I shall adapt one from

Hilary Putnam (1975). On Earth there are two metals, aluminium and molybdenum, which are very similar to the uninformed eye and can be put to very similar uses. However, aluminium is more common and it is thus aluminium which is used for everyday domestic purposes such as pots and pans and cooking foil. Let us suppose that on Twin Earth it is molybdenum which is common and aluminium which is sparse, so it is molybdenum which is used for domestic purposes. Let us also suppose that on Twin Earth the 'English' language is slightly different, so that their word 'aluminium' refers to molybdenum and their word 'molybdenum' refers to aluminium. Now imagine that someone on Earth has a thought which would be correctly expressed by the Earth-English sentence 'Aluminium is not a cause of Alzheimer's', and someone on Twin Earth has a thought which would be correctly expressed by the Twin-English sentence 'Aluminium is not a cause of Alzheimer's'. The Anti-Individualist intuition is that, no matter how similar the two thinkers are both physically and in their past experiences, the terrestrial is thinking about aluminium and the alien is thinking about molybdenum. The two thoughts may even differ in their truth-values. The difference in their environments, whether or not it has registered differentially on the thinkers, is sufficient for a difference in the contents of their thoughts.

Whatever the merits of this view, it is often alleged to generate problems for Self-Knowledge. Some people on Earth have sufficient scientific knowledge to know that they are thinking about the metal aluminium and not molybdenum, and would thus be able to tell the difference between Earth and Twin Earth. However, such knowledge is not a prerequisite for thinking aluminium thoughts: someone who knows nothing of molybdenum or the periodic table, who has only encountered aluminium as the stuff of which cooking utensils are made, can still think that Alzheimer's is not caused by aluminium. Does such a person know that they are having an aluminium thought (rather than a molybdenum thought)? As Andrew Woodfield put the point in the Foreword to an early collection of essays on Anti-Individualism:

> A third person might well be in a better position than the subject to know which object the subject is thinking about, hence be better placed in that respect to know which thought it was. (1982, p.viii)

This 'obvious' incompatibility between Anti-Individualism and Self-Knowledge has become the focus of much recent debate. Davidson (1987) and Burge (1988) have both argued that when one ascribes a thought to oneself, one has to think the content of the thought one is ascribing. Consequently Anti-Individualism does not open up any new possibilities for error, for ascribing oneself a thought content one does not have, for one is no more able to self-ascribe the Twin thoughts one lacks than one is able to have them in the first place. It is a mistake to think that Anti-Individualism introduces a range of undetectable (at least without further empirical investigation) errors into our Self-Attributions.

However, more subtle arguments have been wielded in order to show that there is a problem here. There are three types of argument: the first treats Twin Earth as a sceptical hypothesis and claims that in order to know what I am thinking, I must first know that I am not on Twin Earth (Brueckner 1990); the second considers the possibility of being undetectably switched between Earth and Twin Earth

(Boghossian 1989); and the third tries to deduce an absurdity from the conjunction of Anti-Individualism and privileged Self-Knowledge (McKinsey 1991).

The first line of argument fails because of a disanalogy between the sceptical hypothesis that I am being deceived by an evil demon, and the possibility that I am on Twin Earth where my thoughts have different content. The traditional sceptical argument works by holding my beliefs constant and then considering a situation in which they are all false. In contrast, the hypothesis that I am in fact on Twin Earth does not introduce any falsity into my beliefs, nor any illusion at all. It is very tempting to think that Anti-Individualism entails that being on Twin Earth thinking twin-aluminium-thoughts *would seem the same as* being on Earth thinking genuine aluminium-thoughts. But the Anti-Individualist need not accept this, because there may be no way of characterizing what one seems to be thinking, and hence no way of determining whether two such seemings are of the same type, without reference to the content of the thought one seems to be thinking. On Twin Earth I would not be able to think genuine aluminium-thoughts, and consequently it could not seem to me as if I were.

However, even if we grant that there is some sense in which the conscious experiences of those on Earth and their counterparts on Twin Earth are the same, the sceptical conclusion does not follow: Self-Knowledge is not evidentially based upon such conscious experiences.

The second, slow-switching, argument works by showing that the conjunction of Anti-Individualism and Self-Knowledge is incompatible with some favoured epistemic principle. As a dialectical strategy, this only works if the Anti-Individualist cannot reject the epistemic principle in question. For example, Jessica Brown (2000) persuasively argues that the more attractive, global, reliabilism is incompatible with conjunction of a particular form of Anti-Individualism and Self-Knowledge. But if the Anti-Individualist has already noted the dissimilarities between Self-Knowledge and perceptual knowledge, it is far from obvious that he should endorse reliabilism for Self-Knowledge. Specifically, one might think that reliability is a necessary condition of aetiological epistemologies, but not of the non-aetiological account given above. In general, any adequate epistemology will have to treat Self-Knowledge as a special case, so this form of argument is at best ad hominem.

The third form of incompatibilist argument is very interesting. It was first put forward in a short paper by Michael McKinsey (1991). He summarizes his argument thus:

> In effect it says, look, if you could know a priori that you are in a given mental state, and your being in that state conceptually or logically implies the existence of external objects, then you could know a priori that the external world exists. Since you obviously *can't* know a priori that the external world exists, you also can't know a priori that you are in the mental state in question. It's just that simple. (p.16)

Before evaluating this argument, which has spawned a vast and ever-growing literature, there are a few explanatory notes. First, the Anti-Individualism I introduced above by means of the Twin Earth thought experiment does not have the consequence that your being in certain mental states implies the existence of external objects. To get to that conclusion one needs two further moves. The first is that the best explanation of the differences in thought content between the inhabitants of Earth and of Twin Earth is that they have had different causal histories, in particular that terrestrials have interacted with aluminium where their twins have interacted with molybdenum. The second move is that there is no other way to acquire the concept aluminium than by causal interaction with samples of aluminium. Both moves are quite commonly, if not explicitly, made throughout the literature, though many Anti-Individualists follow Tyler Burge in being more cautious on this point.

Secondly, we should note that by talking of knowing a priori that one is in a certain mental state, McKinsey is not assuming that the epistemology of Self-Attribution is the same as for paradigms of a priori knowledge such as logic and arithmetic. Rather the point is simply that our Self-Knowledge is independent of any investigation of the environment, which is why the consequence, that we know the external world exists, is so absurd.

Thirdly, it is not so absurd that we should know something general and unspecific, namely that the external world exists, a priori. That, after all, has been the hope of many philosophies. In particular, it is not unreasonable to think that Anti-Individualism might provide us with an anti-sceptical transcendental argument. The absurdity McKinsey is pointing to is the absurdity of knowing a priori that particular things or kinds, such as aluminium or water, exist in the external world.

McKinsey's argument provides a recipe for creating trouble. It starts with an argument schema:

1] I believe that p
2] If someone believes that p, then he has causally interacted with e
3] So e exists, or has existed (in my past light cone).

The Anti-Individualist holds that there are true instances of 2], and that we can know this by philosophical reflection alone. If we have Self-Knowledge, I can also know 1] without investigating the environment. Finally, I know the inference is valid, so by the closure of knowledge under known entailment, I can know 3] without investigation of my environment. But this is absurd.

This schema should remind us of Descartes' infamous Trademark Argument:

A] I have an idea of God.
B] My idea of God could only have come from God.
C] So God exists.

The standard objection to this argument is that Descartes' reasons for holding B], if they are to be at all plausible, threaten to undermine A]. Critics such as Hobbes also saw that there is a problem with Descartes *knowing* A]. (See Stout 1998 for an illuminating discussion of the form of Descartes' argument.) Similarly, McKinsey is not primarily intending to challenge the *truth* of 1] or 2], but certain clams about our knowledge of those premises.

We are left with only the following seven options:

i. Accept the 'absurd' conclusion that we can know the existence of something without empirical investigation.
ii. Deny the transmission of epistemic status across known entailment.
iii. Deny the inference is valid.
iv. Deny that we can know 1] without empirical investigation.
v. Deny that we have Self-Knowledge.
vi. Deny that we can know 2] without empirical investigation.
vii. Deny Anti-Individualism.

The very fact that there are so many options means that the argument is unlikely to persuade any Anti-Individualists that there is a problem here, though it is readily agreed that i and iii are hopeless (Pace Peacocke 1996, 152). Option ii has been defended by Martin Davies (1998) and option iv by Bill Brewer (1999), but by far the most common response is vi. The thought is simply this: the thought experiment that persuaded us that Anti-Individualism is true appealed to the fact that our concepts *aluminium* and *molybdenum* successfully name instantiated (natural) kinds. This is not something knowable a priori, as such failures as caloric and the ether make clear.

This response to McKinsey can allow that Anti-Individualism is an a priori philosophical thesis while insisting that any particular claim that a given concept could only be possessed by someone who had been in contact with samples of the kind, depends upon empirical knowledge. The Anti-Individualist *might* even concede that there are one or two concepts (such as Descartes' idea of God, perhaps, or the concept of the physical world) for which the Anti-Individualist thesis 2] can be known a priori. In those cases the conclusion is not absurd but a philosophical triumph.

Both McKinsey and Boghossian (1997) deny this move can be made (but see Stoneham 1999 and Tye and McLaughlin 1999), and the debate quickly moves from the original epistemological question of Self-Attributions, to the vexed issue of what it takes to understand or grasp a concept. The proponent of vi must say that one can grasp a concept, and know that one has grasped it, without knowing its semantic type, and thus without knowing whether 2] is true. This consequence is important for the philosophy of mind and the theory of content, but it takes us beyond the remit of this essay.

Tom Stoneham
Department of Philosophy, University of York

REFERENCES

Alston, W.: 1976, 'Self-Warrant: A Neglected Form of Privileged Access', *American Philosophical Quarterly* **13**, 257-72, reprinted in Alston (1989).

Alston, W: 1989, *Epistemic Justification*, Cornell University Press, Ithaca, N.Y.

Armstrong, D: 1963, 'Is Introspective Knowledge Incorrigible?', *The Philosophical Review* **72**, 417-32.

Boghossian, P.: 1989, 'Content and Self-Knowledge', *Philosophical Topics* **17**, 5-26.

Boghossian, P.: 1997, 'What the Externalist Knows A Priori', *Proceedings of the Aristotelian Society* **97**, 161-175.

Brewer, B.: 1999, *Perception and Reason*, Oxford University Press, Oxford.

Brown, J.: 2000, 'Reliabilism, Knowledge, and Mental Content', *Proceedings of the Aristotelian Society* **100**.

Brueckner, A.: 1990, 'Scepticism about Knowledge of Content', *Mind* **99**, 447-451.

Burge, T.: 1988, 'Individualism and Self-Knowledge', *Journal of Philosophy* **85**, 649-663.

Burge, T.: 1996, 'Our Entitlement to Self-Knowledge', *Proceedings of the Aristotelian Society* **96**, 91-116.

Davidson, D.: 1963, 'Actions, Reasons, and Causes', *Journal of Philosophy* **60**, 685-700.

Davidson, D.: 1984, 'First Person Authority', *Dialectica* **38**, 101-112.

Davidson, D.: 1987, 'Knowing One's Own Mind', *Proceedings and Addresses of the American Philosophical Association* **60**, 441-458.

Davidson, D.: 1991, 'Three Varieties of Knowledge', *Philosophy* **30**, 153-166.

Davies, M.: 1998, 'Externalism, Architecturalism, and Epistemic Warrant', in C. Wright, B. Smith and C. Macdonald (eds.), *Knowing Our Own Minds*, Oxford University Press, Oxford.

Dennett, D.: 1978, *Brainstorms*, Bradford Books, Cambridge, Ma.

Descartes, R.: 1637, *Discourse on the Method*, in *The Philosophical Writings of Descartes*, vol.1, trans. Cottingham, Stoothof and Murdoch, Cambridge University Press, Cambridge, 1984.

Descartes, R.: 1641, *Meditations on First Philosophy*, in *The Philosophical Writings of Descartes*, vol.2, trans. Cottingham, Stoothof and Murdoch, Cambridge University Press, Cambridge, 1984.

Gopnik, A.: 1993, 'How we know our minds: The illusion of first-person knowledge of intentionality', *Brain and Behavioural Sciences* **16**, 1-14.

Hacker, P.: 1990, 'Privacy', in *Wittgenstein: Meaning and Mind,* Part 1, Blackwell, Oxford.

Heal, J.: 1994, 'Moore's Paradox; A Wittgensteinian Approach', *Mind* **103**, 5-24.

Lyons, W.: 1986, *The Disappearance of Introspection*, MIT Press, Cambridge, Ma.

McDowell, J.: 1986, 'Singular Thought and the Extent of Inner Space', in J. McDowell and P. Pettit (eds.), *Subject, Thought and Context*, Oxford University Press, Oxford.

Mellor, D. H.: 1977, 'Conscious Belief', *Proceedings of the Aristotelian Society* **77**, 87-101.

Peacocke, C.: 1996, 'Entitlement, Self-Knowledge, and Conceptual Redeployment', *Proceedings of the Aristotelian Society* **96**, 117-158.

Putnam, H.: 1975, 'The Meaning of "Meaning"', in his *Mind, Language and Reality (Philosophical Papers Volume 2)*, Cambridge University Press, Cambridge.

Reid, T.: 1785, *Essays on the Intellectual Powers of Man*, A. Woozley (ed.), Macmillan, London, 1941.

Ryle, G.: 1949, *The Concept of Mind*, Hutchinson, London.

Sellars, W.: 1956, 'Empiricism and the Philosophy of Mind', *Minnesota Studies in the Philosophy of Science* **1**, 253-329.

Shoemaker, S.: 1996, 'Moore's Paradox and Self-Knowledge', in *The First-Person Perspective and Other Essays*, Cambridge University Press, Cambridge.

Smullyan, R.: 1981, 'An Epistemological Nightmare', in D. Hofstadter and D. Dennett (eds.), *The Mind's I*, Basic Books, New York.

Stoneham, T.: 1998, 'On Believing that I am Thinking', *Proceedings of the Aristotelian Society* **98**, 125-144.

Stoneham, T.: 1999, 'Boghossian on Empty Natural Kind Concepts', *Proceedings of the Aristotelian Society* **99**, 119-122.

Stout, R.: 1998, 'Descartes' Hidden Argument for The Existence of God', *British Journal of Philosophy* **6**, 155-168.

Tye, M. and B. McLaughlin: 1999, 'Is Content Externalism Compatible with Privileged Access?', *Philosophical Review* **107** 349-380.

Williams, B.: 1978, *Descartes: The Project of Pure Enquiry*, Penguin, Harmondsworth.

Williamson, T.: 1996, 'Cognitive Homelessness', *Journal of Philosophy* **93**, 554-573.

Wittgenstein, L.: 1958a, *The Blue and the Brown Books*, Blackwell, Oxford.

Wittgenstein, L.: 1958b, *Philosophical Investigations*, Blackwell, Oxford.

Woodfield, A. (ed.): 1982, *Thought and Object*, Oxford University Press, Oxford.

Wright, C.: 1989a, 'Wittgenstein's Later Philosophy of Mind: Sensation, Privacy and Intention', *Journal of Philosophy*, **86**, 622-634.

Wright, C.: 1989b, 'Wittgenstein's Rule-following Considerations and the Central Project of Theoretical Linguistics', in A.George (ed.), *Reflections on Chomsky*, Blackwell, Oxford.

KEITH YANDELL

THE EPISTEMOLOGY OF RELIGIOUS BELIEF

I. INTRODUCTION

Religion

A religion proposes a diagnosis of, and a cure for, what it takes to be the deep and devastating disease that we all share. Religious traditions differ as their diagnoses and cures differ. Thus different accounts of what there is – an omnicompetent God and self-conscious substances made in God's image; qualityless Brahman and nothing else; or co-dependent momentary states – correlate with accounts of what needs to be made right – separation due to sin removed by God's forgiveness in response to repentance and trust; ignorance of one's identity with Brahman cured by knowledge gained in one sort of esoteric experience; knowledge of one's transitory nature gained in a different sort of esoteric experience. The metaphysics of some religious traditions – and while it is sometimes denied, occasionally even by the traditions, that they have any metaphysic, the denial itself is cast in a context of metaphysical claims – differ vastly from the metaphysics of others.

Religious Belief

A religious belief is a belief that some religious claim is true – a belief that God exists, or that there is such a thing as Brahman without qualities, or that what we call persons are collections of transitory states that stand in certain relations to one another, or that our deep and devastating problem is that we have sinned against God, or the like. For the sake of respecting limits of space, our concern here will be with theistic religious belief. Applying what is said to non-theistic religious belief is left as an exercise for the reader; its application should not be difficult.

Two Questions

Two quite distinct questions are easily conflated: what is the evidence in favor of a proposition being true? and when is a person justified in believing that a proposition is true? Let these be the E(vidence)-question and the B(elief)-question, and consider them relevant to some proposition – say, that (G) God exists – and some person – say, fourteen year old Marti. The E-question, relative to (G), asks about something independent of what anyone believes. It is like the question What is the explanation

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 673-706.
© 2004 *Kluwer Academic Publishers.*

of the avalanche?[1] If there is an avalanche, there is some explanation of that fact, whether anyone discovers it or not. Similarly, whether anyone knows it or not, there is a correct answer to the question as to what the evidence is concerning God's existence – in favor, against, tied, vacuously tied, or whatever. The B-question, relative to Marti, depends on what is true about Marti in the context in which Marti believes that (G) is true. The E-question is a purely epistemological question.[2] No element of value theory properly enters into its answer. It is not a question about duties or virtues, epistemic or otherwise. The B-question may or may not be purely epistemological; many answers offered to it certainly are not. One need not answer the B-question in order to answer the corresponding E-question. Whether one requires an answer to the E-question in order to answer the corresponding B-question is controversial. The E-question is distinct from, and independent of, the B-question.

II. DISTINCTIONS

Internalism and Externalism

The notion that knowledge is justified true belief – that person S knows that proposition P is true if and only if P is true, S assents that P is true, and S's assent is knowingly based on something that is in favor of P's truth – is subject to counter-example.[3] It is hard to see how to fine-tune it to produce a counter-example free product. Hence internalism – the element expressed by *S's assent is knowingly based on something that is in favor of P's truth* – has been widely rejected and replaced by externalism. The rough idea of externalism is that S knows that P if and only if P is true, S assents that P is true, and in S's giving assent S's cognitive capacities are working properly. Then an account is given of *S's cognitive capacities are working properly* that does not require that S's assent is knowingly based on something that is in favor of P's truth.

Among externalist accounts, some favor naturalizing epistemology. What this amounts to depends on what it is to naturalize, and this ranges from replacing epistemology by something scientific – say, psychology in the process of being reduced to physiology – at one extreme to simply removing its non-descriptive components on the other.[4] The most complete version of externalism extant is naturalistic in the latter sense. Strictly, it is naturalistic in the sense that it requires no valuative components not already part of the natural sciences. So far as I can see, what such externalism does – and what any defensible externalism will do – is embed certain elements of internalism in a wider and different sort of context than is typical of purely internalist accounts. I will try to make this clear as we progress. Obviously, both internalism and externalism are answers to the B-question.

Epistemology and Ethics

There is a large element of ethics in many answers to the B- question. A typical assumption is that having knowledge is a matter of having true belief under conditions in which one is also intellectually praiseworthy. Thus to know is said to be a matter of believing in a truth with respect to which one has done one's duty – a truth one has accepted as such only upon whatever sort of perceiving, checking, testing, experimenting, or other epistemic probing (if any) one in one's circumstances is obligated to engage in. Or to know is said to be a matter of believing in a truth with respect to which one has manifested the relevant intellectual virtue – a truth one accepts by virtue of one's truth-seeking cognitive capacities having functioned normally in an environment suited to their exercise. There are various duty-oriented and virtue-oriented accounts, and nothing in principle to prevent someone from adopting a duty-cum-virtue version. I propose to make as small an excursis into the ethics of belief as is possible in a discussion of the epistemology of religious belief. Epistemology is not a sub-discipline of ethics any more than ethics is a sub-discipline of epistemology, though each is sometimes relevant to the other.

Purists (Evidence needs no analysis as probabilistic)[5] *and Robust Probabilists (Evidence necessarily is always quantitatively or qualitatively probabilistic).*

At least two different views of evidence play their roles in epistemology and hence in the epistemology of religious belief.[6] One takes the notion of evidence as primitive; propositions of the form *Q is evidence for P* are taken to be well-formed and in no need of analysis – *being evidence for* being viewed as both understood and unanalysable save perhaps in terms of a small number of tightly related epistemic notions any one of which must be given in order for the rest to receive a proper account. If there is such a tight epistemic community of concepts, making P more probable than it would be without Q is not among them. On this account, often at least when it is natural to say *Probably, P is true* and one cannot quantify, this simply means *The evidence favors P.* The other is robustly probabilistic. Propositions of the form *Q is evidence for P* is regarded as meaning, or at least entailing, *Given Q, P is more probable than were Q not given.* Sometimes one can assign a number to the probability of one proposition, given another; sometimes one can only speak of the probability of a proposition being raised or lowered, given another proposition, without being able to say to what or how much. But the notion of evidence is viewed, in either case, as being analysable into probability concepts. Richard Swinburne is a robust probabilist; Basil Mitchell (apparently) holds a the-notion-of-evidence-is-primitive view.[7] No attempt will be made here to adjudicate between these perspectives, or even to contend that there is more involved than a difference of style.[8]

III. VIEWS ON WHICH THERE IS NO EPISTEMOLOGY OF RELIGIOUS BELIEF

Religious Pluralism

If either the sort of Religious Pluralism embraced by John Hick or the sort of Language Gamism propounded by D. Z. Phillips is correct, there is no such thing as the epistemology of religious belief. In *A Christian Theology of Religions*,[9] John Hick explains two ideas that are essential to his version of Religious Pluralism:

> ... the different objects of worship and foci of contemplation are different manifestations of the Real in itself....This X [the Real] is postulated as that which there must be if religious experience, in its diversity of forms, is not purely imaginative projection but is also a response to a transcendent reality.[10] ... when we speak of the Real, this is not to say that the Real is one in distinction from two or three or more. The Real remains beyond the range of our human conceptuality, including the concept of number.

On the one hand, the notion of the Real has no content; on the other the Real is that to which all religious experience is a response. So the Real both is, and is not, the ground of religious experience. It is, because otherwise religious experience is a grand illusion, a response to nothing at all. It is not, because it is ineffable and no such notion as that to which all religious experience is a response can have any application at all to it.[11] In turn, the ineffability of the Real is necessary in order for none of the religious descriptions of the Real offered by various religious traditions to be more accurate than any other; the view is that it is better that all be utterly mistaken than that any one be more accurate than another. On this view, there is no epistemology of religion to be done. Given that it is logically impossible that there be any such thing as Religious Pluralism's "the Real," and that positing such a thing is essential to Religious Pluralism, this consequence need not, on those grounds anyway, be accepted, and we can continue.

Phillipsian Language Games

D. Z. Phillips holds a view of religion that, if true, would also render the epistemology of religious belief inoperative. He embraces a view of philosophy that is supposed to be merely descriptive. Like any philosophical position that eschews argument and claims merely to look, see what is there, and report it, what we are offered is a view fully laden with philosophical theses. The view in question thinks of religion as not making any claims about what human-mind-independently exists. Theism, on his account, does not hold that before the mountains existed or there was any earth, God existed.[12] He tells us:

> One will never understand what is meant by belief in God if one thinks of God as a being who may or may not exist ... A God who is an existent among existents is not the God of religious belief[13] ... Talk of God's existence or reality cannot be considered as talk about the existence of an object.[14] To ask whether God exists or not is not to ask a theoretical question. If it is to mean anything at all, it is to wonder about

> praising and praying; it is to wonder whether there is anything in all that. ... "There is a God," though it seems to be in the indicative mood, is an expression of faith.[15]

It is easy to underread these comments and thus miss how radical they are. What we are being told is that Christians – at least those who are not superstitious and understand their own religion – do not suppose that, independent of any human person's way of thinking or speaking or living, there is a God. "Object" in Phillips' comments does not mean "physical object;"[16] it means "something that exists." Of course monotheists do not think of God as "an existent among existents" – as being just another part of the world's furniture. They think of God as Creator and Providence, so existing that the world depends for its existence on God who does not depend for existence on the world.[17] Phillips' claim is that they do not think of God as a being that exists in any way at all. Being a Phillipsian expression of faith is incompatible with being an expression of an existential belief, save in the most perverted of senses. One is invited to believe that: "Discovering that there is a God ... is discovering that there is a universe of discourse we had been unaware of."[18]

Coming to believe that God exists is supposed to be identical to coming to be aware that there is theistic discourse which does not assert that there is a God. Similarly, concerning belief that believers who have died shall be resurrected, Phillips says:

> Such a picture may itself be an expression of the belief that people should act towards each other, not according to the status and prestige that people have acquired or failed to acquire, during the course of their lives, but as children of God, in the equality which death will reveal.[19]

His view is that not only may this belief be so taken, but that – save by ignorant Christians – it is so taken. To believe that (i) *Persons survive death* is supposed to be entirely a matter of believing that (ii) *Persons are all equal in that death is the end for everyone and so differences in social status and material acquisition are not important*, just as to believe (iii) *There is a Creator and Providence* is to believe (iv) *There is a way of speaking that talks about God which expresses values one can live by but does not say there is God.* According to Phillips, to reject such reductionistic accounts is to be superstitious.

It is one thing to assert such things, and another to provide any reason at all to think them true. Phillips does not appeal to any explicit theory of meaning as a basis for his claims.[20] He instead describes Christianity as a religious language game and form of life, terms of art that invite one to take the apparently referential portions of religious language and absorb them without remainder into talk about feelings and behavior. There is also the suggestion that his view is correct because Christians don't typically abandon their beliefs upon receiving philosophical critique or require arguments for the truth of their beliefs. So Phillips' sort of view is alleged to be what one comes to if one pays attention to actual religious practice – if one looks and sees rather than offering arguments.

What should one make of this account? The empirical evidence against it is massive. It would not be surprising if the only people who believed it were Oxbridgeans who have lost their faith. There is an obvious concern in the Old Testament, by both prophets and psalmists, with the existence of evil and their efforts to explain evil without giving up the existence or goodness of God. That believers often retain their faith in the presence of what they admit is evidence against it is evidence of their trust in God; whether they are rational in retaining faith or not, they often recognize the relevance and force of the objections. Those who leave the faith because of argument and evidence no more inherently misunderstand what they leave than do those who come to faith because of argument and evidence misunderstand what they come to. A perspective that pervades both Old and New Testaments is expressed by the author of the New Testament Epistle to the Hebrews who says that one would come to faith must believe that God exists and rewards those who seek. The sober fact is that there are libraries of documents, ranging from theological treatises to letters to mission records that testify against Phillips' claims. The number of Christians who have thought of their Christianity in Phillipsian terms must be something like the current number of albino elephants in captivity. As an empirical description, Phillips' account is so plainly preposterous that it can hardly be intended to really be a result of just looking and seeing. It is plainly an artifact of a Phillipsian conceptual lens. What should be said about it as such? I think this: Phillips has described, and perhaps embraced, a language game and form of life that outsiders would describe as an empty shell of Christianity – one that lacks most of its cognitive content that in turn gives rationale for its values and point to its practices. It appears to be a recipe for lapsed Christians who want to retain something of their religious past. Outsiders, of course, are wrong about this. Whereas Phillips seems to be talking about Christians and Christianity, what his terms refer to is his own linguistic and non-linguistic behavior. So while he seems to be saying of Christians that when they speak of God they merely refer to their own linguistic and non-linguistic behavior, what he is actually up to is referring to his own language game and form of life, saying of it that while he cannot accept Christian doctrine he thinks people can still find (and perhaps that he finds) some comfort in saying certain things and meaning things by them that no one not initiated into his form of life would think they mean. Look and see how he is unmoved by appeals to empirical evidence, continuing to embrace things that, taken literally, are plainly false; further evidence is provided by his continuing to say what he says in the face of powerful philosophical criticisms. The criticisms don't apply because he isn't saying what he seems to be saying. When he says that (v) *One will never understand what is meant by belief in God if one thinks of God as a being who may or may not exist* he means (vi) *My language game is unconcerned with theism*; when he tells us that (vii) *"There is a God," though it seems to be in the indicative mood, is an expression of faith* he means only that (viii) *Claims that God exists, while central to Christianity and other monotheisms, are dismissed – as are their denials – from my particular private form of life.* Thus are assertions that apparently are about Christianity absorbed into a sort of conceptual autobiography. Phillips hence offers a sort of second-order esotericism that has no role for the epistemology of religion, just as his first-order esotericism has no role in the epistemology of religion. But that of course is nothing against the epistemology of religion, and no reason whatever why it is a mistake to pursue its various facets.

IV. TWO EXAMPLES OF ATHEISTIC EPISTEMOLOGY OF RELIGION

The terms *atheistic philosophy of religion* and *theistic philosophy of religion* here mean simply this: strategies of argument, neutral in themselves, that have been used in the development of arguments on behalf of, respectively, atheism and theism. With that in mind, we turn to two examples of the former.

The Presumption of Atheism

Anthony Flew contends for what he calls The Presumption of Atheism. He uses "atheism" a bit technically, proposing to understand it as analogous to "amoral," neither moral nor immoral. So construed, an atheist neither asserts nor denies that God exists. But he also uses it in a more inclusive sense in which the atheist either is agnostic – suspending judgment regarding whether or not God exists – or atheist in the usual sense (affirming that there is no God). Here are the relevant passages:[21]

> What I want to examine is the contention that the debate about the existence of God should properly begin from the presumption of atheism, that the onus of proof must lie upon the theist. The word "atheism," however, has in this contention to be construed unusually. Whereas nowadays the usual meaning of "atheist" in English is "someone who asserts there is no such being as God." I want the word to be understood not positively but negatively. I want the original Greek preface "a" to be read in the same way in "atheist" as it is customarily read in such other Greco-English words as "amoral," "atypical," and "asymmetrical." In this interpretation an atheist becomes: not someone who positively asserts the non-existence of God; but someone who is simply not a theist.

> What the protagonist of my presumption of atheism wants to show is that the debate about the existence of God ought to be conducted in a particular way, and that the issue should be seen in a certain perspective. His thesis about the onus of proof involves that it is up to the theist: first to introduce and to defend his proposed concept of God; and second, to provide sufficient reason for believing that this concept of his does in fact have an application.

On the former meaning of "atheism," the presumption is:

(PA1) In the absence of evidence either way, the rational thing to do is to suspend judgment regarding whether or not God exists.[22]

On the latter meaning, we get instead:

(PA2) In the absence of evidence either way, the rational thing to do is either to suspend judgment regarding whether God exists or to deny that God exists.

Obviously (PA1) and (PA2) are different rules; only (PA2) expresses what one would expect of something called the Presumption of Atheism. The idea behind (PA2) is this:

(PA*) For any positive existential proposition P, if one lacks evidence that P is true, then the rational thing to do is to either suspend judgment regarding

P or to deny that P is true (and not rational to believe that P).

It leaves us asking why a negative vote is proper while a positive vote is not; why isn't suspense of judgment the only sensible alternative? There are cases in which, were there an X, we'd obviously have evidence to that effect. If one takes one's Golden Retriever to a veterinarian's small waiting room and sees no elephants there, one's choice of reasonable beliefs concerning elephants includes the belief that there aren't any in the room; belief that there are, or suspense of judgment, are unreasonable in the context. Some philosophers have held that (S) *Necessarily, if God exists, there is evidence that God exists*, and appeal to this claim can back up (PA2). But (S) itself is controversial, and some philosophers also hold that there is evidence for God's existence.[23] In its Flewian uses, I take it, (PA2) is supposed to be plainly true whatever is the case concerning (S) or pro-theistic evidence, and not restricted to cases where one would have positive evidence were the relevant existential claim true. But then there seems to be nothing like a satisfactory answer as to why suspense of judgment is not the only rational option.[24] The idea is that either suspense of judgment or denial is rational, neither more so than the other, while belief is irrational. So understood, (PA*) entails:

(PA**) For any positive existential proposition P, if one lacks evidence that P is true, then it is rational to deny that P (and not rational to believe that P).

Plantinga suggests this counter-example:[25]

(1) There exists a human being not created by God.

Given that *Necessarily, if God exists, then for every human being H, God created H*[26], (1) entails:

(2) God does not exist.

Obviously, (1) entails:

(1a) There exists a human being.

and so (1) is a positive existential proposition. So, by (PA2), it is rational to believe the denial of (1). The denial of (1) is:

(not-1) It is false that there exists a human being not created by God.

This, together with (1a), entails:

(3) There exists a human being created by God.

If it is rational to believe the denial of (1), presumably it is rational to believe what the denial of (1) plus a plainly true proposition entails. The denial of (1) plus the plainly true (1a) entails that God exists. So if (PA2) is true, it is reasonable to believe that God exists.[27]

The Argument from Bambi

William Rowe[28] invites us to consider a hypothetical deer that, unknown to anyone, excruciatingly dies in a forest fire. Since deer lack souls, and no human person knows about this deer's demise, all of the considerations that theists appeal to in order to show that such a death has some point or serves some purpose, fail to apply. Granted, a creative theist can think of possibilities not ruled out by Rowe's conditions – perhaps some demon's exercise of freedom is the cause of the forest fire and it would not be better to remove the demon's exercise of freedom by making this choice impossible or inefficacious; perhaps some angel, tempted to fall, sees the fire and is reminded that the consequences of yielding are a price not worth paying. Perhaps demons or angels – these ones anyway – cannot be deceived by fake fires or have little by way of imagination, so imaginary or illusory fires would not fill the bill. But such proposals are relevant to the question as to whether there is a flat logical inconsistency between *God exists* and *Bambi dies in the fire*, and Rowe grants that there is not. Rowe presses the evidential, not the logical, problem of evil. His claim is that (i) we have good reason to think that there are real cases similar in all relevant ways to the Bambi case, (ii) in the absence of any reason to think that Bambi's demise serves some purpose, it is unreasonable not to think it serves none; (iii) if it is unreasonable to think that Bambi's demise serves some purpose, it is reasonable to think that God does not exist. Rowe takes (i-iii) to be true, and adds (iv) the real cases that are similar to the Bambi example are cases where we have no reason to think that any purpose is served. Hence there are real cases regarding which it is unreasonable to think that the evil they embody serves any purpose or has any point. Hence it is reasonable to think that there is no God.

There are various controversial propositions in the conceptual region occupied by Rowe's powerful argument. Among them are *Necessarily, if God exists, then for any evil that God allows to exist, God will have a sufficient reason for allowing E and There are evils that we have good reason to believe have occurred and that we have no good reason to think serve any purpose.* The former proposition takes it to be logically impossible that God allow gratuitous evil, and assessing it would require considerable discussion of just what it is for an evil to serve a purpose or have a point, and how this is related to God's having a purpose in allowing it, and whether gratuitous evil really is impossible in a world made by God. Here, I leave these aside; suppose, for the sake of the argument, Rowe is right about his first claim. Consider, however, the second claim: *There are evils that we have good reason to believe have occurred and that we have no good reason to think serve any purpose.* Suppose it is true. Is it proper to infer the required Rowean conclusion that *It is reasonable to think, and unreasonable to deny, that there are evils that serve no* *purpose or have no point?* Drawing this conclusion is legitimate only if something like (R) is true. Consider this pattern of inference:

> (CR) If we know propositions of the forms: 1. For all we can tell, it is true that P, and 2. We can find no evidence in favor of not-P, then it is proper to infer to a proposition of the form: 3. It is reasonable to believe that P, and unreasonable not to do so.

Call reasoning along these lines Common Reasoning. Then consider:

> (R) [For Roweanism] If a case in which one applies reasoning of the sort 1-3 exhibits is one dealing with natural objects, artifacts, the means and ends of human persons, or the means and ends of a person whose cognitive capacities, moral goodness, and causal powers vastly exceed ours, then the result of applying it is reasonably believed to be reliable.

Rowe takes (R) to be true. Contrast (R) with:

> (R*) If a case in which one applies reasoning of the sort 1-3 exhibits is one dealing with natural objects, artifacts, or other human persons, then applying it is reasonably believed to be reliable; if a case in which one applies reasoning of the sort 1-3 exhibits is one dealing with a person whose cognitive capacities, moral goodness, and causal powers vastly exceed ours, it is reasonable to think that such reasoning is as likely to be unreliable as it is to be reliable.

There is, so far as I can see, nothing in Rowe's highly interesting papers that gives us better reason to accept (R) than to embrace (R*). Yet, again so far as I can see, it is (R*) rather than (R) that is true or the more plausible approximation to the truth.[29] The issue is not presence or absence of appeal to sheer mystery. The issue is what may be properly inferred from the information we have – what principle of inference it is justified to accept. Since he has not established (R), which his argument requires, that argument fails.

There is another relevant point. It is compatible with Common Reasoning as expressed in (CR) that we have no evidence whatever for P. It is also false regarding the Bambi case that if the evil does serve some point, we would know this or know what the point is. Hence it is proper to hope that what actually represents common reasoning is not (CR) but:

> (CR*) If we have no reason for accepting P, and it is false that were there reason for thinking not-P to be true we'd be aware of that reason, then even if for all we can tell P is true and we can find no reason in favor of not-P, then -- since for all we can tell not-P is true, it is reasonable simply to suspend judgment regarding P.

Note that (CR*) is of no aid to Rowe's argument. Rowe takes the Bambi case to be analogous to one in which we both see a friend who once had a mustache and who sits fully visible at the bar. We can see that he is clean-shaven, and – having no reason to think he wears a mustache-hiding mask, has died his mustache flesh-colored, or has visited a wizard who has rendered it invisible – we should conclude he no longer has the mustache. I take the Bambi case to be analogous to a case in which our friend sits at the bar with a paper bag over his head. From what we can see, we are neither justified in concluding he still has his mustache, or that he does not.

V. THREE EXAMPLES OF THEISTIC EPISTEMOLOGY OF RELIGION

Probabilistic Epistemology of Religion

In *The Existence of God* Richard Swinburne (1979) appeals to the explanatory power of theism – its truth explains there being something rather than nothing, the order of nature, the accessibility of that order to us, the objectivity of morality, and the occurrence of numinous religious experiences. The distinctiveness comes from his enthusiastic probabilism.[30] He takes theism to have an intrinsic probability – probability given only "tautological knowledge" or (better, since not all necessary truths are tautologies) necessary truths. Prior probability "depends on fit with empirical background knowledge, and scope"[31].

Prior probability is distinct in concept from intrinsic probability, but sometimes identical with it in extension, as Swinburne takes it to be in the case of theism[32] since no empirical background knowledge arises here (the data of such knowledge depending on its being created by God).[33] His view is that theism has great simplicity and hence, as it were, starts the race for truth with an advantage over its main competitor, namely the view that the existence of the world is a brute fact. The explanatory power of theism raises its probability. The occurrence of numinous religious experience then raises its probability to more than .5 Much of Swinburne's argument could be stated without Bayesian or other probabilism.

Swinburne takes it that (i) every proposition has an intrinsic probability; (ii) for any two propositions P and Q, P has some probability on (given) Q and Q has some probability on (given) P; (iii) these probabilities are objective, mind-independent features of propositions. While it is a necessary truth that given a fair throw of a fair die in a fair environment the chances of a one are one in six, and plainly true that given *Every one of a million cars are green, save one* the probability of *The car parked outside is green* is higher than it is given *Of the million cars there are, one is green*, it isn't clear that *My left knee is sore now* has some specific objective probability given *The oldest collie in Switzerland is drinking water now*. Further, the idea that logically contingent propositions have any intrinsic probability is controversial. Why should one think of *Montana contains more goats than Massachusetts* as possessing some degree or other of probability, given only necessary truths? Plantinga claims to show that the idea is flatly contradictory.

Here is his argument:

> ... there are many large classes of propositions such that there seems to be no way in which intrinsic possibility can be distributed over their members in a way that accords both with the calculus of probability and with intuition. Consider, for example, a countably infinite set S of propositions that are mutually exclusive in pairs and such that necessarily, exactly one of them is true: S might be, for example, the set of propositions such that for each natural number n (including 0), S contains the proposition "There exist exactly n flying donkeys." Given nothing but necessary truths, one number should be as probable as another to be the number of flying donkeys. But the members of this countably infinite set can have the same probability only if each has the probability 0. That means, however, that the proposition "That there are no flying donkeys" has intrinsic probability 0; hence its denial – "There are some flying donkeys" – has an intrinsic probability of 1 ... The only way to avoid this unsavory result is to suppose that intrinsic probabilities are distributed in accordance with some series that converges to 0: for example, "There are no flying donkeys" has an intrinsic probability of 1/2, "There is just one flying donkey" gets 1/4, and so on. But then we are committed to the idea that some numbers are vastly more likely (conditional on necessary truths alone) to be the number of flying donkeys.[34]

Plantinga's quoted argument is an enthymeme. It requires some such additional premises as these: (A1) In distributively assigning probability to S, give the probability of 0 to each of its members. (A2) If the probability of a proposition P is 0 and P is assigned that probability due to its being one among an infinite number of contraries, then it is logically impossible that P be true.

In his overall argument he considers another possibility to which a Swinburnean might appeal: (A1*) In distributively assigning probability to S, give the probability of .5 to its simplest member and for every succeeding member give a probability of half of the probability of the preceding member.

There is a neglected strategy that one might adopt:

(A1**) In distributively assigning probability to S, give to each proposition an infinitesimal probability (a probability smaller than any finite number but greater than 0).

or even:

(A1**a) In distributively assigning probability to S, give the probability of .5 to its simplest member and give to each other member an infinitesimal probability.

On (A1), each proposition in S has 0 probability.

On (A1*) the proposition of the form *There are no x* receives a .5 probability, the proposition of the form *There is one x* receives a probability of .25, the proposition *There are two x* receives a probability of .125, and so on. On (A1**), each of these propositions, and each sibling proposition, receives an infinitesimal probability. On (A1**a), each proposition in S that entails there being one or more x receives an infinitesimal probability but the proposition that entails that there are no x receives a probability of .5.

If Plantinga is right that contingent propositions have no intrinsic probability – and this seems to me the truth of the matter – then none of these strategies is the correct way to assign intrinsic probabilities to propositions. There is no more a correct way to do that than there is to say how many square roots there are in the largest apple pie ever baked in Alaska. But the critique is supposed to be, so to say, Swinburne-internal, appealing on to propositions that are part of his theory. As such, it can be rebutted by way of denying (A2) and replacing it by:

(A2*) Only if P is correctly assigned the probability 0 as a member of a infinite set of contraries is it logically impossible that P be true and that procedure is incorrect.

What (A2*) does is to deny that correct distributive probability assignments of 0 to the members of an infinite set of contraries renders that set a collection of impossibilities. Including (A2*) rather than (A2) among Swinburnean claims will enable one to elude the argument intended to show that on his view plainly contingent claims would be necessarily false. Further, (A2*) seems the right choice among (A2) and (A2*). Further, Swinburne could assign the relevant probabilities in accord with an (A1**) or an (A1**a) recipe (presumably preferably the latter) and escape the objection that it is twice as likely that there be one flying donkey as to be two. I take it, then, that there are alternatives fully available to Swinburne that will escape the intended internal critique offered by Plantinga. This, of course, is no defense of the idea that such propositions as *There is one cow* has any intrinsic probability.

Another crucial element in Swinburne's argument is his Principle of Credulity, roughly stated as follows:

> I suggest that it is a principle of rationality that (in the absence of special consideration) if it seems (epistemically) to a subject that x is present, then probably x is present; what one seems to perceive is probably so.[35]

Here, too, Plantinga is unpersuaded. Considering the proposition *It seems to Sam that Zeus is present* he writes:

> Wouldn't it be just as likely that Sam was mistaken, the victim of a Cartesian demon, or an Alpha Centaurian scientist, or any number of things we can't even think of? What would be a reason for thinking this? That in most possible worlds, most pairs of such propositions are such that the second member is true if the first is? But is there any reason to think that?[36]

His claim is that:

(PC) For any proposition of the form *X is present to S*, that proposition has a probability of greater than .5 on a proposition of the form *X (epistemically) seems to S to be present.*[37]

has this feature: there is at best no reason to think it true. The argument suggested seems to be something like the following. One begins with a claim along the lines of:

(PC*) If there is an indefinitely large set Q of propositions such that if any member of Q is true, then even if a proposition of the form *X (epistemically) seems to S to be present* is true, it is not the case that the

> corresponding proposition of the form *X is present to S* is true, then even if S has no reason to think that how things seem to S are not as they are, the truth of the proposition of the form *X (epistemically) seems to S to be present* does not raise the probability of the truth of the corresponding proposition of the form *X is present to S* to greater than .5.[38]

One continues with a claim of this sort:

> (PC1) For any propositions of the form *X (epistemically) seems to S to be present* and *X is present to S*, there is an indefinitely large set Q of propositions such that if any member of Q is true, then even if a proposition of the form *X (epistemically) seems to S to be present* is true, it is not the case that the corresponding proposition of the form *X is present to S* is true.

It concludes that:

> (PC2) Even if S has no reason to think that how things seem to S are not as they are, the truth of the proposition of the form *X (epistemically) seems to S to be present* does not raise the probability of the truth of the corresponding proposition of the form *X is present to S* to greater than .5.

Perhaps we should also think of Plantinga's critique as including the claim:

> (PC**)If there is an indefinitely large set Q of propositions such that if any member of Q is true, then even if a proposition of the form *X (epistemically) seems to S to be present* is true, it is not the case that the corresponding proposition of the form *X is present to S* is true, then even if S has no reason to think that how things seem to S are not as they are, the truth of the proposition of the form *X (epistemically) seems to S to be present* does not raise the probability of the truth of the corresponding proposition of the form *X is present to S* to any degree at all.

and proceeding analogously to the previous argument. Plantinga's claim is that only if such propositions as (PC*) and (PC**) are themselves embedded in a theistic context is there any reason to think them true, and that of course would rather ruin them as premises in any argument for theism from descriptions of experience plus claims about the probability of experiences being reliable.

Swinburne's typical argument for claims such as (PC) is that if they are not true, the alternative is scepticism, and that scepticism is false. Leaving this last claim aside, it seems false that the alternative to (PC) is scepticism regarding the reliability of our experiences.[39] Plantinga suggests his own Warrant Epistemology of Religion. It seems right that Swinburne's Principle of Credulity is not a necessary truth, and once one sees that it is a logically contingent proposition it is hard to see why one should think it true.

One can abandon robust probabilism[40] and offer this claim:

(PE) For any person S, experience E, and proposition P, if E is a matter of S's phenomenologically seeming to encounter an X, and P is the proposition *There is an X*, then E is evidence that P is true.[41]

Given (PE) – more accurately, given its more carefully formulated successor – the occurrence of such experiences as at least seeming to see a pig will be (defeasible) evidence that there is a pig and at least seeming to experience God will be (defeasible) evidence that there is a God.[42] If some version of (PE) is true, it is a necessary truth, and – unlike (PC) – it seems to me true in all possible worlds. In any case, there are in fact alternatives to (PC).

I take there to be a good deal of force in Swinburne's arguments; I also take that force to be independent of the robust probabilism in which his case is cast. The analogous claim seems to me right about William Alston's arguments and the doxastic practice format within which they are made.

Properly Basic Epistemology

In a few papers[43], Alvin Plantinga drew a map and offered a refutation. His map was simple. There is Ancient-and-Medieval Foundationalism: for any person S and proposition P, S properly believes that P without evidence if and only if P is self-evident to S or is evident to S's senses. There is Modern Foundationalism: for any person S and proposition P, S properly believes that P without evidence if and only if P is self-evident to S or P is incorrigible to S[44]. Both the older and the newer foundationalist is an evidentialist: for any person S and proposition P, if S believes that P and P is not foundational for S, then if S properly believes that P it is the case that S has evidence for P and that evidence is some proposition P* such that P* is foundational for S and it is the case that the probability of P on P* is greater than .5. Traditional epistemology has been foundationalist in one of two ways, and evidentialist in the same way. The challenge to the foundationalist is to provide a justification for itself that satisfies its own criterion. Plantinga's claim is that any attempt to answer his challenge will fail because there are no propositions that meet the challenge and are self-evident, evident to the senses, incorrigible, or made more probable than not by foundational propositions. This absence of relevant propositions is of course not a matter of chance or due to lack of ingenuity; the idea is that it is logically impossible that there be any such propositions. Thus foundationalism suffers the inelegance of self-referential incoherence – no foundationalist can, on her own terms, offer a justification of being a foundationalist. While this does not show that foundationalism is false, it does show that no one can offer a justification for foundationalism that meets its own standards – no one properly accepts it. Further, the view that one cannot justifiably believe anything without having propositional evidence for it faces the problem that one either would have to regress infinitely through layers of propositional evidence – an impossibility – or allow circular inference to be justifying – a clear mistake. So some beliefs must be properly basic – held without propositional evidence on their behalf – and among these will be propositions that do not satisfy foundationalist standards.[45]

Plantinga's Basic Criticism of Foundationalism

Modern[46] Foundationalism, Plantinga claims, is self-referentially incoherent, which is to say that it proclaims a standard for knowledge that it does not itself meet. That is, it tells us that:

(MF) For any person S and proposition P, S knows that P if and only if P is a necessary truth or P is incorrigible to S.

But consider the claim that:

(SMF) Person S knows that (MF).

Since (MF) is not a necessary truth or incorrigible to S, S cannot know (MF) if (MF) is true; if (MF) is false, she cannot know (MF). But (MF) is either true or false, and so S cannot know (MF). This does not show that (MF) is false, but that any Modern Foundationalist cannot know it to be true. It is something of an inelegance in a position concerning what knowledge is that its proponents cannot know it to be true.

Descartes is a Modern Foundationalist if anyone is; he has pride of precedence, and arguably of influence, over Locke, at least so far as philosophers go. He held:

(CK) For any person S and proposition P, S knows that P if and only if it is logically impossible that S believes that P be true and *P is true* be false.

This allows knowledge of necessary truths; they can't be false at all, and so can't be false if someone believes them to be true. It allows knowledge of belief-entailed propositions:

(BE) For any person S and time T, there is a set of logically contingent belief-entailed propositions – a set K of propositions such that P is a member of K if and only if S believes that P entails P is true.

For Kim at time T, the set K/Kim will include such propositions as I exist at T, I am conscious at T[47], I am not a prime number at T and the like. For Karen at time T, a corresponding set of propositions will be true, namely those saying of Karen what the members of Kim's set say of Kim.

Descartes also holds:

(CK*) For any person S and proposition P, if it is logically impossible that S believe that P at time T and P be false, then P is either a necessary truth or a belief-entailed proposition relative to S and tensed to T.

Now (CK) and (CK*) entail (MF). Each of (CK) and (CK*) is, if true, then necessarily true. Hence (MF) is, if true, then necessarily true. If (MF) is necessarily true, there is no reason why a Modern Foundationalist cannot, on her own terms, know it to be true. So Modern Foundationalism is not, pace Plantinga, self-referentially incoherent.

Descartes' *Meditations* in effect run as an experiment to see whether or not there is some way to begin with only necessary and belief-entailed propositions and move on with security to propositions that entail the existence of God and of physical objects.[48] As it turns out, the program founders at least regarding propositions expressing mind-body relations and competing scientific hypotheses that seem intractable save though appeal to crucial experiment. The program is generally deemed a failure even regarding God and physical objects. Suppose all this is so. What follows is that either we don't know that God exists and that there are physical objects or that we do and Descartes' definition of knowledge is faulty. Then of course no one can know it is correct.

Descartes beat the sceptic at the sceptic's own game. But he couldn't, on the terms he used to do so, know that there were actual sceptics rather than simply sceptical positions he might himself take. He had secured knowledge of necessary truths and himself, but not knowledge of his physical environment and his fellow humans. That, not self-referential incoherence, is his problem.[49]

Properly Basic Theism

Plantinga suggests that:

> Perhaps the theist is entirely within his epistemic rights in starting from belief in God, taking that proposition to be one of the ones probability with respect to which determines the rational propriety of other beliefs he holds.[50]

He takes this suggestion to be exactly right, in contrast to the famous claim of W. K. Clifford that:

> To sum up: it is wrong always, everywhere, and for anyone to believe anything upon insufficient evidence.[51]

Clifford's maxim would apparently require that extraordinary effort be spent in avoidance of inadequately based belief, and one might wonder whether such avoidance has the stellar and always preeminent status that Clifford affords it. Suppose, for example, that Susan finds that every other Saturday at noon some thought about Australians pops into her head and she believes it to be true – that Einstein loved Australians more than Europeans, that Australians love kangaroos more than New Zealanders, and the like. She believes these propositions to be true without having any evidence and she never checks up as to what can be said for or against them. On Clifford's account, this is wrong; Susan should check out her Saturday noon beliefs and try to rid herself of those that do not pass muster. It isn't that consequence of Cliffordianism that Plantinga protests against.

What Plantinga claims is that it is not always obligatory that one have any (propositional) evidence at all for at least some of what one believes – that some beliefs are properly basic in the sense of their being believed without (propositional) evidence without one violating any obligations at all in so doing. He takes belief that God exists to typically have this status. Plantinga's basic contention, then, in the interests of which the map is drawn and the refutation proffered, is that belief in God is properly basic. Strictly, Plantinga contends that if a Christian believes that God has forgiven him for some sin, guided him to take this job rather than that, or convicted him (made them feel guilty) about some good deed left undone or some bad deed done, he believes with full propriety even if there is no other proposition from which he infers those beliefs or offers as evidence for them. Call these everyday Christian in-house beliefs. Plantinga's example of epistemic propriety in accepting such beliefs is a fourteen year old who accepts such in-house propositions. God forgives, guides, or convicts only if God exists. Belief that God exists is an obvious entailment of the sorts of beliefs noted above, and is held to be properly basic only by virtue of these other beliefs entailing it. Those other beliefs aren't inferred from still other beliefs. Being basic entails not being inferred and it is this not-being-inferredess that so links basicality and evidence that in properly basic belief epistemology evidence and propositional evidence become interchangeable.

Basicality and Grounds

Plantinga adds that in-house beliefs typically have non-propositional grounds, namely those conscious states that give rise to the beliefs. The experiential grounds are such states as feeling forgiven upon having asked for forgiveness, feeling led when one asked for guidance, and having a sense of outside disapproval upon having acted wrongly. Presumably the idea is that these states are grounds for their corresponding beliefs, not (or not merely) in the sense that (for some people, in some contexts) they give rise to such beliefs but in the sense that they are what it is hard not to call non-propositional evidence for them. Analogies are proposed that in which other non-inferred beliefs are alleged to be related to their grounds as are in-house beliefs to their grounds. Thus we are offered these examples as analogies: upon being appeared to treely, one believes that one sees a tree; upon having a memory image of eggs moving on a fork toward one's mouth, one (at least properly believes that one) remembers having eggs for breakfast. Since *There is a tree* entails *There is an external world* and *I had breakfast today* entails *There was a past*, these propositions too are properly basic.

Insofar as one is asking whether the fourteen year old theist is involved in intellectual inelegance- failing some duty or exhibiting some vice – say, in believing that God has forgiven some committed sin – it is very easy to answer negatively. It seems harsh to expect that every theist who reaches the age of fourteen ought then to begin to engage in as rigorous a program of theism-assessment as her intellectual and other resources allow, or that she is intellectually nonvirtuous unless she does so.[52] This seems unrelated to whether or not she has evidence or grounds for that belief. Few people in human history seem to have had the inclination and the opportunity to engage in a very comprehensive program of assessment of their beliefs about ultimate matters. Plantinga of course thinks that typically the fourteen year old theist believes truly in the cases described. Further, given the analogies, he takes the grounds for those beliefs to be truth-preferring. The criterion for being a properly basic belief in S's noetic system seems, then, to include having grounds, which is tantamount to being based on non-propositional evidence.[53] If the *based on* relation includes justification and awareness of the justifying content as, so to say, fitting the proposition believed, then the account includes a characteristically internalist element and the resulting position is a non-Classical, non-Cartesian foundationalism. Further, any justification provided is defeasible; the subject's later experience in the same environment, conceptual considerations, contrary testimony from others, discoveries of relevant misleading circumstances, recognition of defects in the subject's perceptual apparatus or belief-processing faculties, and the like, can make revision of belief the proper thing; Plantinga refers to such things as defeaters. Both having in-principle-accessible evidence for P and being in principle sensitive to in-principle accessible contrary evidence, should there be any, are internalist requirements for justified belief. Thus internalist criteria appear in considering whether a belief is justified and/or in considering whether retaining a belief is justified. Both sorts of criteria appear in proper-basic-belief epistemology.

Thus one can state Cartesian foundationalism in what is orthodox Cartesian terms without having stated a self-referentially incoherent position; the price of accepting this view as an account of knowledge is that most of what we think we know, we don't know. While it seems true that proper-basic belief epistemology does not make any belief you like part of any particular proper-basicist's commitment, or leave her unprotected from their addition, it is also true that propositions incompatible with hers are equally properly basic for others with nothing in properly-basic epistemology available to decide among the set of incompatible but equally properly basic beliefs. It is not to properly-basic epistemology *per se* that one should look for help in this matter – it is an attempt to answer the B-question, not the E-question, and it is to answers to the E-question that one should refer matters about what is true, other than what is true about the right analysis of things such as *S knows that P* and *S is justified in believing that P without having propositional evidence for P.* Properly-basic epistemology is a form of foundationalism insofar as it expects that non-basic beliefs within a person's noetic system be justified by appeal to basic beliefs in that system and evidentialist insofar as it requires non-propositional grounds either as justification for the forming or the retaining of properly basic beliefs.

A Point About Phenomenology

Properly basic belief epistemology makes no distinction between experiences in which one has a sense of feeling forgiven[54] but lacks any phenomenology that can properly be called having an awareness of God beyond the phenomenology of feeling forgiven, and experiences in which, whether there is any feeling forgiven phenomenology or not, there is a phenomenology of having an awareness of God. Obviously, how one is to understand the notion of having an awareness of God is an interesting and difficult question,[55] and it is plausible that there is a continuum of possible (and also of actual) experiences on one end of which there isn't any, and at the other end there is intense, divine-awareness phenomenology with no exact formula for where the mid-point lies. Nonetheless, there are clear cases of the divine-awareness phenomenology. Where it is entirely lacking, why think that any experience that occurs provides grounds for belief in God? Feeling that was one was in the presence of King Henry VIII, without any Henryish-phenomenology, would presumably be no grounds whatever for believing that he had appeared.[56] Why, in the absence of – if I may be permitted the term – Godish phenomenology should feelings of being forgiven be any better regarding their subject experiencing God?

Put the point another way. Some religious experiences, as they occur and are understood within Christian traditions, have no Godish phenomenology and are, phenomenologically, sheer feelings of being forgiven. Perhaps God is the cause of such experiences; if so, one might properly infer to God's existence. Whether one can properly do so or not depends on the status of the claims that fill in the gap between feeling forgiven and God forgiving. Other experiences in the same tradition include Godish phenomenology. If any experiences provide (non-inferential) grounds for belief in God, it seems to me, it is these. Whether it is the experience itself, or a sufficiently accurate description of the experience, that is the believer's warrant that God exists seems irrelevant. Unless some relevant principle of experiential evidence is true regarding such experiences, then they don't provide grounds for belief in God. If the comments of this paragraph are substantially right, it is at least the case – even if properly basic belief epistemology is right – that the Boston Celtics version of that epistemology (the version that is most defensible, richest in correct content, and the like) will fine-tune its discussion of which experiences provide (non-inferential) grounds and which do not. Analogous comments apply to Doxastic Practice Epistemology of Religion, though they will not be repeated in our discussion of it.

Further, unless some principle connecting grounds and belief is true, it is puzzling why our having an experience with a Godish epistemology should be thought of as grounds for belief in God (taking grounds to have justificatory force, not merely causal impact); some such principle, it seems, will be required somewhere in the account of warrant or as a presupposition thereof.

Warrant Epistemology

Plantingean warrant is the stuff which, when added to acceptance and truth, yields knowledge. (S1) *S knows that P* entails, but is not entailed by, (S2) *P is true and S assents to P*.

Consider propositional function (S3) *It is true of S that X*. If (S3*) – the result of properly replacing "X" in (S3) – and (S2) entail (S1), what has replaced "X" describes warrant.[57] Plantingean warrant is a complex affair and we cannot do it justice here. The basic idea is to avoid internalism's problems by not being an internalist, avoid externalism's problems by offering a view as naturalistic as are those natural sciences whose resources the view uses when it comes to describing our cognitive mechanisms, and avoiding counterexamples that are produced by the possibility of cognitive misfirings by making the notion of our cognitive faculties properly functioning central to the account of knowledge. Plantinga offers this summary of his view:

> ... a belief B has warrant for you if and only if (1) the cognitive faculties involved in the production of B are functioning properly (and this is to include the relevant defeater systems as well as those systems, if any, that provide propositional inputs into the systems in question); (2) your cognitive environment is sufficiently similar to the one for which your cognitive faculties are designed; (3) the triple of the design plan governing the production of the belief in question involves, as purpose or function, the production of true beliefs (and the same goes for the elements of the design plan governing the production of input beliefs to the system in question); and (4) the design plan is a good one: that is, there is a high statistical or objective probability that a belief produced in accordance with the relevant segment of the design plan in that sort of environment is true.[58]

Explaining, let alone assessing, this notion of warrant goes beyond what is possible here. Plantinga claims that the concepts central to warrant are inherently teleological, incapable of being understood in ways totally divorced from the notion of conscious design. Further, he claims that since evolution selects for survival rather than true belief and true belief is not requisite to survival – lots of collections of false beliefs would be as survival-promoting as a collection of true ones – only if theism is true is it the case that it is more probable than not that our properly functioning faculties produce true belief. We cannot prove something like *Our cognitive faculties are reliable* since our alleged proof will be such that, for at least some set C of these very faculties, this is the case: unless they are reliable, we cannot tell how good the proof is. But the reliability of the faculties in C is part of what was supposed to be proved.

Of course not just any old sort of theism will do here. It must be a theism which favors a multiplicity of true beliefs. So Plantinga comments that:

... qua traditional theist – qua Jewish, Moslem, or Christian theist – he believes that God is the premier knower and has created us human beings in his image, an important part of which involves his endowing them with a reflection of his powers as a knower.[59]

A Plantingean argument will at most work for a variety of theism that holds or entails something of this sort:

(H) All human persons are created in God's image.

(H1) For any person S created in God's image, and any time T, the set B of beliefs S holds at T is such that there is at least one more truth in B than there are falsehoods in B.

or:

(H2) For any person S created in God's image, and the set B of beliefs that include all the beliefs that S ever has, there is at least one more truth in B than there are falsehoods in B.

Of course, (H1) and (H2) will need to be made more complex. Their more precise successors must refer to properly functioning faculties in friendly environments under conditions in which the design plan dictates that truth is the goal, not something else. How this will affect claims about the overall percentage of true beliefs within a total system of beliefs held by one made in the divine image I do not pretend to know. It is only in cases where friendly environment, properly functioning capacities, and truth specified by the design plan as the goal combine that an objective probability of truth applies. What per cent of belief-cases fall under this rubric I have no idea. It is not so very clear how to tell and so not at all obvious whether people typically have true beliefs in a majority of such cases. My interest here is in whether theism entails that they do. In what follows, I will simply take it that the indicated qualifications, and what others are necessary, are implicitly present.

There is a long distance between the general thesis that to be created in God's image includes having cognitive capacities to the thesis that if God creates human beings those capacities will be reliable in the sense of producing a percentage of true belief that exceeds .5 or that makes it more probable than not that, for any belief B and cognitive process C, B is more probably true than not by virtue of being produced by C, or any other similar claim of use to a Plantingean argument. While (H) is a part of standard theism, (H1) and (H2) are not. Plantinga's argument reminds one of Descartes' claims that *Necessarily, God is not a deceiver* and *Necessarily, if beliefs we cannot but have given our nature as human beings are false, then God is a deceiver.* Since beliefs that there are mind-independent physical objects are beliefs we cannot but have given our nature as human beings[60] it follows that there are mind-independent physical objects. To this, Leibniz replied that what is true is *Necessarily, God does not cause or allow us to be deceived unless God has sufficient reason to do so.* Since Leibniz thought that there are no physical objects – perception being unclear and unreliable thought and so not truth-revealing – he presumably also thought that in allowing almost all of us to be deceived about this matter, God had sufficient reason. Plantinga's argument is obviously, in this respect, more Cartesian than Leibnizean. But the Leibnizean sort of argument is worth pursuing.

Suppose idealism is true: there are perceptual experiences in the sense of one's being (as Chisholm would say) objectly-appeared to, but no physical objects. Then all of our beliefs whose truth entails that there are physical objects would be false. Would this be incompatible with theism being true? I take Berkeley to be right here in thinking that it would not.

Suppose colors and other secondary qualities are mind-dependent, as many have held. Then a great many of our beliefs about the properties of physical objects are false in the presence of our faculties functioning quite properly. What the percentage of such beliefs is among the whole class of our beliefs about objects I do not know, but lots of our beliefs about objects (or at least lots of people's beliefs about objects who have not taken secondary qualities to be mind- dependent, which must be the vast majority of humanity) have been false. Is this compatible with theism? I've never seen an argument from false belief regarding sensory qualities to the non-existence of God. Is this merely philosophical sloth and neglect?

Perhaps more importantly, a significant number of the religious beliefs held over time, including now, have been false because they are logically incompatible with one another. To be clear about this, consider only the beliefs that express the diagnosis and cure proffered by a particular religious tradition. Compare them with those of another religious tradition, and then another. I suggest that it becomes obvious that these diagnosis and cure propositions are not logically compatible. But then (given historical facts about the distribution of religious traditions among human populations) it follows that the majority of diagnosis and cure religious beliefs – the ones that religious traditions themselves take to be most basic – have been false. It does not follow directly that, to put it in familiar terms, most people go to hell if one of the diagnoses and cures is correct. But it does follow that being religiously mistaken is a widespread condition. Is this compatible with theism? This question is more complex and the answer more controversial. Plantinga, I assume, takes the answer to be affirmative. I agree. But then having cognitive capacities by virtue of being created in God's image is not as robustly truth-preferring as one might think. Perhaps it is not as robust as Plantinga's argument requires; perhaps theism does not entail that over half of our properly functioning faculty produced beliefs are true. The matter needs a closer look.

Plantinga recognizes, of course, that in various contexts the design plan does not make truth the primary goal; the proper exercise of one's cognitive faculties is sometimes not truth. When the primary goal is truth, Plantinga takes theism to entail that more often that not, we succeed. God would not make an organism whose goal was typically X and whose properly-functioning faculties typically yielded not-X or non-X. Regarding this claim, however, note two things. First, there is always the Leibnizean qualification: God would not do this without sufficient reason; that, says Leibniz, is what theism entails. But that isn't enough for Plantinga's argument.[61] Second, there is the question of how much of the time, according to theism, the goal is true belief. I wouldn't have thought that theism strictly had any entailments about that. There have to be what a philosopher of science might call auxiliary hypotheses concerning that matter, and the question would be how to rationally decide among them, and which hypothesis won the race. My concern is that Plantinga has too easily accepted one or another auxiliary hypothesis to theism as part of what theism entails.[62]

Doxastic Practice Philosophy of Religion

As we noted earlier, internalist accounts of *S knows that P* are subject to counter-examples in which S meets all of the specified conditions but it is an accident that S is right about P. Externalist accounts of *S knows that P* are subject to counter-examples in which S is in conditions that meet all of the specifications but S's belief-producers misfire so that S is just lucky that P is true. These pessimistic claims have more plausibility than an epistemologist would like, and thus there is an attraction to mixing internalist and externalist elements into a new epistemic brew. William Alston's doxastic practice philosophy endeavors to do this, though strictly his purpose in his most thorough discussion of matters directly relevant to our concerns is with epistemic justification of theism. All sorts of interesting issues arise as the exercise unfolds, arguing, not for the claim *God exists* but for the contention that *Some beliefs about God are justified.* Only a few can be explored here.

Alston writes:

> ... I think of a doxastic practice as the exercise of a system or constellation of belief-forming habits or mechanisms, each realizing a function that yields belief with a certain kind of content from inputs of a certain type. Such functions differ in the width of the input and output types involved. The input type could be something as narrow as a certain determinate configuration of specific sensory qualia, and the output type something as narrow as a belief to the effect that the object in the center of the visual field is Susie Jones.[63]

Doxastic practices are belief-forming practices. Tea-leaf reading, crystal ball gazing, and entrails-examining for non-physiological purposes, are all doxastic practices. Each moves from certain experiences to particular beliefs against a background of assumptions. Not all doxastic practices are reliable; Alston's concern in part is to distinguish reliable from unreliable ones.

A central motivating notion[64] for doing things doxastically is this claim:

A1. We are unable to give an adequate non-circular argument for the reliability of sense perception.[65]

And this:

A2. For no doxastic practice P, or type T of experience central thereto, are we able to give a non-circular argument A such that A is sound and valid and A's conclusion is *P is reliable* or *Experiences of type T are typically reliable* or any other conclusion that entails the reliability of P or the typical reliability of T.

The apparent consequence of A2 is the ultimacy of doxastic practices – each moves from its own base of operations and produces its own products, none intrinsically better or worse than any other. As this is Alstonianly unwelcome, he tries to avoid it. One of the most interesting questions concerning doxasticism is whether it entails relativism regarding doxastic systems, or among some favored but exclusive[66] set thereof.

Circular Justifications

Alston attaches great importance to the at least alleged fact that believing our perceptual experiences to be on the whole reliable cannot be justified without appeal to perceptual experience (taken to be reliable, of course); that our acceptance of logic cannot be justified without appeal to logic; and so on through the sources of our various types of belief. The alleged fact seems genuine. The question, then, is whether this actually creates a problem. Consider the claims: (PNC1) For any proposition P, it is not the case that P is both true and false and (PNC2) For any substance S, quality Q, and time T, it is false that at T S both has and lacks Q. Consider also the claim: (A) It is not possible that one provide a non-circular justification of either (PNC1) or (PNC2), where justification J of proposition P is non-circular if and only if *The propositions included in J are true* does not presuppose P. Suppose that there is no non-circular justification of (PNC1) or (PNC2). What problem arises? The standard view regarding (PNC1) and (PNC2) is something like this: they are necessary truths so fundamental to thought that any attempt to prove them will be either circular or appeal to something no more evident than they are. Since their truth is typically accessible to those who reflect on them, no general epistemological problem arises, though a problem of some sort may arise for those who do not see that they are true. It is logically impossible to offer a non-circular proof of every proposition one believes. It has not typically been supposed that this posed any problem. One is perfectly justified in believing without proof such propositions as (PNC1) and (PNC2) – propositions true in all possible conditions and all possible worlds. One is not justified in rejecting them.[67]

Similarly, consider the principle of experiential evidence (PE) For any person S and sensory experience E, if S's having E is a matter of its sensorily seeming to E that there exists some object O, then E is evidence that O exists.[68] If (PE) or some epistemological sibling is true, it is necessarily true. Whether one is having an experience in which it at least sensorily seems to one that one sees a cow or a tree is not typically beyond one's powers of discernment; just such experiences have been offered as providing examples of things one cannot be mistaken about.[69] If P plus some necessary truth N entails Q, then P by itself entails Q. (SC) *S at least seems to see a cow*, plus (PE), entails (SE) *S has evidence that there is a cow*. If (PE) is true, then (SC) entails (SE). It is logically impossible that one sensorily seem to see a cow and not have evidence that there is a cow. If this is so, it is so independent of whether some doxastic practice is reliable. A doxastic perceptual practice, however, presumably will embed these considerations into its epistemic core.[70] The practice depends on the principle, and not conversely.

Analogously, consider (RE) For any person S and religious experience E, if S's having E is a matter of its phenomenologically seeming to that God exists,[71] then E is evidence that God exists. Suppose further that (SG) *S at least phenomenologically seems to experience God*. (SG) plus (RE) entail (SEG) *E is evidence that God exists*. If (RE) is true, it is necessarily true. So (SG) by itself entails (SEG). It is logically impossible that one phenomenologically seem to experience God and not have evidence that God exists. If this is so, it is so independent of whether some doxastic practice is reliable. A theistic doxastic experiential practice, however, presumably will embed these considerations into its epistemic core.

If the above comments are correct, it is hard to see that there being no non-circular arguments for various fundamental claims is the deeply problematic problem it was advertised as being. Some propositions are basic in the sense that there is nothing more obviously true than they are from which they may be inferred. Some can be defended by noting that attempts to refute them assume them. Others have the feature that we assume them if we take ourselves to have experiential evidence of one sort or another. It is not clear that there is any logically possible alternative to this – any logically possible world in which this is not so.

Doxastic Presuppositions

Alston claims that:

> The existence of physical objects and the general reliability of sense perception are basic presuppositions of SP [our sensory doxastic practice of forming beliefs regarding the existence and properties of physical objects]; we couldn't engage in it without at least tacitly accepting these propositions.[72]

Perhaps so. It does not follow that our having evidence that there are physical objects occurs only in a context in which we must take for granted that there are. This is so even if (as seems plausible) our evidence for *There are physical objects* comes from our evidence that *There is this object* and *There is that object.*

If there are necessarily true principles of experiential evidence, they are true independent of doxastic practices and such practices provide conceptual contexts within which the principles are applied. On this view, doxastic practices are not ultimate in the sense of being the bottom line of appeal as far as justification is concerned. That status goes to necessarily true principles of evidence which are themselves capable of being embedded in more than one doxastic practice.

Defeaters

Crucial to Alston's notion of a properly structured doxastic system is the presence of defeaters – for example, inconsistency among formed beliefs entailing that at least one is false.[73] A doxastic system in action that produced ordered pairs of beliefs of the forms P and not-P would be radically defective, and a large proportion of such cases would be enough to reveal unreliability of the practice that yielded this result.[74] At the very least, then, an initially promising belief B – one that looks likely true – may be stripped of its promise by the occurrence of an equally qualified belief B* whose truth is incompatible with that of B. Inconsistency as problematic apparently is shared among doxastic systems as Alston conceives them, apparently without concern about there being no non-circular justification for so conceiving them.

Similarly, two doxastic systems D1 and D2, where D1 regularly produced belief that P and D2 produced belief that not-P in perfect correlation or high degree of the cases of belief-production would be such that at least one system was defective. The proposed remedy for such a case – for deciding which system to accept – is this: take the more widely accepted, more definitely structured, more important to our lives, possessed of more of an innate base, more difficult to abstain from, graced with more obviously true principles system among D1 and D2, and embrace it.[75]

It is not clear that these features, save perhaps the last, have anything to do with reliability – that is, producing true beliefs. Social entrenchment need have no such connection, nor clarity of structure, nor felt importance or actual importance, nor being difficult to abstain from; further, these may well yield different choices in different cultures and sub-cultures. "Innateness" is too vague a criterion to comment on without considerable preliminaries, and if the principles of D1 are more obviously true than those of D2 why isn't that enough to decide things in D1's favor unless D1's products are very loosely connected to its principles – unless its practices don't embody those principles well? The gist of the criteria give every appearance of being simply pragmatic. Pragmatism being but relativism in the hands of an entrepreneur, they give every appearance of capitulating to relativism. Put differently, they look like advice given in a context where reliability, however important, is no longer an operative consideration. Talk of what is "more firmly established"[76] is ambiguous. In order to yield justified confidence regarding reliability, it must mean something epistemic; in fact, its content seems basically sociological. Further, what if beliefs about what is more widely accepted, more definite in structure, more innately based, more important to our lives, and more difficult to abstain from themselves differ among those whose beliefs are D1-produced and those whose beliefs arise within D2? This is obviously logically possible, and not this-world unlikely. Further still, can't there be a relatively successful doxastic practice that favors conceptual elitism, loose structure, non-innateness, suspicion of what is widely valued, and disdain for what is hard to resist? Nothing in the Alstonian response[77] to the possibility of D1/D2 cases suggests anything other a relativistic result. This would be less troubling were there not sources of belief that generate commitments not compatible with the beliefs those in Anglo-American and European culture accept. At least until the arrival of New Agers, celebration of gods and goddesses, tea leaf reading, the curative power of magnets and crystals, searching the entrails of deceased animals as a source of accurate predictions, and the like were not taken seriously; now, in some sub-cultures, they are. Peyote ceremonies, schools of meditation claiming to lead to the recognition of truths about ultimate reality and the structure of personal identity, consulting the elders who in turn consult visions – these are but a few types of doxastic practices whose resultant beliefs do not accord well with those that Alston accepts. Appeal to doxasticism is supposed to provide an escape from the otherwise inescapable dilemma of there being distinct doxastic practices none of which is more or less justifiably accepted and each of which produces beliefs not consistent with those yielded by other practices.

Self-purification

A doxastic practice can be self-purifying (Alston suggests as example religious doxastic practices that once predicted datable apocalyptic events and no longer do so). One who thinks of magic as science badly done can argue that science is self-purified magic and one who thinks of religion as science badly done can think of science as self-purified religion. I suspect that such considerations are not worth pursuing in the absence of clear identity conditions for doxastic practices and for sub-practices within a practice, and these are not on offer.

Self-support

A different tactic is to argue that some practices give "significant self-support." Alston writes:

> ... consider the following ways in which SP supports its own claims. (1) By engaging in SP and allied memory and inferential practices we are enabled to make predictions many of which turn out to be correct, and thereby we are able to anticipate and, to some considerable extent, control the course of events. (2) By relying on SP and associated practices we are able to establish facts about the operation of

> sense perception that show both that it is a reliable source of belief and why it is reliable ... It cannot be assumed that any practice whatever will yield comparable fruits.[78]

Fair enough. Can we assume that there is no doxastic practice that (say) focuses on experiences of smoke patterns as they arise from cave fires. Different patterns signify different events in the overworld, the underworld, and the spirit world. Overworld events are (also) observed via interpreting cloud patterns, underworld events are observed by watching water eddies in sacred streams, and spirit world events are observed by interpreting the sacred dreams of community leaders. There are even accounts of how and why these epistemic sources are reliable. The various doxastic practices mesh nicely to form an overall coherent body of beliefs, and in turn the contents of these beliefs are used to understand local history and provide the conceptual context within which daily life is lived. But in none of this do we meet any of the theoretical entities or the lawlike counterfactuals of science. Why should we suppose that the linkages of common perception with science is a better indicator of the reliability of perceptual doxastic practice than is the linkage of common perception with overworld, underworld, and spirit world doxastic practices? Each collection of practices presumably will embody criteria that favor its sort of self-support over every other sort.

Once one ascends to as cumbersome and complex an item as a doxastic practice, one treats of the epistemological analogues of worldviews in metaphysics. If one thinks of practices or worldviews as essentially independent and autonomous, and denies that there is any set of propositions that lies beyond their jurisdiction, relativism is inevitable. The notion of propositions beyond the jurisdiction of doxastic practices, like that of propositions beyond the jurisdiction of worldviews, involves the ideas of at least two modalities of *being beyond.* A proposition P may be beyond the jurisdiction of every doxastic practice by virtue of (i) being a proposition every doxastic practice must in some manner include – as an unexpressed assumption, part of background beliefs, something entailed by the descriptions of its epistemically significant experiences, or whatever, or (ii) of being necessarily true regarding every doxastic practice and so inconsistent for any to deny.[79] Perhaps something can be done to adjudicate between doxastic systems along the lines of considering what sorts of propositions might fall within (i) and/or (ii). But while perhaps not every doxastic system can provide significant self-support, presumably various practices can do so with comparable impressiveness. If so, one can be particularly grateful that doxastic systems, however interesting, are not the ultimate constituents of the epistemological world.

Keith E. Yandell
University of Wisconsin – Madison U.S.A.

NOTES

[1] I here use "explanation" in its metaphysical sense, not its epistemological sense. In the latter sense, but not the former, *There is an explanation* entails *Someone has offered an explanation.*

[2] If one objects that issues of logic – of whether certain issues have one modality as opposed to another, whether certain inferences are valid or invalid, whether certain claims in probability theory are correct and whether they are properly applied in certain cases, and the like – arise relative to E-question, I have no quarrel. If one says that the E-question actually is a metaphysical question at least as much as it is an correct answer to which must mix value theory with epistemology, and that there is a central and crucial part at least of epistemology that is not mix of value theory of some sort with something else (say, epistemology).

[3] I take internalist answers to the E-question to be already naturalized in the latter sense and unnaturalizable in the former sense, but this is not the place to argue for these claims.

[4] Unfortunately, at the time of writing, Plantinga's *Warrant and Christian Belief* is but a gleam in a publisher's eye. This is the volume in which Plantingian externalism is to be applied to religious belief.

[5] Purists of course allow that often there is probabilistic evidence – evidence that some proposition is probably true or probably false. They simply deny that all evidence is inherently or even profitably put in terms of the quantitative or qualitative probability. For example, they see neither need to think of (i) *The chair's seat being warm even though no one was home is evidence that the family Golden Retriever has been napping in his favorite spot* in such terms as (ii) *The probability of* ***The family Golden Retriever has been napping in his favourite spot*** *is raised given* ***The chair's seat is warm even though no one was home*** nor gain in so doing. Robust probabilists will take it to be a necessary truth that (ii) captures (i) and that (i) is simply (ii) cast in vague terms, and purists concerning the notion of evidence will think robust probabilists are in chains to a dubious theory.

[6] They are far more evident in their application than in their statement, appearing in the way claims are cast rather than as explicitly stated axioms.

[7] He manifests none of the putting-things-in-probabilist-terms characteristic of robust probabilists.

[8] I take it to be reasonably clear that more is involved than simply style; a bit of evidence that this is so comes up in our discussion of Plantinga's critique of Swinburne.

[9] Hick 1995. Originally published as *The Rainbow of Faiths* (London: SCM Press, Ltd. 1995). The quotations are from pages 68, 60 and 71.

[10] The book is presented as a series of questions to which Hick responds. Here the questioner (Phil) asks "Or, presumably, realities" and Hick replies "I don't think so ..." Here Hick hues to the ghost of his former monotheism, as he does again in this passage from page 69: " So it seems to me that the most reasonable hypothesis is of a single ground of all savific human transformation, rather than of a plurality of such grounds." In the last passage cited here, it turns out that *Is there but one transcendent reality, or two or more, or maybe an infinite number?* is sheer nonsense, like *Does quigliness fastigate pruntically?* This is but one multitude of contradictions that comprise Religious Pluralism.

[11] The elements that together yield the inconsistencies of Religious Pluralism, and the ways in which attempts are made to avoid the inconsistencies, receive detailed attention in Yandell 1993, pp. 187-211, and 1998, Chapter Five.

[12] The idea is not that he thinks of God as eternal, but that he thinks that *God exists* is neither true nor false, and that this is the Christian view of the matter.

[13] Phillips 1965, pp. 1, 2.

[14] Phillips 1976, p. 174.

[15] *Ibid.*, p. 181.

[16] Though Phillips subtly suggests that those who think that God exists must be thinking of God as an exotic physical object.

[17] It seems typical of Phillips' writing that his descriptions of the beliefs that Christians don't hold are true if one understands those descriptions in one way and false if one understands them in another. The conclusion he draws requires that the descriptions be understood in both ways at once.

[18] Phillips 1967, p. 69.

[19] Phillips 1970, p. 66.

[20] This is just as well, given that such theories typically are meaningless on their own standards and/or dismiss from meaningful discourse things their authors plausibly think reasonable to believe – e.g. the current deliverances of the more theoretical claims of science.

[21] Flew 1976, pp. 14, 15.

[22] Plantinga describes (PA1), requiring one to be neutral regarding theism and atheism (in the ordinary sense) as "entirely correct, if something of a truism". This is surprsing, as it (if it is taken to cover belief formation) is incompatible with accepting theism being a properly basic belief. Perhaps Plantinga is understanding (PA1) simply as a rule covering debates or arguments concerning whether God exists or not, not as a principle covering belief.

[23] Some among them accept the theological claim that the evidence is very strong, and we fail to see its strength due to sin.

[24] The only such response, so long as we leave aside prudential considerations – e.g. that an optimistic patient may be more likely to recover even if, given the evidence, recovery is no more likely than non-recovery.

[25] Strictly, to a similar view suggested by Michael Scriven.

[26] Or: *Necessarily, if the being B that is identical to God who has created all human beings exists, then for every human being H, if H exists then B created H.*

[27] One can then try to restate (PA2) so that it applies to a subclass of existential propositions so defined as not to allow propositions like *There exists a human being not created by God* as members. Here, it seems more profitable to turn to Plantiga's defence of Properly Basic Beliefs, which includes various sorts of existential propositions within its scope.

[28] Rowe has other contributions to the philosophy of religion. His *The Cosmological Argument* and *Thomas Reid on Freedom* are superb treatments of cosmological arguments and libertarian freedom.

[29] Quite possibly both (R) and (R*) need qualification before either is the best representative of the type of view it represents.

[30] See Swinburne 1979, cited hereafter as **EG**. Strictly, Swinburne appeals to Bayesian probabilism – this is not the place to distinguish this variety from others, nor do our purposes require this.

[31] **EG**, page 282.

[32] "Where the prior probability is intrinsic probability, the second factor [empirical background knowledge] does not play any role..." **EG**, page 106.

[33] Scope in fact plays little role here; "where we are dealing with a theory of large scope, scope is of far less importance than simplicity in determining prior probability. The intrinsic probability of theism seems to depend mainly on just how simple a theory it is." **EG**, page 282. Nor is it clear on his view how to view the scope of theism; see **EG**, page 106.

[34] Plantinga 1990, p. 60.

[35] **EG**, page 254.

[36] Plantinga 1990.

[37] This is, of course, not to be read as entailing that X exists; it is to be read as a statement about how things appear to S, whether they are that way or not.

[38] The corresponding proposition to *It (epistemically) seems to S that Zeus is present* is of course *Zeus is present to S.* Perhaps (S*) should be more complex – e.g. perhaps it should include the qualification *S has no reason, independent of its seeming that x is present, for thinking the members of Q to be globally false.* Perhaps the idea is so that so long as one member of Q is not eliminated, the probability that Swinburne desires will not be forthcoming. Arguments composed of questions are not easy to track.

[39] Swinburne would, I assume, claim that Alston's *Doxastic Practice Epistemology of Religion* assumes a principle of credulity, as in effect it does, though not necessarily in a probabilistic version.

[40] I.e. not claim that every proposition has an intrinsic probability (a probability given only necessary truths), and not claim that for every proposition P and Q, P has some probability on Q. This would free one as well from the apparent artificiality of attaching a probability to necessary truths and necessary falsehoods. After all, *If they have any intrinsic probabilities, necessary truths have a probability of 1. and necessary falsehoods have a probability of 0* does not entail *Necessary truths have a probability of 1. and necessary falsehoods have a probability of 0.*

[41] This principle of experiential evidence of course is incomplete. For one thing, it mentions no defeaters. It does not consider the possibility of illusion. The right principle of experimental evidence will take account of two simple ideas: that things seem so-and-so is some evidence that they are so-and-so and that it is logically possible that things seem so-and-so when they are not. In *The Epistemology of Religious Experience* (Yandell 1994), I offer a better principle of experimental evidence. Since my purpose there is to discuss what force religious experience has as evidence for religious belief, I offer a version of the principle that is, I suspect, too strong since if numinous religious experience is evidence even given standards higher than necessary it will also be evidence on lower standards. But something like the principle offered there seems to me to be right and whatever version is true is necessarily so.

[42] Obviously various questions about appropriate phenomenology, effability, social science explanation of religious experience, and the like, need to be taken into account. Cf. Yandell 1994, 1999, and 1997.

[43] Plantinga 1979, 1983, 1986.

[44] P is incorrigible to S if and only if S believes that P and it is logically impossible that S believe that P and P be false.

[45] For all that Plantinga is a proper foundationalist and evidentialist if grounds count as evidence, as they should.

[46] He takes an exactly analogous argument to apply to Ancient-Medieval Foundationalism.

[47] *I (Kim) am not a prime number at T* is to be understood as to entail *I (Kim) exist at T* so it is contingent even though *If Kim exists at T then Kim is not a prime number at T* is a necessary truth.

[48] Other minds don't come into the story directly. Descartes, I think, takes it that if we get physical objects, we get human bodies, and then can offer the argument by analogy for the existence of other minds.

[49] There is no reason why an evidentialist need hold that proper belief requires *propositional* evidence. Experimental evidence – which presumably can be propositionally

expressed, but need not be – will do. More importantly, experimental evidence can typically be overcome; it is *defeasible*. The core idea is:

(D) For any person S, experience E, and logically contingent proposition P, if S's having E at T is evidence for P, it is logically possible that S have an experience E* such that if E* is reliable either (i) P is false, (ii) E is not evidence for P or (iii) E* is as good evidence against P as E is for P.

If all experimental evidence for logically contingent propositions is defeasible, the no one who accepts such propositions on the basis of experimental evidence does so with perfect safety from error. Some logically contingent propositions are belief-entailed, where a proposition is belief-entailed relative to person S if and only if S *believes that P* entails *P is true*. If a person S at is , say aware of feeling fatigued and thinks *At least I still exist now*, S's belief that she exists now is not defeasible because it is belief-entailed, and her (even apparently) feeling fatigued id indefeasible grounds for that belief. Thus (D) needs revision via:

(D) For any person S, experience E, and logically contingent proposition P provided E is not the ground of belief-entailed belief by S that P, if S's having E at T is evidence for P, it is logically possible that S have an experience E* such that E* is reliable either (i) P is false, (ii) E is not evidence for P or (iii) E* is as good evidence against P as E is for P.

Relevant to this are two further truths:

(D1) Necessarily, if for some person S and non-belief-entailed proposition P, S's having experience E is evidence for P, then P is defeasible.

(D2) If (D1) is true, then Necessarily, no one who accepts a logically contingent non-belief entailed proposition on the basis of experimental evidence does so with perfect safety form error.

Every logically contingent proposition is possibly false; every experience one has that is evidence, for a logically non-belief-entailed contingent proposition is possibly misleading in the sense that it is logically possible that the experience be just as it is and the proposition for which it is evidence be false. Further, this is the case in every possible world; it is logically impossible that things be otherwise. Epistemological efforts to change this are like attempts to square the circle; one may learn from them, but their success is not an option.

[50] Plantinga 1983, p. 24

[51] Clifford 1879, p. 183.

[52] Analogous remarks hold for atheists, Buddhists, Jains, and so on.

[53] Where it is not so clear whether *based on* means "is caused by" or "is justified by" and where the experiential or non-propositional evidence is expressible in propositions.

[54] Similarly for feeling led, feeling loved, and so on.

[55] Rudolph Otto's classic *The Idea of the Holy* (Otto 1990), perhaps with the qualifications suggested by Yandell 1971; Pike 1992.

[56] How much Henryish phenomenology would make a difference is no doubt controversial.

[57] Where (S3) by itself does not entail (S1) and (S1) does not entail (S2).

[58] Plantinga 1993, p. 194.

[59] *Ibid.*, p. 236.

[60] I waive the question as to how Descartes could, on his own terms, know this.

[61] One cannot, of course, appeal to this sort of consideration in dealing with whatever problem there is in the majority of persons having had false religious beliefs and ignore it when offering a Plantinga-type argument. I do not suggest that Plantinga does this.

[62] Not what bare theism entails, but what Christian, Jewish or Islamic theism entails.

[63] Alston 1993, p. 155.

[64] Both rationally motivating (as an essential element in the argument – **PG** 103) and psychologically motivating.

[65] And, similarly, religious and other forms of experience.

[66] "Exclusive" in this sense: doxastic practices A and B are exclusive of one another if there is a set of beliefs SA such that practicing A according to A's rules yields SA and a set of beliefs SB such that practicing B according to B's rules yields SB, and the set C of beliefs produced by conjoining SA with SB in logically inconsistent.

[67] One might nonculpably be sufficiently confused or otherwise cognitively underfed as to not be unjustified in not accepting them.

[68] No doubt the actually correct version of a principle of experiential evidence regarding sensory experience is more complex; I've considered various alternatives in the *Epistemology of Religious Experience* (Yandell 1994).

[69] For example, it is often asserted that S *believes that S sensorily at least seems to see a cow* entails S *at least sensorily seems to see a cow* and thus one cannot be wrong about one's beliefs concerning how things sensorily seem to one.

[70] Of course, as noted, things will get much more complex than anything stated here before one has anything like a really defensible account of the epistemology of sense perception.

[71] Obviously, comments about (PE) apply with all the more force regarding (RE).

[72] Alston 1993, p. 164.

[73] If formed beliefs P and Q are contradictions, one must be false; if they are contraries, both may be.

[74] See **PG**, 170. Presumably *one* case in which applying a practice according to its own rules yielded both P and not-P would at least call for a revision. Identity conditions among doxastic practices are not clear enough, I think, for one to tell whether this would be revision or revolution. The same thing can be true, of course, for theories.

[75] The list is at **PG**, 171.

[76] See **PG, 172.**

[77] The one on **PG**, 172.

[78] **PG**, page 173.

[79] Perhaps one could think of (ii) along the lines of there being some minimal doxastic practice D such that it is necessarily true for any doxastic practice D* such that D* produces a range of beliefs distinct from those produced by D, D* must contain or presuppose or entail D. I'm not clear enough about identity criteria for doxastic practices to be very confident about how good a way this would be to make the relevant sort of suggestion.

REFERENCES

Alston, W.: 1993, *Perceiving God*, Cornell University Press, Ithaca.

Clifford, W.K.: 1879, *The Ethics of Belief, Lectures and Essays*, Macmillan, London.

Flew, A.: 1976, *The Presumption of Atheism*, Pemberton, London.

Hick, J.: 1995, *A Christian Theology of Religions*, Westminster John Knox Press, Louisville, KY.

Otto, R.: 1990, *The Idea of the Holy*, Oxford University Press, London.

Phillips, D. Z.: 1965, *The Concept of Prayer*, Routledge, London.

Phillips, D. Z.: 1967, 'Faith Sceptisism, and Religious Understanding', in D.Z. Phillips, *Religion and Understanding*, Oxford University Press, Oxford.

Phillips D. Z.: 1970, *Death and Immortality*, St. Martin's Press, New York.

Phillips, D. Z.: 1976, *Religion Without Explanation*, Blackwell, Oxford.

Pike, N.: 1992, *Mystic Union,* Cornell University Press, Ithaca.

Plantinga, A.: 1990, 'Justification and Theism', in M. D. Beatty, *Christian Theism and the Problems of Philosophy*, Notre Dame University Press, Notre Dame.

Plantinga, A.: 1979, 'Is Belief in God Rational?', in C.F. Delaney (ed.), *Rationality and Religious Belief,* Notre Dame University Press, Notre Dame.

Plantinga, A.: 1981, 'Is Belief in God Properly Basic?' *Nous*, **15**, 41-51.

Plantinga, A.: 1983, 'Reason and Belief in God', in *Faith and Rationality*, Unversity of Notre Dame Press, Notre Dame

Plantinga, A.: 1986, 'Epistemic Justification', *Nous* **20**, 3-18.

Plantinga, A.: 1993, *Warrant and Proper Function,* Oxford University Press, Oxford.

Swinburne, R: 1979, *The Existence of God*, Oxford University Press, Oxford.

Yandell, K.: 1971, *Basic Issues in the Philosophy of Religion*, Allyn and Bacon, Boston.

Yandell, K.: 1993, 'Some Varieties of Religious Pluralism', in James Kellenberger (ed.), *Inter-Religious Models and Criteria*, St. Martin's Press, New York.

Yandell, K.: 1994, *The Epistemology of Religious Experience*, Cambridge University Press, Cambridge.

Yandell, K.: 1997, 'Religious Experience', in P. L. Quinn and C. Taliaferro (eds.), *The Blackwell Companion to the Philosophy of Religion*, Blackwell, London, 367-375.

Yandell, K.: 1998, *The Philosophy of Religion*, Routledge, London.

KENT JOHNSON AND ERNIE LEPORE

KNOWLEDGE AND SEMANTIC COMPETENCE

I INTRODUCTION: WHAT IS THE PURPOSE OF A THEORY OF MEANING?

1.0 Motivation and General Considerations

This discussion is about linguistic competence – the ability of speakers to understand their language. Our focus, in particular, is on *semantic* competence, an ability to *interpret* language. To see its theoretical interest, consider an unusual description of a familiar type of phenomenon. John sees Mary searching for something in her living room. He surmises she has misplaced her scarf. Remembering recently having seen it under the table, he believes that if she knew what he remembered it would facilitate her search. He takes a short breath; the air in his lungs releases at a slow steady rate; his vocal folds contract and relax in an elaborate fashion; and as the air passes into his mouth, his jaw, lips and tongue move in complicated ways, all of which serve to create a specific vibratory pattern, which sounds like an utterance of, 'I saw your scarf under the table'. The sound pattern bounces off sensitive bits of tissue in Mary's inner ear, and shortly afterwards, her search ceases with the scarf recovered.

Many of these details are of theoretical interest, but our focus will be on what enables Mary to recognize that John's utterance means at some time prior to it, *John saw Mary's scarf under the table*. The most common answer is that linguistic competence equals *knowledge* of a theory, or 'grammar' (see §1.1). In what follows we will review influential answers to the question, 'Are speakers able to understand their (first) language in virtue of bearing a doxastic relation to a grammar?' Before we start, we will outline some technical terms we will employ throughout.

1.1 Terminology.

A *grammar* is an abstract entity; in particular, it produces syntactic structures of a language and assigns them meanings and phonological forms. Though only part of a grammar generates interpretations of sentences, we will frequently speak of grammars and knowledge of grammars. Occasionally, when more specificity is required, we will speak of semantics and semantic knowledge. A *psychogrammar* is a mental state of knowing a grammar (if such a state exists); it is 'a mental condition on a par with the state of *thinking of the number 3*' (George 1989b, 90). In short, a speaker's grammar is an object he knows, and his psychogrammar is his state of knowing it. Thus, as George notes, '[w]e might come to be able to articulate the object of a speaker's knowledge, the grammar, without thereby being able to say

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 707-731.
© 2004 *Kluwer Academic Publishers.*

how that object is represented by the speaker. The grammar is what is represented, not what is doing the representing' (George 1989b, 91).

A *physiogrammar* is a physical state (if such a state exists) of the speaker that realizes the psychogrammar. Just as a correct theory of a speaker's grammar does not render one theory of her psychogrammar more plausible than all others, a correct theory of her psychogrammar does not render one theory of her physiogrammar more plausible than all others either. On this picture, if knowledge of language enters into an explanation of behavior, then, if such explanations are causal, a psychogrammar enters into the explanatory causal chain. If the psychogrammar is identical to the physiogrammar, then of course the latter is a part of that causal chain. Grammars, though, since abstract, *cannot* be causally efficacious; they are objects of knowledge, and so they can be no more causally responsible for behavior than Santa Claus should little Billie become joyful when he anticipates that Santa Claus is coming to town (George 1989b, 92).

It is common ground that there is a systematic relationship between knowledge of a grammar (i.e., one's psychogrammar) and whatever other beliefs one forms as a result of linguistic competence. Suppose John hears Mary utter, 'Hesperus burns bright tonight', and according to his knowledge of grammar, (roughly) an utterance of 'Hesperus burns bright tonight' is true iff Hesperus burns bright on the evening of the utterance. He will, *ceteris paribus*, believe that Mary said Hesperus burns bright that night. A *processing algorithm* (if one exists) is an abstract object that describes processes of linguistic perception and production. Such an algorithm takes John from his grammar-induced belief and his perceptual belief (something to the effect that Mary produced an utterance of a certain form) to his belief about what was said.

In §2 we will in discuss the plausibility of supposing speakers have some sort of knowledge of a grammar of their language. The views we consider address whether speakers are doxastically related to grammars of their languages, and if so, what the nature of that relation is. In §3, we focus on an argument designed to show that linguistic competence cannot be adequately explained by describing a grammar (as characterized above) and the relation one bears to it.

II TACIT KNOWLEDGE

2.0 Introduction: Cognitivism.

Do competent speakers know a grammar of their language? We speak of 'knowledge of meaning' and 'knowledge of language'. Yet whatever we mean by such locutions talk of this sort of knowledge differs from the knowledge that 2 + 2 = 4, or that one's favorite cup is filled with coffee. Unlike the latter two, the former seems to be rarely (if ever) explicitly statable by its knower. Most speakers cannot state principles which would explain the ungrammaticality of (2.1) and the grammaticality of (2.2).

(2.1) *John believed that any senators were drunk.
(2.2) John doubted that any senators were drunk.[1]

Most speakers, though capable of using (2.3) and (2.4), cannot explain why only (2.3) permits substitution of identities *salva veritate.*

(2.3) Mary saw the student leave.
(2.4) Mary saw that the student left.

If the student who left is the happiest girl in Newark, then Mary saw the happiest girl in Newark leave; not so for (2.4). Furthermore, unlike (2.3), (2.4) can be true even if Mary saw no one leave. Perhaps she noticed that the previously occupied chair was empty (for further discussion of differences between (2.3) and (2.4), see Higginbotham 1983; 1995).

Many authors dismiss the ascription of knowledge to speakers as 'unnatural' or 'incoherent', since we lack conscious access to it (Foster 1975, 2; cf. Schiffer 1987, 255-261, Dummett 1975). A traditional and still common way to deal with this problem is to attribute *tacit knowledge* of a grammar (Chomsky 1965, 8; 1986, 266). Speakers have propositional knowledge of a grammar, but such knowledge is inaccessible to consciousness. Speakers understand a sentence of their language, because they exploit a grammar to (unconsciously) compute a meaning theorem for the sentence. Positing tacit knowledge is justified if so doing explains linguistic behavior better than any rival account.

Any theory which treats speakers as linguistically competent *in virtue of* tacit knowledge of a grammar we shall call **cognitivism**. Cognitivism is the received view in linguistics, as can be seen by a glance at the introductory chapters to linguistics textbooks: (Culicover 1997, 1-3; Cowper 1992, 1-4; Hagemann 1990, Larson and Segal 1995, 9-22). In this section, we will discuss an attempt to justify cognitivism, as well as some famous objections. (We will then discuss theories like cognitivism classified according to the relation they posit between a semantic theory and a speaker's mental state.)

Two points about cognitivism are relevant. First, noted by Higginbotham (1994), talk of knowledge of meaning can be misleading: 'knowledge of meaning is a *phenomenon*, not a hidden *explanandum* [sic]. A psychology for me that simply omitted to state that I knew the words 'snow is white' meant in my speech that snow is white would be in so far forth a false psychology' (p. 88; cf. Segal 1994, 115-116). However, at stake is not whether linguistic competence *per se* requires semantic knowledge. Our concern is whether such competence requires (tacit) knowledge of *all* of a semantic theory, which, *prima facie,* is not what a psychological theory should predict. Secondly, 'cognitivism' is ambiguous between requiring that a grammar only specify knowledge one has of the structure, meaning, and phonological properties of sentences (a position endorsed by Chomsky (1965), pp. 8-9 and Samuel Keyser[2]) or requiring that whatever procedures a grammar uses to derive appropriate meaning theorems are psychologically real as well, in the sense that they are 'mirrored' by a process in one's mind (i.e., in one's psychogrammar) (cf. Davies' discussion of his 'mirror constraint' (1981, 53-55; 1987, 446-447; Chomsky 1986, 263-273). The latter requires (something like) propositional knowledge of the axioms of a particular meaning theory, whereas the former is quiet about the nature of linguistic knowledge. Our discussion of cognitivism will focus exclusively on the latter, though much of what we say here and in a later discussion of dispositionalism will apply, *mutatis mutandis*, to the former view.

2.1 Justifying Tacit Knowledge of a Grammar.

In articulating how cognitivism might be justified, we appealed to a 'best theory' principle. While defending this principle would require delving into more philosophy of science than we have space for, Fodor's (1968) defense is worth commenting on. Fodor attempts to justify a general principle for positing tacit knowledge, the crux of which is that one way to explain how a type of behavior might occur is by building a machine that simulates the behavior. His argument divides into three stages. First, he argues that a computer's programming language 'can be thought of as establishing a mapping of the physical states of a machine onto sentences of English such that the English sentence assigned to a given state expresses the instruction the machine is said to be executing when it is in that state' (p. 639). Second, if the programmed machine 'optimally' simulates an organism's behavior, then the machine exhibits a type of behavior (if and) only if the organism does, and for each type the machine can exhibit, the sequence of (computationally relevant) states of the machine resulting in that behavior can be mapped onto a sequence of English sentences, such that the latter constitutes a true etiology of the machine's output. Finally, he invokes a general principle of inductive inference, namely,

> If D is a true description of the etiology of an event e, and if e' is an event numerically distinct from e but of the same kind, then it is reasonable to infer, *ceteris paribus*, that D is a true description of the etiology of e' (p. 639).

He concludes,

> If X is something an organism knows how to do but is unable to explain how to do, and if S is some sequence of operations, the specification of which would constitute an answer to the question 'How do you X?', and if an optimal simulation of the behavior of the organism X-s by running through the sequence of operations specified by S, then the organism *tacitly knows* the answer to the question, 'How do you X?', and S is a formulation of the organism's tacit knowledge (p. 638).

To be sure, Fodor's defense is schematic. Filling in details would involve resolving a number of issues, for instance, what counts as behavior. Since Chomsky's review of Skinner's *Verbal Behavior* (Chomsky 1959), it has been widely acknowledged that there is more to behavior than what behaviorism included. But, as is also well known, including more than overt physical behavior in an *explanandum* engenders other sorts of problems. Another question requiring an answer concerns how to construct a theory of event types in a principled way so that relevant human and machine behaviors get typed together. This problem also increases in complexity when the extension of 'behavior' is expanded. A third question concerns what counts as optimal simulation of behavior? Since there have been but a finite number of human behaviors, there are infinitely many different ways of producing those behaviors. For that matter, there are infinitely many different ways of producing reasonable infinite extensions of those behaviors. So, beyond extensional equivalence, we need additional criteria for what counts as optimal simulation. What these criteria are and what justifies them is well nigh tantamount to explaining what makes for a good theory, or why one theory is to be accepted over another. (Further discussion of Fodor's argument is found in Graves et al. 1973.)

Wright notes that machine simulations of complex behavior do not always license intuitively plausible ascriptions of tacit knowledge. It is possible to write a program that simulates a homing pigeon's ability to find its way home from indefinitely many distant locations, but Wright contends that the bird lacks any sort of tacit knowledge of a homing theory that issues in homing theorems about where it should fly to next (Wright 1986a, 41-42, 1986b, 235-37). (This type of argument will receive detailed discussion in §2.2 and §2.3.)

Quine put forward a powerful and influential objection to positing tacit knowledge of grammar Quine (1972). First, he observes that any finitely axiomatizable theory can be finitely axiomatized in infinitely many ways. So, if there is one finite grammar of a language, there are infinitely many. Furthermore, such grammars are *extensionally equivalent*; they all generate the same sentences, and assign them the same meanings and phonological forms. Thus, if English has a finite grammar, it has infinitely many. Second, Quine distinguishes two relations one might bear to a grammar. In his terminology, either it *fits* the linguistic behavior of competent speakers; or, if an adult learned, say, English (for the first time) by memorizing a particular grammar, then that grammar – unlike extensionally equivalent ones – *guides* his behavior. Positing tacit knowledge of a grammar presumably amounts to linguistic competence in virtue of speakers being guided by the grammar in some sense.[3] Thus, according to cognitivism, one grammar is 'special' in the sense that it is the one used, i.e., it correctly describes the mental processing that underlies sentence comprehension in a way that its extensional equivalents do not. But then even where there is complete agreement about sentences of the target language, one grammar still must be singled out from its extensional equivalents. What justifies selecting one over another? As Quine puts it,

> If it is to make any sense to say that a native was explicitly guided by one system of rules and not by another extensionally equivalent system, this sense must link up somehow with the native's dispositions to behave in observable ways in observable circumstances (Quine 1972, 444).

Thus, the task Quine sets for the cognitivist is to find 'a criterion of what to count as the real or proper grammar, as over against an extensionally equivalent counterfeit' (p. 448; cf. George 1986, 493-496 for further discussion of how the ascription of tacit knowledge of a grammar is not fully justified by the kind of behavioral data Quine is concerned with).[4]

As a point of scholarship, Quine's wording is ambiguous. His text supports characterizing the project as what Davies calls 'Quine's challenge', which involves answering how there *can* be empirical evidence to warrant attributing tacit knowledge of one theory rather than another, extensionally equivalent, one (Davies 1987, 442). But it also supports a reading under which Quine's attack on tacit knowledge centers around the plausibility of there actually being evidence favoring one grammar over its extensionally equivalent counterfeits.[5] One could satisfy the former and not the latter. In some possible world, when supplied with hypnotic suggestion, we immediately write down a particular grammar. This scenario only shows how there can be empirical evidence, not that there is empirical evidence. The latter view seems more Quinean in spirit, and it is also the more difficult and pertinent challenge. Thus, unless explicitly noted otherwise, references to Quine's challenge will be to the latter interpretation.

In reply to Wright (1981), which presents a version of Quine's challenge, Evans suggests that the challenge can be met by 'providing a causal, presumably neurophysiologically based, explanation of comprehension' (Evans 1981, 127). When such explanation is available, Evans claims, 'we can simply see' which theory is correct (*ibid.*). Evans goes on to suggest three additional plausible types of empirical evidence for one of a set of extensionally equivalent grammars as tacitly guiding a speaker.[6] First, empirical evidence for the theory we actually use could come from the patterns in which we acquire dispositions, and second, from the patterns in which we lose dispositions, perhaps due to linguistic impairment. Thirdly, evidence can be culled from our (empirically testable) perceptions of linguistic structure in sentence perception (Evans 1981, 127-29; cf. Chomsky 1986, 252-87, Larson and Segal 1995, 56-62). (A clever thought experiment designed to show such evidence could be misleading is in Davies 1987, 451-53.)

2.2 Do All Processes Involve Tacit Knowledge?

We turn now to a well-worn argument against any attempt to explain linguistic competence with tacit knowledge. The argument has more critics than defenders, though Searle has employed versions of it (Searle 1983, 262-272; 1984, 28-31, 47-50). It goes something like this:

> Suppose you posit a cognitive state called tacit knowledge to explain linguistic competence. If the general line of reasoning for positing this state is sound, why can't we invoke cognitive states to explain digestion? Just as competent speakers cannot explain how they know which strings are meaningful and which are not, so too proficient digesters cannot explain how they alter their stomachs to appropriately digest some food and reject indigestible food. In short, they 'interpret' their digestible input correctly and 'judge' the indigestible input as not part of their dietary corpus. But since the ability to digest is not *cognitive, we* should not posit a cognitive state to explain it. *Mutatis mutandis,* we should not posit tacit knowledge of a semantic theory to explain linguistic abilities.

Discussions of versions of this argument are in, *inter alii*, Nagel 1969, pp.172-174, Fodor 1975, p.74, fn.15, Chomsky 1986, pp.239, 241, and Wright 1986, pp.41-43.

A primary response is to defend differing general structures of the best theories of linguistic competence and digestion: unlike digestion, the best theory of linguistic competence entails that 'a representation of the rules they follow constitutes one of the causal determinants of their behavior' (Fodor 1975, 74; cf. Chomsky 1986, 244, 253-257). Employing linguistic capacities produces or requires certain belief-like states, such as whether 'Sta nevicando' means that it's snowing, or whether a string is a sentence of one's language. For linguistically competent organisms, their competence involves such beliefs. (This is an empirical defense, and so it would not follow that such beliefs are constitutive of one's competence, only evidence for it.) On the other hand, there is no reason to impute beliefs to digestively proficient organisms as such. We can account for the ability to digest good food and reject bad food without positing beliefs, explicit or implicit. (Cf. Lepore 1996 for a discussion of the epistemological import of linguistic beliefs.)

Nagel offers additional support for tacit linguistic knowledge, which invokes consciousness. He argues that 'In the case of language-learning...conscious apprehension of the data...is essential; and what the individual can do as a result of his linguistic capacity is to speak and understand sentences' (Nagel 1969, 174). He compares statements of a tacitly known grammatical theory to statements that express cognitive attitudes revealed by psychoanalytic techniques, and he suggests that what they share is that it is often possible (at least in principle) to evoke a sense of recognition in the subject of the correctness of the attribution of the belief (or other attitude), and that this recognition will be, as it were, 'from the inside' (p.176). Nagel's aim is to drive a wedge between phenomena like digestion and linguistic competence by urging crucial connections with consciousness for the latter which the former lack (although cf. Chomsky 1986, 230).

However, Nagel never clarifies why we should suppose consciousness plays a role in language acquisition or competence. The literature on formal learning theory contains numerous descriptions of algorithms that can 'learn' small fragments of natural languages. When proposed algorithms fail to converge on the correct language, the problem is not that the system implementing the algorithm lacks consciousness. To take an example, Gibson and Wexler's Trigger Learning Algorithm learns any grammar in a hypothesis space of languages defined by a few parameters, and does so simply by ('unconsciously') reacting to its own failure or success at parsing the current input string (Gibson & Wexler 1994).[7] In this literature, 'learning' is a technical term, though the aim is to model human learning. Nagel can always reply that consciousness is crucial to the actual learning of a grammar by a human. However, since there are attempts to uncover what is needed for learning not requiring consciousness, further defense of the connection between consciousness and learning is needed before any connection can differentiate linguistic competence from digestion.

2.3 Do Speakers Really Know a Grammar?

In addition to asking what justifies positing a distinctively cognitive capacity to account for linguistic competence, one might wonder whether the capacity is knowledge. As noted earlier, there is a difference between typical cases of knowing and so-called knowledge of a grammar. Various philosophers argue for psychological differences between typical beliefs and the information bearing states constitutive of 'knowledge' of grammar, and that these differences rule out the latter as beliefs (so, *a fortiori*, as knowledge as well) (Evans 1981, 131-32; Wright 1986a, 33-34, 41-43; Stich 1978). We will focus on Stich (1978).

One difference is that typical beliefs are accessible to consciousness: Attention 'suitably directed to the content of the belief' leads to 'a certain sort of conscious experience' (Stich 1978, 504). You may not be thinking about how you brush your teeth, but, if asked, you will have a conscious episode that involves reflecting (perhaps in detail) about how you do so.[8] On the other hand, if asked to articulate the semantics of 'every', or just the part that explains why 'Every plane landed together' is ill-formed while 'All the planes landed together' is not, you might not know. In fact, even if told why 'every' behaves this way, you still might not believe it (in some sense, at least).[9]

Secondly, a typical belief 'inferentially integrates' with other beliefs, but states carrying grammatical information need not. For typical beliefs, if a subject believes that p, and comes to believe that if p, then q, she will also come to believe that q. Similarly, for other common deductive and inductive inferential schemata.[10] To some extent, 'beliefs' about grammar share this property. For instance, the state of 'believing' that predicative noun phrases obey rule R may be inferentially connected to one's explicit belief that 'He is stupid and a liar' is fine, but 'He is a liar and John' is not.[11] However, though grammatical 'beliefs' may enter into inferential relations, Stich's point is that it is nonetheless severely restricted as to what kinds of inferences they can enter into. So, suppose you 'believe' predicative noun phrases obey rule R, and you also explicitly believe (perhaps because a wealthy theorist told you so) that if predicative noun phrases obey rule R, you will receive a million dollars. Despite the ingredients for a simple *modus ponens*, you do not come to believe you will receive a million dollars. You don't, Stich suggests, because grammatical 'beliefs' do not enter into inferential relations with other beliefs in the 'promiscuous' ways typical beliefs can. Similarly, most of us never feel an incompatibility between a tacit belief and an obviously contradictory conscious belief (Stich 1971, 489).

The foregoing argument challenges whether linguistic information bearing states are beliefs, and also whether it *matters* if they are. To see this, note that Stich assumes that the relevant states represent a theory of the language, and they are causally efficacious in linguistic comprehension. Whether such states are 'subdoxastic' or full-fledged beliefs depends largely on how beliefs function. So what is achieved by endorsing a theory that requires that Xs are beliefs (cf. Stich 1978, 514-515)? Are we seeking the true nature of reality or of our concepts? Are we trying to develop a useful concept for cognitive science? Whether these states are beliefs might be important to someone like Dummett, who believes that a theory of meaning must explain how language use is rational (cf. the opening pages of Dummett 1975, Dummett 1976; 1978, 104; cf. also Smith 1992, 124-31, Wright 1986b, 215-216, and Lepore 1996, 50). If linguistic competence is located primarily in subdoxastic states, perhaps we should concede that it is 'outrageous' to suppose that the type of propositional attitude speakers bear to their grammar is knowledge, in the usual sense (McGinn 1981, 290). We might instead follow Chomsky invoking the term of art 'cognize' for a sort of propositional attitude speakers bear to grammars (Chomsky 1986, 265-69).[12] (As a point of procedure, we will use the traditional 'tacit or 'implicit' knowledge, with no presumptions as to the nature of the type cognitive state it is. If you doubt such states are knowledge, treat our uses as privative adjectives, as McGinn suggests (McGinn 1981, 290).) A principal way to justify that speakers cognize grammars continues to be that assuming so better explains linguistic competence than any other hypothesis.

2.4 Dispositionalism: Two Alternatives to Cognitivism.

In this section, we will sketch two alternatives to cognitivism, what we shall call **unstructured** and **structured dispositionalism** (UD and SD, for short). We shall begin with UD. The cognitivist supposes that the hypothesis of tacit knowledge of a grammar is part of the best theory of linguistic competence, and so she posits tacit knowledge, thereby freeing herself to exploit any advantages of the hypothesis (as well as incurring its disadvantages). UD differs from cognitivism because it makes no strong claim about the relation between a grammar and a speaker. According to UD, a speaker may not tacitly know (or cognize) a grammar of her language. Its task is to construct a grammar that 'fits' (in Quine's sense) a speaker's dispositions to verbal behavior (where 'behavior' need not be understood in Quine's sense) (cf. Quine 1975).

UD is a *methodological* alternative to cognitivism, differing from it only about the scope of the project of devising a semantic theory for a natural language. Cognitivism requires a theoretical description of the semantic features of the target language that expresses the content of a representational state of the speaker which explains semantic competence. UD, on the other hand, requires a true semantic theory, but posits no psychological mechanisms. (At the other end of the spectrum is what we shall call non-cognitivism, according to which we lack tacit knowledge of a semantic theory.[13] (We will return to this position below.)

Although UD is less bold than any account that purports to specify the psychological mechanisms that underwrite linguistic competence, its modesty also buys stability: a UD theory can be correct regardless of how a physical system like the human brain realizes dispositions constitutive of linguistic competence. Questions about realization are someone else's concern, perhaps the neuroscientist's. In this sense, then, the semanticist determines (in detail) the goal of what is an empirical problem for the neuroscientist and a design problem for the AI researcher. Furthermore, this naturally divides the theoretical work in accounting for linguistic competence. A UD defender might argue that cognitivism has semanticists strongly constraining the *architecture* of psychological and perhaps even neuroscientific theories. UD, on the other hand, only has semanticists constraining the *goals* of such theories. UD requires semanticists to inform psychologists about the semantic data to be explained, while cognitivism further requires semanticists to inform psychologists how to construct a theory that accounts for the data. Of course, the UD theorist is not suggesting that tacit knowledge posited by cognitivism is wrong; the essence of UD is quietism.

If UD is the correct methodological stance, why should finiteness concern us? One might object that the finite amount of our mental storage space, computational powers, and language acquisition time are all (strictly speaking) empirical hypotheses (cf. Davidson 1965). What justifies attention to these empirical data and not others? In response, first note that the dispositionalist is devising a theory to be used by the psychologist; he is not devising the theory used by a speaker. So, though the finiteness constraint is justified by the attention span of psychologists, it is also justified by the sorts of empirical data mentioned above. Although a UD theorist is quiet about the nature of the psychogrammar (in George's sense), he needn't be completely oblivious – knowing basic finiteness facts about humans, he can try to respect this very modest empirical constraint. If other facts became as uncontroversial, they too might be incorporated into the dispositionalist's agenda. Perhaps, then, the rubric of dispositionalism houses a spectrum of theories, depending on how uncontroversial other data are.

In contrast to UD, which broadly characterizes dispositions to verbal behavior without a stance about which dispositional components comprise this larger collection (or how they do), **structured dispositionalism** (SD, for short) does take a stance. According to SD, corresponding to each axiom in a correct meaning theory is a unique disposition. Following Evans (1981), consider a finite language L, with ten proper names and ten one-place predicates, for a total of one hundred sentences. A speaker S has dispositions corresponding to a base clause (in a meaning theory for L) that says that 'a' refers to John just in case S has a disposition such that,

> (2.5) For any quote-name Φ of any predicate of L and any predicate Ψ of the metalanguage of L, if S has the disposition corresponding to a clause that says something satisfies Φ iff it is Ψ, and S hears an utterance of the form Φ^'a', S will judge the utterance true iff John is Ψ.[14]

'connectedly', Evans writes, S has a disposition corresponding to the clause that says that something satisfies 'F' iff it is bald just in case S has a disposition such that,

> (2.6) For any object x and any quote-name α of a name in L, if S has the disposition corresponding to the clause that says that α refers to x, and S hears an utterance of the form 'F'^α, S will judge the utterance true iff x is bald (Evans 1981, 124-25).

In addition to hypothesizing individuation conditions for dispositions that constitute a grammar, Evans recommends such talk to be understood in a 'full-blooded' sense: S's dispositions are states of S appropriately causally responsible for the relevant patterns of behavior. Thus, SD posits a network of possibly non-cognitive dispositions constitutive of semantic competence. If they are non-cognitive (i.e., independent of any cognitive apparatus), SD is more than a methodological alternative to cognitivism. Despite using 'tacit knowledge', this is how Evans construes SD (cf. Evans 1981, 120-121, 124, and especially 133-134). On the other hand, SD may be a mere methodological alternative to cognitivism, if one is quietist about underlying the dispositional or categorical bases.

Wright raises three problems for SD. First, it is circular about understanding names and predicates: competence with a name is given in terms of competence with predicates, but competence with a predicate is given in terms of competence with names (Wright 1986a, 39-40; 1986b, 232-233). Secondly, when axioms are replaced with their corresponding dispositions, Quine's challenge remains: any empirical data that supports ascribing a set D of linguistic dispositions corresponding to a grammar can be made to support the ascription of a distinct set

D' of linguistic dispositions corresponding to an extensionally equivalent grammar, by exploiting 'appropriate hypotheses, of a *non-semantical* sort, about the presumed causal substructure' of the two sets of dispositions' (Wright 1986a, , 1986b, 231; cf. Davies 1987, 451-453). Finally, Wright notes that in the object language under consideration, it is natural to construct a compositional meaning theory using a compositional axiom, such as Evans',

> (2.7) A sentence coupling a name with a predicate is true iff the object denoted by the name satisfies the predicate (Evans 1981, 123).

Wright then argues that a meaning theory would be 'crippled' without something like (2.7), but that SD need not postulate a disposition corresponding to (2.7). A speaker with the dispositions in (2.5) and (2.6) 'is thereby disposed to attach the proper significance to name-predicate coupling – since he is thereby disposed to attach the proper significance to sentences formed by coupling names and predicates' (Wright 1986a, 38; 1986b, 232). But now there is discordance between the details of the meaning theory and how SD says the meaning theory is 'realized'. If (2.7) is crucial to articulating a meaning theory, but its corresponding disposition is otiose in an account of linguistic competence, then the dispositions SD posits bear no simple one-one relationship to the axioms of the theory SD advertises. (For further discussion of dispositionalism and Wright's objections to SD see Davies 1987.)

2.5 Semantic Non-cognitivism and Transductionist Theories.

We began §2 with cognitivism and a battery of arguments against it. We turned to various forms of dispositionalism, which, to varying degrees, are alternatives to cognitivism. We turn now to another alternative to cognitivism, which we shall call **non-cognitivism**.

Strictly speaking, non-cognitivism is a form of dispositionalism, because it suggests that the best explanation of linguistic competence does not require cognitive relations to a semantic theory. As noted in §2.4, non-cognitivism entails that we lack tacit knowledge of a semantic theory. The standard way to support this entailment is to produce a theory which explains linguistic competence without appeal to tacit knowledge. Behind non-cognitivism is the idea that if linguistic competence can be so explained, then, assuming tacit knowledge does no theoretical work elsewhere, positing it is idle, and so, by Occam's razor, its existence should be denied. Although we will consider only one form of non-cognitivism, what we shall call **transductionism**, other types are available, such as those developed or suggested within a connectionist paradigm (cf. Elman, Bates, *et al* 1996, Rumelhart, McClelland *et al.* 1986, Langacker 1990).

To render knowledge of a semantic theory unnecessary, it suffices to show how competent users of a natural language could plausibly engage in the kinds of (linguistic and mental) activities they do without recourse to tacit semantic knowledge. Fodor articulates such a view in *The Language of Thought*, and still endorses its relevant parts (Fodor 1975; 1990b; 1998; cf. also Schiffer 1987). We begin by sketching his position, and then turn to its criticism.

Transductionism. The main tenet of transductionism is that mental processing has the form of operations based on nomic properties of certain possibly complex mental objects. To be more precise, mental processing takes place because of operations on the syntactic features of expressions in a language of thought (LOT).[15] On this view, the primary explanandum concerning natural language is how we communicate. According to transductionism, communication is the process whereby a sentence in a speaker's LOT, called a 'message' (Fodor 1975, 106), is mapped onto a phonetic string of English (say), which when produced in the vicinity of a hearer is in turn mapped onto (another token of) the message the speaker wished to communicate in the hearer's LOT. Other aspects of the hearer's processing algorithm (cf. §1.1) function to produce a belief about what the speaker said (i.e., a belief whose content is something like 'x said that P'). Successful communication lies in whether speaker and hearer share sufficiently similar transducing mechanisms between messages and heard strings (cf. p. 103). This is where transductionism becomes 'Gricean in spirit': expressions of a natural language like English acquire meaning in virtue of interpersonal similarities concerning the range of phonetic strings that can be used to communicate a given message (p. 104). However, this does not mean linguistics plays no role: a generative grammar for a natural language specifies for each message, 'the descriptions (morphological, phonological, syntactic, etc.) that a token [heard string] must satisfy if it is to conform to the linguistic conventions' for that natural language (p.109). Thus, one need not know (even though one surely does) that 'the dog' denotes the dog to be competent in English; one need only share with other speakers 'a knowledge of the descriptions that a written form must satisfy if it is to serve to communicate references to the dog to people who belong to that community' (p. 105).

According to Fodor, then, linguistic competence consists in an ability to map expressions of English onto correct expressions of one's LOT, and *vice-versa*, where correctness is a matter of conformity to the conventions of the community. Most interesting questions, such as 'What constitutes competence with respect to LOT?' and 'How do LOT expressions get their semantics?' are for the philosophy of mind and metaphysics (not epistemology and linguistics). Tokens of LOT get their meanings however they do, and have whatever meanings they have. Linguistic competence is just an ability to transduce objects of one sort (phonologically individuated strings) into objects of another sort (tokens of LOT). Fodor acknowledges this when he writes 'English *has no semantics*' (Fodor 1998, 9), other than whatever it inherits from the semantics of LOT. (A similar semantics-free view of linguistic competence is championed by, among others, Chomsky and Hornstein (Chomsky 1986, 1995, Hornstein 1984, 1988, 1989, 1991).

Fodor's view has been challenged by, among others, Lepore (1996). Lepore argues for epistemic consequences of linguistic competence that transductionism fails to explain. His point is that transductionism challenges the need to ascribe semantic knowledge by arguing that linguistic competence is constituted by a transduction relation between English and LOT. If someone hears you utter 'It's raining', she will reliably come to believe you said it's raining, because the transduction process from English to LOT is reliable, as well as are the other 'algorithmic' processes (cf. §1.1) needed to generate her belief. Lepore argues that if this were all there is to belief acquisition about what others say, transductionism would provide no account of one's own *reasons* for these beliefs. On the one hand, beliefs about what is said may be justified, at least on an externalist theory of justification, of the sort associated with, e.g., Goldman (1986). But on the other hand, transductionism provides no reason for why the interpreter acquires the particular belief she does about what you said. Compatible with transductionism, a speaker might be utterly 'clueless' as to how she acquired the belief that you said it's raining when you uttered to her 'it's raining', and she might also be clueless as to whether this belief is justified (Lepore 1996, 52). Following Davidson, Lepore suggests that 'nothing can count as a reason for holding a belief [about what's said] except another belief [about what the words uttered mean]' (p. 53; cf. Davidson 1986, 123), and that 'a belief that p (partly) rationalizes a belief that q only if the belief that p is (partly) causally responsible for the belief that q' (p. 53).

Although Lepore's argument is directed against transductionism, it also challenges various forms of dispositionalism. If the dispositions that constitute linguistic competence are non-cognitive, then although partly causally responsible for someone coming to have a belief about what another said, they cannot provide reason for one's having those beliefs.

III DOES KNOWING A GRAMMAR EXPLAIN LINGUISTIC COMPETENCE

3.0 Is Modesty Enough?

In this section, we will contrast modest and full-blooded meaning theories, and then review some objections to modest theories.

A **modest** meaning theory for a language L associates concepts with words and issues in meaning assignments to every sentence of L (cf. Dummett 1975, 102, 127; McDowell 1987, 72-73; Dummett 1987, 263-264; McDowell 1997, 119-120). Any theory that aims solely to derive theorems of forms (M) or (T) for every sentence S of L is modest,

(M) S in L means that p
(T) S is L is true iff p

where 'p' specifies the meaning of 'S'.

Dummett favors full-blooded theories over modest ones. The former not only associate words with concepts, but explain 'what it is to have the concepts expressible by means of that language' (Dummett 1975, 101). Where a modest theory might tell us only that something satisfies 'red' iff it is red, a full-blooded one 'explains...to someone who does not already have the concept' red what grasping the concept of red is. For more on modesty and full-bloodedness, see Dummett 1975, 102; McDowell 1987, 62; 1997, 105-106.

Why would anyone want more than modesty? Harman answers as follows,

> [W]e might know that the sentence 'All mimsy were the borogroves' is true if and only if all mimsy were the borogroves. However, in knowing this we would not know the first thing about the meaning of the sentence,
>
> (3) 'All mimsy were the borogroves' (Harman 1974, 6; our numbering; cf. also, Dummett 1975).

Knowing (3) is insufficient for understanding 'All mimsy were the borogroves' unless one already understands or has the concepts expressed by 'mimsy' and 'borogroves' (cf. Block 1986, 110, for a related argument). The theories under attack by Harman are modest theories, even though he couches his objection in terms of truth theories. Harman's objection (the **mimsy argument,** for short) is driven by an assumption that he and Dummett share, namely,

(D) A theory of meaning for a language L is a theory of understanding for L (Dummett 1975, 99).

Dummett emphasizes (D) (Dummett 1975, 99, 100-101; 1976, 69ff; cf. Smith 1992), 112. The role (D) plays in the mimsy argument is evident in its (schematic) reconstruction,

(M1) A meaning theory for L must explain understanding sentences of L [from D].
(M2) Modest theories do not explain understanding sentences of L.
(M3) Nothing else about such theories (e.g., how they were constructed or justified) explains this understanding.
(M4) ∴ [by 1,2, 3] Modest meaning theories are defective.

3.1 A Standard Reply.

Dummett and Harman both anticipate a reply to (D) and (M1)-(M4) that denies (M2) (Dummett 1975, 114; Harman 1974, 6). Modest theories explain understanding, because they are couched in a metalanguage the speaker understands (or at least she already has the concepts expressible in this metalanguage). So, a speaker's grammar will generate an interpretation of (3) only if her grammar has the axiom that something satisfies 'mimsy' iff it is mimsy. But a grammar with this axiom requires the speaker already to understand (or have the concept expressed by) the word 'mimsy'. Since a speaker's grammar can interpret (3) only if she already understands, or has the concept expressible by, 'mimsy', (M2) is false, and the mimsy argument is unsound.

Anticipating some such reply, both Harman and Dummett rebut that assuming prior understanding or conceptual grasp puts modest meaning theories on a par with translation manuals. A translation manual consists 'in the statement of an effective method for going from an arbitrary sentence of the alien tongue to a sentence of a familiar language' (Davidson 1973, 129). Translation theories *qua* of theories of understanding have been criticized on the grounds that one can know a translation of every sentence of one language into another language without understanding any sentence in the former, and so, without understanding what any sentence means (Lewis 1970, 18-19; Davidson 1973). Imagine a manual in English that translates Greek into Latin. It will contain items like ''ανθρωπos' translates into 'homo'', ''ιππos' translates into 'equus'', ''κλεπτω' translates into 'claudo''. One could use this manual to interpret Greek only if one already understood Latin (and English). Similarly, urge Harman and Dummett, any modest theorist must be presupposing that a speaker already understands (or has the concepts expressible in) the language in which the theory is specified. Dummett and Harman rebut that were this presupposition correct, a translationist could make it as well. When explaining linguistic competence, presuming a translation manual can explain linguistic competence is incorrect.

3.2 Higginbotham's Reply.

Higginbotham 1989b, p. 165 contests (D) by arguing that a semantics for a language and a speaker's understanding of it can come apart. Consider Putnam's speaker who cannot distinguish elms from beeches. This speaker might fully understand his language, but his language might induce only a partial interpretation of 'elm' and 'beech'. So, the speaker fully understands 'beech' and 'elm' in his idiolect, but what they mean is not what they mean in English, since in English their extensions differ (though cf. Burge 1979). Higginbotham suggests this is not how we think of reference.

> Our words do refer to certain things...even when our knowledge of reference is incomplete. Moreover, it appears that incomplete understanding does not even prevent attribution of the same *concept* to the ignorant as to the learned. As we learn, we seem to come to know, or to know more fully, what things we refer to and through what concepts we refer to them (p.155).

Thus, he recommends we consider the language fully interpreted, with a speaker having only a partial grasp. If he is right, it is unclear whether a semantic theory ought to account for what one knows when one understands language, particularly if understanding a language despite is compatible with said deficiencies with respect to 'elm' and 'beech'.[17]

So for Dummett a theory for L is correct only if its meaning theorems *explain* understanding L, whereas for Higginbotham it might be correct even without any such explanation (which he doubts it can (Higginbotham 1989b, 166)). Nonetheless, for Higginbotham knowledge of such a theory could constitute *partial* linguistic competence. Invoking partial constitution is supposed to support Higginbotham since it explains how speakers can use expressions they only partly understand: according to Higginbotham, one can know what 'x carried out a leveraged buyout of y' means and not know what leveraged buyouts are. No such explanation is available to Dummett, since he demands a meaning theorem to explain one's having knowledge of a homophonic meaning theorem. According to Higginbotham, then, there cannot be a fully explicit full-blooded theory of the sort Dummett envisages because our words have fixed meanings (your use of 'beech' doesn't have elms in its extension even if you cannot distinguish elms from beeches) despite our lacking the appropriate understanding constituitive of a full-blooded meaning theory. To see why, it will be useful to discuss Dummett's attack on modest theories (1975, 105-108).

Dummett observes one can know 'the Earth moves' is true without knowing that the Earth moves. He calls the latter *knowledge of the proposition expressed.* With truth-conditional meaning theories, the goal is to explain knowledge of the proposition expressed by (3.1),

(3.1) 'The Earth moves' is true iff the Earth moves.

Knowledge of (3.1) is not disquotational, such as (3.2),

(3.2) "The Earth moves' is true iff the Earth moves' is true.[18]

What must one know to know the proposition expressed by (3.1)? Dummett suggests one must know the meanings of its used component words. In a truth-conditional framework, this means knowledge of base axioms. Hence, by a similar line of argument, something besides knowledge of the truth of the axioms is required for knowledge of the propositions expressed by axioms. What could this something else be? If we suppose it to be knowing the truth of the axioms used in a derivation of (3.1), then we have started a regress. The additional knowledge must be of a different sort if it is to explain knowing the proposition expressed by (3.1). This final claim is the primary argumentative engine driving him to the conclusion that meaning theories must be full-blooded.

A crucial aspect of Dummett's position is that knowing the proposition expressed by a meaning theorem depends on knowing the propositions expressed by the axioms from which it is derived. So failure to understand a term like 'beech' amounts to failure to know which proposition is expressed by 'x satisfies 'beech' iff x is a beech'. However, *prima facie*, speakers have varying degrees of knowledge of the meanings of expressions; furthermore, over time, they may acquire increased degrees of knowledge of these meanings. So any account of partial knowledge that Dummett offers must account for these phenomena too. Either Dummett can argue against treating imperfect speakers as partially knowing a fully interpreted language or he can account for partial understanding. The latter must show how full understanding can be achieved, and be consistent with a theory of meaning being a theory of understanding. (If one only partially understands 'beech', what effect does this have on its meaning?) If he accepts partial knowledge of our language and adopts the latter approach, then the account still must be simpler than Higginbotham's, since Higginbotham provides a simple explanation of the phenomenon. Thus, stories involving complex structures of related propositions (e.g., structures that relate a proposition that amounts to total knowledge of a word to propositions that amount to partial knowledge – which themselves may have to be interrelated) will not work.

3.3 McDowell's Reply.

McDowell offers two arguments against Dummett (McDowell 1987). One supports modest theories directly, and the other indirectly by, in effect, assaulting the mimsy argument.

McDowell defends modest theories as such,

(McD1) Meaning theories are modest or full-blooded.
(McD2) They cannot be full-blooded.
(McD3) ∴ They must be modest.

His second argument is only a bit more complicated,

(McD4) An explanatory meaning theory (in Dummett's sense) must be full-blooded.
(McD2) Meaning theories cannot be full-blooded.
(McD5) ∴ No meaning theory is explanatory (in Dummett's sense).
(McD6) There can be a correct meaning theory.
(McD7) ∴ [denial of (M1)] Meaning theories need not explain our understanding (in Dummett's sense) of an object language.

Which feature of the mimsy argument McDowell's second argument challenges depends on what counts as explanation. If explanations are Dummettian, the second argument attacks either (D) (i.e., that a theory of meaning is a theory of understanding) or the inference from (D) to (M1), depending on how one understands 'understanding', an issue we discuss below.

Turning to (McD2), why reject full-blooded meaning theories? Dummett replies that though modest theories pair expressions with concepts, by failing to explain concept possession they fail to explain linguistic understanding (Dummett 1987, 258-60; McDowell 1997, 111-12). For McDowell this dilemma is false: the issue is not about explaining concept possession, but whether we can do so and still respect the constraint that a meaning theory be full-blooded. McDowell argues that one feature of this constraint concerns the sort of explanation of linguistic competence that is required by a full-blooded theory (McDowell 1987, 61). Full-bloodedness requires explaining what it is to possess concepts associated with words. So suppose we have a full-blooded theory for some language L. Understanding this theory must suffice for one previously unacquainted with L to come to understand L (Dummett 1975, 103-104; 1987, 265-266). But that a full-blooded theory must be in language, McDowell's argument runs, creates problems. First, by the response to the standard objection (in §3.1), the theory is on a par with translationist theories. Secondly, since the current move requires us to explain how one understands another language, progress on the task of explaining understanding a language is nil. Thus, McDowell seems to be using a version of the mimsy argument, one that attacks full-blooded theories and their demands on the explanatory work of such a theory. However, endorsing any such argument does not prevent McDowell from attacking the other form of the mimsy argument, one which attacks modest theories. We will discuss this below.

McDowell responds that 'a proper theory of meaning for a language would be formulated "as from outside" content altogether' (McDowell 1987, 61). This requires that a full-blooded meaning theory not use expressions which specify or presuppose a specification of the contents (of words, expressions, utterances, thoughts, etc.).[19] Although this is opaque, it appears that one has specified the content of an expression as from outside content altogether, if the specification does not include a use of an intensional context. This restriction prevents an explanation of possessing the concept 'square' along the lines: One has the concept *square* iff one is disposed to believe of all and only square things one encounters *that they are square.*[20]

Thus, a full-blooded meaning theory must

(i) explain, what it is to have concepts denoted by expressions in the language, and

(ii) it must accomplish using a vocabulary that does not specify the contents of words, utterances, thoughts, etc.

But McDowell (and almost everyone else) also rejects behaviorism, so a theory that purports to explain concept possession in terms of 'outward behavior' is untenable (McDowell 1987, 65). This entails a further constraint on full-blooded meaning theories:

(iii) the theory cannot be behavioristic (McDowell 1987, 63-65).

McDowell doubts any theory can satisfy (i)-(iii).[21] One might try by ascribing tacit knowledge of a meaning theory (that avoids behaviorism) where such knowledge 'shows itself partly by manifestation of the practical ability, and partly by a willingness to acknowledge as correct a formulation of what is known when it is presented' (Dummett 1978, 96). [22] However, McDowell notes that any such appeal guarantees that the meaning theory will be indeterminate: when all possible data are in, with every other relevant theory as precisely determined as can be, extensionally non-equivalent theories equally compatible with the data still exist (McDowell 1987, 66-67; 1997, 112-115). (Cf. George 1986 for the differences between underdetermination and indeterminacy.)[23] That is, no matter how much empirical data we have concerning e.g. the meaning of 'square', it can be accommodated equally well by theories according to which 'square' does not mean *square.* For suppose we hypothesize that the meaning of 'square' is *square,* because speakers of the object language are disposed to call only squares 'square'. This evidence is equally well explained by the hypothesis that speakers are disposed to call squares or pieces of mud from the bottom of the ocean 'squares'. Even if there *is* evidence that they are not so disposed, other Goodmanesque hypotheses compatible with the data will always be available (e.g., the disposition to use 'square' to pick out squares or numerals more than 1,000 digits long). (Cf. Goodman 1954, chapter 3.)

Regardless of how much evidence is available for positing tacit knowledge, it will be finite, and so infinitely many extensionally non-equivalent grammars that account for the data equally well will exist. (The same result holds even if there were (*per impossible*) an infinite amount of data.)[24] The *locus classicus* for problems of indeterminacy is Quine 1960.

To sum up: McDowell's arguments for modesty rely on (McD2). A full-blooded theory explains linguistic competence only if one can learn it without already understanding a language. This suggests that a full-blooded theory can be given "as from outside' content' altogether, thus rendering full-blooded theories behavioristic. Finally, invoking tacit knowledge is no help, for to do so renders the theory unacceptably indeterminate. So, if meaning theories must be modest or full-blooded (a premise that aches to be clarified and challenged), they must be modest.

Furthermore, the mimsy argument fails, because if the explanatory task of a meaning theory is as Dummett says, then either no meaning theory is correct or the inference to (M1) is unsound. Since the former is implausible (though adopted in Schiffer 1987), the second must be adopted, which entails the unsoundness of the mimsy argument. On the other hand, perhaps one need not demand as much as Dummett about what suffices for explanation in (M1)-(M3). It may be that arguing to (M1) is legitimate, but one's alternative conception of an explanation is such that the justification for (M2) is thereby undermined. This seems to be McDowell's negative position regarding full-blooded theories. We will not discuss his positive view, but he does argue that the theorems of a modest meaning theory suffice to explain linguistic competence (and do so without incurring the indeterminacy of a theory that posits tacit knowledge of a full-blooded theory) (McDowell 1987, 67-70, 73-76; 1997, 116-119).

In conclusion, the diversity and difficulty of the replies we have reviewed show that mimsy argument to be anything but simple. It combines independently problematic issues including disquotational theories of truth, theories of truth as theories of meaning, lexical semantics, the structure and possession conditions of concepts, the nature of explanation, and the interface between one's psychogrammar and one's other capacities for the rational use of language. These issues are more fundamental than the mimsy argument because it can be understood only when these other issues are better understood.

IV SEMANTICS AND LINGUISTIC COMPETENCE

In this paper, we have discussed major issues concerning semantic competence. However, space prevents treating every relevant issue. We will conclude by merely mentioning three issues a longer paper on knowledge and semantic competence should discuss. (i) The concept of tacit knowledge was central in §2. A variety of analyses of this concept, and of the related concept of tacit belief, are in Lycan 1986, Dennett 1987, Kirsh 1990 and Crimmins 1992. (ii) The 'Kripkenstein' problem about whether past evidence can determine that we are currently following a rule (of grammar, for instance), and more specifically, whether there can be any fact of the matter about what we mean by our words. This problem first appeared in Kripke 1982, and a good overview of the problem can be found in Loar (1985). (iii) Quine's 'indeterminacy of translation' is often explained in terms of a speaker's ability to translate utterances from another language, though the translated language may be taken to be the translating language. In this latter situation, the problem purportedly shows that no single correct translation manual (or set of extensionally equivalent translation manuals) exists. The problem originates with Quine 1960; further discussion is in Root 1976 and Lepore 1977.[25]

Kent Johnson and Ernie Lepore
Center for Cognitive Science
Rutgers University

NOTES

[1] For an argument that the distribution of negative polarity items cannot be characterized syntactically and must be characterized semantically see Ladusaw 1980 (cf. Higginbotham 1995a, 5-7).

[2] Cowper 1992, p. 2 reports Keyser as saying, "We are trying to figure out what it is that people *act as if* they know".

[3] Quine suggests the relevant form of guidance is 'an intermediate condition, between mere fitting and full guidance in my flat-footed sense...' (Quine 1972, 442). Whether he's right is irrelevant here.

[4] In the terminology of George 1989b, Quine's challenge is, 'What evidence selects one theory of a psychogrammar over another?'

[5] In addition, Quine writes, '...the new doctrine of the grammarian's added burden raises the problem of evidence whereby to decide, or conjecture, which of two extensionally equivalent systems of rules has been implicitly guiding the native's verbal behavior' (pp. 443-44); 'The problem of evidence for a linguistic universal is insufficiently appreciated' (p. 446); 'The enigmatic doctrine under consideration says that one of these analyses is right, and the other wrong, by tacit consensus of native speakers. How do we find out which is right?' (p. 448).

[6] It is not clear Evans intended to defend cognitivism. Nonetheless, his remarks may be so construed. The details of his position are taken up below.

[7] See also Niyogi & Berwick 1996.

[8] Stich's claim about typical beliefs' principled accessibility to consciousness is about what would (likely) happen were the subject and her situation normal. Unconscious beliefs of psychoanalytic theory do not count, because the antecedent is not satisfied, inasmuch as (we may suppose) some psychological mechanism interferes with ordinary processes leading from a belief to conscious awareness of it (Stich 1978, 505).

[9] It may be that nobody knows why 'every' and 'all' distribute as they do. The example is from Christine Brisson's dissertation, 'Some Wider Consequences of Narrow Scope' (Linguistics, Rutgers University, 1998).

[10] Evans agrees that inferential integration is *constitutive* of belief, 'To have a belief requires one to appreciate its location in a network of beliefs' (Evans (1981), p.132). He also ascribes it to Wittgenstein 1969, §141.

[11] Although Evans denies this point (1981, 133). It is hard to see how our linguistic competence could be explained by appeal to information bearing states that cannot interact with one another (assuming a relatively simplistic theory of individuation of the relevant

information bearing states) and, more importantly, could not produce further explicit or implicit beliefs. This would render tacit beliefs unable to explain, e.g., a speaker's coming to believe explicitly what a particular utterance means.

[12] We will remain silent about the relation between knowing and cognizing; Chomsky himself vacillates on the extent to which cognizing and knowing overlap; however, he is consistent about the unimportance of overlap for explaining linguistic competence (Chomsky (1980), Chomsky (1986), pp. 265-69).

[13] It would be an interesting project to compare the notion of dispositionalism (and perhaps even some versions of non-cognitivism) with the notion of "knowledge how". Doing this would require developing a clear account of the cognitive structure of the latter notion, which would take us too far afield from the present project.

[14] '^' means 'concatenated with'.

[15] A footnote of Fodor's on the syntax of LOT is relevant here, '*Any* nomic property of symbol tokens...any property in virtue of the possession of which they satisfy causal laws...would, in principle, do just as well. (So, for example, syntactic structure could be realized by relations among electromagnetic states rather than relations among shapes; as, indeed, it is in real computers.)' (Fodor (1987), p. 156, fn. 5)

[16] Lepore and Loewer (1981) respond to Harman by arguing that one no more needs to understand the metalanguage in which (A) is written to know what (A) expresses than Galileo needed to know English for him to have believed that the earth moves.

(A) 'La terra si muove' is true in Italian iff the earth moves.

They do not disagree with Dummett, however, that knowing (A) requires concepts of the earth and movement. Whether this excludes modest theories as theories of linguistic competence is a topic for the rest of this section.

[17] There may be a way to reconcile Dummett and Higginbotham, because there are several ways to understand crucial terms both in (D) and in the argument in which (D) is employed (cf., Smith 1992 for extensive discussion of Davidsonian, Dummettian, and Chomskian interpretations of (D)). One might suppose Dummett has something special in mind by a theory of understanding: 'once we can say what it is for someone to know a language, in the sense of knowing the meanings of all expressions of the language, then we have essentially solved every problem that can arise concerning meaning' (Dummett 1975, 133).

[18] This point is not unique to Dummett. Cf., Chomsky 1986, 266, Fodor 1968, 633-34.

[19] The formulation in the text preempts appeal to contents in explaining concept possession. However, a weaker restriction is available: for any name ϕ of any expression of the object language, one cannot use ϕ in a content clause (i.e., in an intensional context) in an account of the possession of the concept denoted by ϕ. This permits using other kinds of content clauses in accounting for possessing the concept denoted by ϕ. Further restrictions on this second proposal are needed; how does Dummett avoid psychologism if possessing the concept denoted by ϕ is explained *via* a content clause containing a use of ψ, the possession of which is explained *via* a content clause containing a use of ϕ? One might restrict the expressions that can occur in the content clause(s) that explain the possession conditions of the concept denoted by ϕ to those expressions taken to denote innate concepts, or to those expressions that have the possession conditions of their concepts explained 'earlier' in some recursively described hierarchy. It is not clear Dummett would take the first option, given his reluctance to develop his theory so that it becomes more than minimally answerable to empirical psychological hypotheses. However, if one had an acceptable means for defending some class of expressions as usable in content clauses in explaining concept possession, it might help with difficult cases, such as explaining theoretical concepts.

[20] The restriction does not prohibit using 'square' in accounting for possessing the concept *square*; cf. McDowell 1987, 62.

[21] The core of McDowell's reply is that meaning theories cannot have properties (i)-(iii). But his attack is primarily about the compatibility of (ii) and (iii); he argues that any theory formulated 'as from outside' content must be behavioristic. If behaviorism is unacceptable, theories cannot be specified 'as from outside' content. But this undermines full-blooded theories only if they must be formulated 'as from outside' content, which has yet to be established. Thus, endorsing McDowell's argument does not require rejecting full-blooded meaning theories.

[22] Two points are relevant here. First, the second part of Dummett's claim about how tacit knowledge might be partly manifested is false: if we have tacit knowledge of a meaning theory, there are many principles of this theory we are unlikely to acknowledge as correct when presented with their correct formulation. (For further discussion and examples, see §2.0.) Second, appeal to tacit knowledge of the present sort places an additional constraint on the formulation of meaning theories: where C is any concept expressed by an expression of the object language, explaining what it is to possess C must not use an expression that expresses C. Since the current suggestion is that one has a kind of knowledge of the theory, violating this restriction amounts to explaining possession of C by appeal to an epistemic state one has only if one already has C (McDowell 1987, 66).

[23] Is a theory's vulnerability to indeterminacy much of a criticism, since every theory suffers as such? McDowell believes his view is immune from indeterminacy because content is 'present in the words... [T]he thought (say) that some table-tops are square can be heard or seen in the words 'Some table-tops are square', by people who would be able to put their own minds into those words if they had occasion to do so' (McDowell 1987, 69).

[24] George glosses indeterminacy as follows: 'Where there is slack between observation and theory we have underdetermination, but slippage between total theory (all facts, known or unknown) and theory is indeterminacy. If any choice among the many present or future, explanatorily adequate, underdetermined theories of the world would leave unsettled the truth or falsity of linguistics' claims, then we cannot make sense of there being objectively correct evaluations of these' (George 1986, 489).

[25] Special thanks to Matti Sintonen and Barry Smith for their generous comments on earlier drafts of this paper. Ned Block (1986), 'Advertisement for a Semantics for Psychology', in Stich and Warfield (1994).

REFERENCES

Burge, T.: 1979, 'Individualism and the Mental", in P. French, T. Uehling, and H. Wettstein (eds.) *Midwest Studies in Philosophy*, vol. IV, University of Minnesota Press, Minneapolis, pp. 73-121.

Brandon, R. N. & N. Hornstein: 1986, 'From Icons to Symbols: Some Speculations on the Origins of Language', *Biology and Philosophy* **1**, 169-189.

Cappelen, H. and E. Lepore: 1997, 'On an Alleged Connection Between Indirect Speech and the Theory of Meaning', *Mind and Language*, **12**, 278-296.

Chomsky, N.: 1959, 'A Review of B. F. Skinner's *Verbal Behavior*, *Language*', in Block (ed.), *Readings in Philosophy of Psychology*, Vol. 1, HUP, Cambridge, 1980, pp.48-63.

Chomsky, N.: 1965, *Aspects of the Theory of Syntax*, MIT Press, Cambridge, MA.

Chomsky, N.: 1986, *Knowledge of Language*, Praeger, Westport, Conn.

Chomsky, N.: 1995, 'Language and Nature', *Mind*, **104**, 1-61.

Cowper, E. A.: 1992, *A Concise Introduction to Syntactic Theory*, University of Chicago Press, Chicago.

Crimmins, M.: 1992, 'Tacitness and Virtual Beliefs', *Mind and Language* **7**, 240-263.

Culicover, P. W.: 1997, *Principles and Parameters*, OUP, Oxford.

Davidson, D.: 1965, 'Theories of Meaning and Learnable Languages', in Davidson, 1984, pp. 3-15.

Davidson, D: 1970, 'Semantics for Natural Languages', in Davidson, 1984, pp. 55-64.

Davidson, D.: 1984, *Inquiries into Truth and Interpretation*, Clarendon, Oxford.

Davidson, D.: 1986, 'A Coherence Theory of Truth and Interpretation', in Lepore (ed.), *Truth and Interpretation*, Blackwell's, Oxford.

Davies, M.: 1987, 'Tacit Knowledge and Semantic Theory: Can a Five per cent Difference Matter?', *Mind* **96**, 441-62.

Dennett, D.: 1978, *Brainstorms*, MIT Press, Cambridge, MA.

Dennett, D.: 1987, *The Intentional Stance*, MIT Press, Cambridge, MA.

Dummett, M.: 1975, 'What is a Theory of Meaning?', S. Guttenplan (ed.), *Mind and Language,* 1974, OUP, Oxford, reprinted in P. Ludlow (ed.), *Readings in the Philosophy of Language*, MIT Press Cambridge, MA, 1997, pp. 129-155.

Dummett, M.: 1976, 'What is a Theory of Meaning? (II)', in G. Evans and J. McDowell (eds.), *Truth and Meaning: Essays in Semantics*, Clarendon, Oxford, pp. 67-137.

Elman, J., E. Bates *et al.*: 1996, *Rethinking Innateness*, MIT Press, Cambridge, MA.

Fodor, J.A.: 1968, 'The Appeal to Tacit Knowledge in Psychological Explanation', *The Journal of Philosophy* **65**, 627-640.

Fodor, J. A.: 1975, *The Language of Thought*, Harvard University Press, Cambridge, MA.

Fodor, J. A.: 1987, *Psychosemantics*, MIT Press, Cambridge, MA.

Fodor, J. A.: 1990a, *A Theory of Content and Other Essays*, MIT Press, Cambridge, MA.

Fodor, J. A.: 1990b, 'Review of Stephen Schiffer's *Remnants of Meaning*', in Fodor, 1990a.

Fodor, J.A.: 1990c, 'A Theory of Content, II: The Theory', in Fodor, 1990a.

Fodor, J. A.: 1998, *Concepts*, Clarendon, Oxford.

Fodor, J. A. and E. Lepore: 1998, 'The Emptiness of the Lexicon', *Linguistic Inquiry*, **29**, 269-288.

George, A.: 1986, 'Whence and Whither the Debate Between Quine and Chomsky?', *The Journal of Philosophy*, 489-99.

George, A. (ed.): 1989a, *Reflections on Chomsky*, Blackwell, Oxford.

George, A.: 1989b, 'How Not to Become Confused about Linguistics', in George, 1989a, pp. 90-110.

Gibson, E. and K. Wexler (1994), 'Triggers', *Linguistic Inquiry*, **25**, 407-54.

Goldman, A.: 1986, *Epistemology and Cognition*, Harvard University Press, Cambridge, MA.

Goodman, N.: 1954, *Fact, Fiction, and Forecast*, Harvard University Press, Cambridge, MA.

Graves, C. J., J. Katz, Y. Nishiyama, S. Soames., R. Stecker, and P. Tovey (1973), 'Tacit Knowledge', *The Journal of Philosophy*, **70**, 318-330.

Harman, G.: 1974, 'Meaning and Semantics', in M. K. Munitz and P. K. Unger (eds.), *Semantics and Philosophy*, New York University Press, New York, 1-16.

Higginbotham, J.: 1983, 'The Logic of Perceptual Reports: An Extensional Alternative to Situation Semantics', *The Journal of Philosophy*, 100-27.

Higginbotham, J.: 1985, 'On Semantics', in Lepore (ed.), *New Directions in Semantics*, Academic Press, London, 1987, pp.1-54.

Higginbotham, J.: 1986, 'Linguistic Theory and Davidson's Program in Semantics', in E. Lepore (ed.), *Truth and Interpretation*, Blackwell, Oxford's, pp. 29-48.

Higginbotham, J.: 1989, 'Elucidations of Meaning', *Linguistics and Philosophy* **12**, 465-517.

Higginbotham, J.: 1989b, 'Knowledge of Reference', in George, 1989a, pp. 153-174.

Higginbotham, J.: 1993, 'Grammatical Form and Logical Form', in J. E. Tomberlin (ed.), *Language and Logic*, Atascadero, CA: Ridgeview, pp. 173-196.

Higginbotham, J.: 1994, 'Priorities in the Philosophy of Thought', *Proceedings of the Aristotelian Society*, Supp. Vol. **20**, 85-106.

Higginbotham, J.: 1995a, 'Sense and Syntax', *An Inaugural Lecture delivered before the University of Oxford*, Clarendon, Oxford.

Higginbotham, J.: 1995b, 'Some Philosophy of Language', in L. Gleitman and M. Liberman (eds.), *Language: Invitation to Cognitive Science* (2nd ed.), MIT Press, Cambridge, MA., 1995.

Hintikka, J.: 1980, 'Theories of Truth and Learnable Languages', in S. Kanger and S. Ohman (eds.), *Philosophy and Grammar*, D. Reidel, Dordrecht, Holland:, pp. 37-57.

Holtzman, S. & C. Leich (eds.): (1981), *Wittgenstein: to Follow a Rule*, RKP, London.

Hornstein, N.: 1984, *Logic as Grammar*, MIT Press, Cambridge, MA..

Hornstein, N.: 1988, 'The Heartbreak of Semantics', *Mind and Language*, **3**, 9-27.

Hornstein, N.: 1989, 'Meaning and the Mental: the Problem of Semantics after Chomsky', in George, 1989a, pp.23-40.

Hornstein, N.: 1991, 'Grammar, Meaning, and Indeterminacy', in Kasher, 1991, pp. 104-121.

Kasher, A.: 1991, *The Chomskyan Turn*, Blackwell, Oxford.

Kirsh, D.: 1990, 'When Is Information Explicitly Represented?', in P. Hanson (ed.), *Information, Language, and Cognition*, University. of British Columbia Press, Vancouver, pp. 340-365.

Kripke, S. A.: 1982, *Wittgenstein on Rules and Private Language*, Harvard University Press, Cambridge, MA.

Ladusaw, W.: 1980, 'On the Notion Affective in the Analysis of Negative-Polarity Items', *Journal of Linguistic Research*, **1**.

Langacker, R. W.: 1990, *Concept, Image and Symbol*, Moutin de Gruyter, Berlin.

Larson, R. and G. Segal: 1995, *Knowledge of Meaning*, MIT Press, Cambridge, MA..

Lepore, E.: 1977, 'Reply to Professor Root", *Philosophical Studies*, **32**, 211-215.

Lepore, E.: 1996, 'Conditions on Understanding Language', *Proceedings of the Aristotelian Society*, 41-60.

Lepore, E. and B. Loewer: 1981, 'Translational Semantics', *Synthese*, **48**, 121-133.

Loar, B.: 1985, 'Critical Review of Saul Kripke, *Wittgenstein on Rules and Private Language*', *Nous* **19**, 273-280.

Lycan, W.: 1986, 'Tacit Belief', in R.J. Bogdan (ed.), *Belief*, Clarendon, Oxford, pp. 61-82.

Niyogi, P. & R. C. Berwick (1996), 'A Language Learning Model for Finite Parameter Spaces', *Cognition* **61**, 161-193.

Peacocke, C.: 1994, 'Content, Computation, and Externalism', *Mind and Language*, **9**, 303-335.

Quine, W.V.: 1960, *Word and Object*, MIT Press, Cambridge, MA..

Quine, W.V.: 1972, 'Methodological Reflections on Current Linguistic Theory', in Davidson & Harman (eds.), *Semantics of Natural Language*, D. Reidel, Dordrecht, pp. 442-54.

Root, M.: 1976, 'Speaker Intuitions", *Philosophical Studies*, **29**, 221-234.

Rumelhart, D., J. McClelland, *et al.*: 1986, *Parallel Distributed Processing, vol. 1*, MIT Press, Cambridge, MA.

Schiffer, S.: 1987, *Remnants of Meaning*, MIT Press, Cambridge, MA..

Segal, G.: 1989, 'Seeing What Is Not There', *Philosophical Review*, **XCVIII**, 189-214.

Segal, G.: 1991, 'Defence of a Reasonable Individualism', *Mind* **C**, 485-494.

Segal, G.: 1994, 'Priorities in the Philosophy of Thought', *Proceedings of the Aristotelian Society*, Supp. Vol. **20**.

Smith, B.: 1992, 'Understanding Language', *Proceedings of the Aristotelian Society,* **XCII**, 109-141.

Smith, B.: 1998, 'On Knowing One's Own Language', in C. Wright, B. Smith, and C. Macdonald (eds.), *Knowing Our Own Minds*, Clarendon Press, Oxford:, pp.391-428.

Stich, S.: 1971, 'What Every Speaker Knows', *Philosophical Review*, **80**.

Stich, S.: 1978, 'Beliefs and Subdoxastic States', *Philosophy of Science*, **45**, 499-518.

Stich, S. and T. Warfield (eds.): 1994, *Mental Representation: A Reader*, Blackwell, Oxford.

Wittgenstein, L.: 1969, *On Certainty*, Blackwell, Oxford.

Wright, C.: 1981, 'Rule-Following, Objectivity and the Theory of Meaning', in Holtzman & Leich, 1981, pp. 99-117.

Wright, C.: 1986a, 'How Can the Theory of Meaning be a Philosophical Project?', *Mind and Language* **1**, 31-44.

Wright, C.: 1986b, 'Theories of Meaning and Speaker's Knowledge', in Wright, 1993, pp.204-38.

Wright, C: 1993, *Realism, Meaning and Truth* (2nd ed.), Basil Blackwell, Oxford.

PART V: SPECIAL TOPICS

MICHAEL BRADIE

NATURALISM AND EVOLUTIONARY EPISTEMOLOGIES

TRADITIONAL EPISTEMOLOGIES

Traditional epistemology has its roots in Plato and the ancient skeptics. One strand emerges from Plato's interest in the problem of distinguishing between knowledge and true belief. His solution was to suggest that knowledge differs from true belief in being justified. Ancient skeptics complained that all attempts to provide any such justification were hopelessly flawed. Another strand emerges from the attempt to provide a reconstruction of human knowledge showing how the pieces of human knowledge fit together in a structure of mutual support. This project got its modern stamp from Descartes and comes in empiricist as well as rationalist versions which in turn can be given either a foundational or coherentist twist. The two strands are woven together by a common theme. The bonds that hold the reconstruction of human knowledge together are the justificational and evidential relations which enable us to distinguish knowledge from true belief.

The traditional approach is predicated on the assumption that epistemological questions have to be answered in ways which do not presuppose any particular knowledge. The argument is that any such appeal would obviously be question begging. Such approaches may be appropriately labeled "transcendental."

The Darwinian revolution of the nineteenth century suggested an alternative approach first explored by Dewey and the pragmatists. Human beings, as the products of evolutionary development, are natural beings. Their capacities for knowledge and belief are also the products of a natural evolutionary development. As such, there is some reason to suspect that knowing, as a natural activity, could and should be treated and analyzed along lines compatible with its status, i. e., by the methods of natural science. On this view, there is no sharp division of labor between science and epistemology. In particular, the results of particular sciences such as evolutionary biology and psychology are not ruled *a priori* irrelevant to the solution of epistemological problems. Such approaches, in general, are called naturalistic epistemologies, whether they are directly motivated by evolutionary considerations or not. Those which are directly motivated by evolutionary considerations and which argue that the growth of knowledge follows the pattern of evolution in biology are called "evolutionary epistemologies".

THREE DISTINCTIONS: TWO PROGRAMS

There are two interrelated but distinct programs which go by the name "evolutionary epistemology." One focuses on the development of cognitive mechanisms in animals and humans. This involves a straightforward extension of the biological

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 735-745.
© 2004 *Kluwer Academic Publishers.*

theory of evolution to those aspects or traits of animals which are the biological substrates of cognitive activity, e. g., their brains, sensory systems, motor systems, etc. The other program attempts to account for the evolution of ideas, scientific theories and culture in general by using models and metaphors drawn from evolutionary biology. Both programs have their roots in 19th century biology and social philosophy, in the work of Darwin, Spencer, James and others. There have been a number of attempts in the intervening years to develop the programs in detail (see the comprehensive bibliography constructed by Campbell and Gary Cziko at http://www.ed.uiuc.edu/facstaff/g-cziko/). Much of the contemporary work in evolutionary epistemology derives from the work of Konrad Lorenz (1977, 1982), Donald Campbell (1960, 1974a, et. al.), Karl Popper (1968, 1972, 1976, 1978, 1984) and Stephen Toulmin (1967, 1972, 1974, 1981). I have labeled these two programs EEM and EET (Bradie 1986). EEM is the label for the program which attempts to provide an evolutionary account of the development of cognitive structures. EET is the label for the program which attempts to analyze the development of human knowledge and epistemological norms by appealing to relevant biological considerations. Some of these attempts involve analyzing the growth of human knowledge in terms of evolutionary (selectionist) models and metaphors (e. g., Popper 1968, 1972; Toulmin 1972; Hull 1988). Others (e. g., Ruse 1986, Rescher 1977) argue for a biological grounding of epistemological norms and methodologies but eschew selectionist models of the growth of human knowledge as such.

The EEM program starts from the fact that human beings have evolved from ancestral forms as a result of natural selection and other evolutionary forces in much the same way as any other organisms. The ancestral pre-human forms presumably differed from us not only in physical form but in sensitive and cognitive capacities as well. It is easy to suppose that various increases in these capacities were selectively advantageous and so became fixed in the human lineage. These sensory and cognitive capacities are located in the brain. They include not only the sense organs but also the ability to use language, to formulate hypotheses and to engage in other forms of what we call higher abstract reasoning. At present, our understanding of how brains work to produce all these marvelous results is quite limited. Even more limited and speculative are the evolutionary scenarios that we construct to account for their emergence in human phylogenies. However, given our conviction that the Darwinian picture of the development of life on earth is broadly correct, we have no doubt that some such scenario must be correct. This is the gist of what I mean by the EEM program.

A clear statement of the EEM program can be found in Vollmer (1975, 102):

> Our cognitive apparatus is a result of evolution. The subjective cognitive structures are adapted to the world because they have evolved, in the course of evolution, in adaptation to that world. And they match (partially) the real structures because only such matching has made such survival possible (quoted by Bunge 1983, 8).

Lorenz expresses a similar sentiment:

> I consider human understanding in the same way as any other phylogenetically evolved function which serves the purposed of survival, that is, as a function of a natural physical interaction with a physical external world. (Lorenz 1977, 4)

Lorenz saw Kant as the intellectual ancestor of his view. The *a priori* categorical structures which organisms use to form their cognitive pictures of reality are to be understood as the *a posteriori* evolutionary products of phylogenetic development. Thus,

> One familiar with the innate modes of reaction of subhuman organisms can readily hypothesize that the *a priori* is due to hereditary differentiations of the central nervous system which have become characteristic of the species, producing hereditary dispositions to think in certain forms. (Lorenz 1982, 122)

This line of thinking treats Kant as a precursor of contemporary evolutionary epistemology. While there is much to recommend this point of view one should not ignore the arguments by Ruse (1986) to the effect that Hume and his naturalistic approach to ethics and epistemology is more properly understood as the "true" precursor of evolutionary epistemology.

Popper's endorsement of the EEM program can be seen in the following:

> The specifically human ability to know, and also the ability to produce scientific knowledge, are the results of natural selection. They are closely connected with the evolution of a specifically human language. The first thesis is almost trivial. (Popper 1984, 239)

In his 'Reply to My Critics', Popper notes some further consequences of evolutionary theory. From the fact that man is an animal and that animal senses have evolved from primitive beginnings, it follows, Popper thinks, that human knowledge is almost as fallible as animal knowledge and that human senses, like animal senses, are part of a "decoding mechanism." (Schilpp 1974, 1059) Elsewhere in that volume Popper reverses the metaphor and claims that human sensory organs are "conjectures!" (Schilpp 1974, 111) This line of argument is developed in considerable detail by Munz (1993).

Donald Campbell, in particular, notes with approval, Lorenz's biologizing of Kant and the implication that the categories, etc., are to be read 'descriptively' and not 'prescriptively.' He also advocates the view that "...evolution – even in its biological aspects – is a knowledge process, and ...the natural-selection paradigm for such knowledge increments can be generalized to other epistemic activities, such as learning, thought and science." (Schilpp 1974, 412)

Campbell, in addition, consistently endorses the applicability of a 'blind-selection-and-retention' model to explain not only the evolution of all biological structures (not merely cognitive structures) but also the growth of scientific knowledge which is more properly viewed as part of the complementary EET program.

The EET program addresses the relevance of biological considerations for understanding the growth of knowledge and the development of epistemological norms. The two programs are interrelated and one often finds the same authors arguing for both. The general trend (exemplified by, e. g., Lorenz, Campbell, Popper, Toulmin and Hull) is to attempt to develop selectionist models for the growth of scientific knowledge. Rescher (1977) and Ruse (1986) demur.

Campbell endorses Popper's treatment of the succession of theories in science as due to a selective elimination process analogous to the eliminative role of natural selection in biological evolution. In addition, trial and error *learning* by animals, including man, brings the evolutionary model to the ontogenesis of knowledge. (Schilpp 1974, 415f)

Stephen Toulmin has also developed a version of "epistemological Darwinism":

> Darwin's populational theory of 'variation and natural selection' is one illustration of a more general form of historical explanation; and ... this same pattern is applicable also, on appropriate conditions, to historical entities and populations of other kinds. (Toulmin 1972, 135)

Science, according to Toulmin, develops in a two-step process analogous to biological evolution. At each stage in the historical development of science, a pool of competing intellectual variants exists along with a selection process which determines which variants survive and which die out. (Toulmin 1967, 465)

David Hull defends a similar position. (Hull 1973, 1975, 1978, 1982, 1988) On Hull's view, neither biological evolution nor the growth of knowledge serves as the primary model in terms of which we are to understand the other. Hull prefers to develop a general analysis of "evolution through selection processes which applies equally to biological, social and cultural evolution." (Hull 1982, 275; Hull 1988) Hull's rationale for treating both biological evolution and conceptual evolution as exemplifications of some common general selectionist model is to undercut objections to selectionist accounts of conceptual change which emphasize the disanalogies between biological and conceptual change. (Hull 1988, 418) Although the specific mechanisms of change are not the same in the two cases (Hull 1988, 431) and there is no clear evidence that there is any "significant correlation between genetic and conceptual inclusive fitness," (Hull 1988, 282f), Hull argues that both processes can be profitably analyzed in terms of interaction, selection and differential replication.

More recently, Gary Cziko (1995) has echoed this theme. He advocates a "universal selection theory" as the best available account of the emergence of "adaptive complexity" in all its ramifications including the evolution of knowledge systems. Like Hull and Campbell, Cziko sees biological evolution by natural selection as just one exemplification of a general model that can be profitably used to explain the dynamics of a wide range of systems.

In contrast to these approaches, Michael Ruse, although a fervent critic of evolutionary epistemologies which promote selectionist models of conceptual change, nevertheless sees an important role for Darwinian insights into the development of scientific methods and traditional epistemological problems in general. (Ruse 1986) Ruse, in urging us to take Darwin seriously, argues against the attempts by "evolutionary epistemologists" to boldly lift the model of variation and selection which characterizes evolution by natural selection and use it to model scientific methodology. Ruse contends that if we are to take Darwin seriously, we must stick to fundamentals and eschew the facile application of selectionist models drawn from evolutionary biology. Scientific reasoning is a specimen of culture, and if we are to give a Darwinian account of it, that account must rely on the building blocks of culture which modern Darwinian thought has bequeathed to us. Those building blocks, according to Ruse, are the "epigenetic rules" which shape the ways in which our minds work.

Taking Darwin seriously in questions epistemological, for Ruse, involves determining what the epigenetic rules of scientific reasoning are. To ascertain what those rules are, we need to examine the practice of scientific method and infer from that what the rules must be in order to produce the practice that we observe. The model of science that Ruse adopts is an admittedly noncontroversial version of the standard model which includes elements of inductive and deductive reasoning, i. e., logic, mathematics, reasoning by analogy, the generation of laws, the attribution of causes, and appeals to simplicity and consilience. (Ruse 1986, 156ff.) A very proper nineteenth century view, indeed. These principles and methods are "rooted in our biology" and justified by their adaptive value to us or our proto-human ancestors. (Ruse 1986, 155)

Taking Darwin seriously involves explaining how we come to use these rules and methods. We do so, Ruse argues, because the methods and principles of scientific reason mirror or mimic general intellectual tendencies that we would expect would be of selective advantage to those of our ancestors who happened to have the good fortune to act on them.

This brief survey should serve to illustrate that there are, in fact, two quite distinct programs parading under the name "evolutionary epistemology" which are, nevertheless, interrelated. (But, cf. Cziko 1995 and Plotkin 1994) They are not, however, identical. There is a sense in which some version of the EEM program must be true if our current understanding of evolutionary processes is anywhere near correct. What remains to be seen is what useful insights, if any, will be forthcoming about the evolution of the cognitive mechanisms of organisms. A further question is what, if anything, any such results have to do with epistemology, whether narrowly or broadly conceived.

The success of the EET programs is much more problematic. Even if we could demonstrate that our brains and cognitive apparatuses in general have evolved under selection because of their cognizing abilities, it is not clear what follows from this. It is by no means a straightforward extension from our recognition that our organs of knowing are evolved structures to the conclusion that we endorse epistemic and methodological norms because of their selective advantage. Nor does it straightforwardly follow that a selectionist model of conceptual change is either a correct or even fruitful way of thinking about conceptual change. Ruse's view, on the other hand, while not endorsing a selectionist model of conceptual change, has an air of *post hoc* reconstructionism about it. Given that we endorse certain cognitive and epistemological methods, it is easy enough to claim that we do so for selective reasons but not so easy to see what we learn by so doing.

Phylogeny versus Ontogeny

A second distinction concerns ontogeny versus phylogeny. Biological development involves both ontogenetic and phylogenetic considerations. Thus, the development of specific traits, such as the opposable thumb in humans, can be viewed both from the point of view of the development of that trait in individual organisms (ontogeny) and the development of that trait in the human lineage (phylogeny). The development of knowledge and knowing mechanisms exhibits a parallel distinction. Thus, the growth of an individual's corpus of knowledge and epistemological norms is an ontogenetic process as is the maturation of his or her brain and nervous system. On the other hand, the growth of human knowledge and establishment of epistemological norms across generations is a phylogenetic process as is the development of brains in the human lineage. The EEM/EET distinction cuts across this distinction since we may be concerned either with the ontogenetic or phylogenetic development of, e.g., the brain or the ontogenetic or phylogenetic development of norms and knowledge corpora. One might expect that since current orthodoxy maintains that biological processes of ontogenesis proceed differently from the biological processes of phylogenesis, that evolutionary epistemologies would reflect this difference. Curiously enough, however, for the most part they do not.

The evolutionary considerations addressed in the previous section were directed towards phylogenetic change. Here I want to briefly discuss some of the applications of selectionist models to ontogenetic processes. These come in two varieties depending upon whether one focuses on the ontogenetic development of individual brains (an EEM project) or the ontogenesis of knowledge or norms in the individual (an EET project). We turn first to selectionist accounts of neural development.

The theory of neural Darwinism, as presented by Edelman and Changeaux, applies populational and variational models to the ontogenetic development of neuronal networks in the brains of individual organisms (Edelman 1985, 1987; Changeaux 1985; cf. Cain and Darden 1989) The basic idea is that the neurons of the brain do not develop according to a program "hardwired" in the genes. Rather, the specific interconnections and topology of neural networks form in accordance with selection pressures of various sorts. These views are still somewhat controversial and it is an ongoing research project to determine the exact nature of the neurophysiological processes that may be involved.

What of the ontogenesis of knowledge in an individual? Campbell, as was noted earlier, endorses Popper's "extension" of the trial and error model to include learning processes by individual organisms. In 'Of Clouds and Clocks,' Popper had claimed that organisms as well as phyla were "problem solvers." In fact, he puts forward there the curious analogy that as the actions of an individual organism are tentative solutions to problems faced by it in its environment, so individual organisms are tentative solutions to problems faced by the phylum of which they are members. (Popper 1972, 243) Elsewhere he draws the connection between the two in the following way. From an ontogenetic point of view, scientific explanations rest on the expectations of the newborn child. Children allegedly grow into critical adulthood by the well known Popperian process of conjecture and refutation, with the initial conjectures formed on the basis of innate expectations. These innate expectations are the result of phylogenetic development. So, from a phylogenetic point of view, today's science rests on the "expectations" of ancestral unicellular organisms. This is epitomized by the quip that "There is, as it were, only one step from the [ancestral] amoeba to Einstein" (Popper 1972, 347; Campbell, in his pre-Popperian days expresses a similar sentiment in Campbell 1960). This way of putting the point blurs the difference between the two programs. The phylogenetic development of innate expectations in organisms in a lineage is a question appropriate to EEM. The ontogenetic development of knowledge in an individual (as opposed to the ontogenetic development of the biological structures necessary for an individual to become a competent adult discussed above in connection with neural Darwinism) is a question in the EET program. The corresponding phylogenetic EET question concerns the historical evolution of science from, say, Aristotle to Einstein. Toulmin characterizes the distinction within the EET program in a clearer way:

> We ... face questions about the social, cultural, and intellectual changes that are responsible for the historical evolution of our various modes of life and thought – our institutions, our concepts, and our other practical procedures. (These questions correspond to questions about *phylogeny* in evolutionary biology.) Individually speaking, we ... face questions about the manner in which maturation and experience, socialization and enculturation shape the young child's capacities for rational thought and action – how the child comes to participate in his native society and culture. (These questions correspond to the questions about *ontogeny* in developmental biology.) (Toulmin 1981, 26)

Descriptive versus prescriptive epistemologies

A third distinction concerns descriptive versus prescriptive approaches to epistemology and the growth of human knowledge. Many have argued that neither the EEM programs nor the EET programs have anything at all to do with epistemology properly (i. e., traditionally) understood. The basis for this contention is that epistemology, properly understood, is a normative discipline, whereas the EEM and EET programs are concerned with the construction of causal and genetic (i. e., factual) models. No such models, it is alleged, can have anything important to contribute to normative epistemology.

Both evolutionary and naturalized epistemologies challenge the tradition in arguing that the description of cognitive processes is a more central epistemological concern than the search for foundations and principles of justification. Traditionalists have responded by challenging the legitimacy of the descriptivist's claim to be epistemologists at all. (E.g., Dretske 1971, Dretske 1985, Kim 1988, Stroud 1981, Stroud 1984, Hull 1982, Hull 1988.)

One way of sorting out the relationship between descriptive and traditional epistemology proceeds as follows:

(1) Descriptive epistemology is a competitor to traditional epistemology. On this view, both are trying to address the same concerns and offering competing solutions to similar problems. Insofar as the tradition has been concerned with normative and prescriptive claims, the traditionalists have argued that descriptive epistemology fails to address these traditional questions and is epistemology in name only. As Kim puts it, "For epistemology to go out of the business of justification is for it to go out of business." (Kim 1988, 391)

(2) Descriptive epistemology might be seen as complementary to traditional epistemology. On this view, the focus of traditional epistemology remains the justificational questions of the tradition. Descriptive epistemology (either a narrowly construed evolutionary epistemology or a more broadly construed naturalized epistemology) supplements this account with a psychological account or a genetic account of the origin of human knowledge. This is Donald Campbell's view as expressed in a number of papers including Campbell (1974) and Campbell (1977). Quine, who has argued that the tradition is at best misguided and at worst corrupt rejects both option (1) and (2).

(3) Descriptive epistemology might be seen as a successor discipline to traditional epistemology. On this reading, descriptive epistemology does not address the questions of traditional epistemology because it deems them irrelevant or

unanswerable or uninteresting. Many defenders of naturalized epistemologies fall into this camp including the early Quine (Quine 1960, 1969; but, see also, e. g., Davidson 1973, Dennett 1978, Harman 1982, Kornblith 1985, Bartley 1976, 1987a, 1987b, Munz 1985, Hull 1982, 1988; cf., Dewey 1910).

Insofar as option (3) entails the rejection of all the traditional normative questions associated with epistemology, it is open to the charge leveled against option (1). What remains when questions of justification are set aside, it has been charged, is epistemology in name only. For radicals like Rorty, who argue that much of the tradition in philosophy is wrongheaded, this suggests that there is no longer any point in doing epistemology under any name. For moderates like Quine, such an approach smacks of throwing the baby out with the bath water. In more recent papers, Quine has retreated from his apparently more radical earlier view, that naturalized epistemology must be purely descriptive, to a more tempered view which endorses, in a transformed way, the justificational questions of traditional epistemology. This has led some critics to charge that, in effect, Quine wants to have his cake and eat it too. But, this criticism is too harsh. There is room for a, suitably modulated, account of norms in an evolutionary epistemology.

Campbell, for one, insisted that his interests were "decidedly normative" as well as descriptive (Campbell, 1988, 374). In the William James lectures, presented in 1977, Campbell maintained that despite its focus on the empirical question of how organisms come to have knowledge, descriptive epistemology is *hypothetically* normative. Part of its task is to produce a theory about why certain practices such as science produce knowledge and other practices do not and "how one should go about . . . [doing science] . . . if one wants valid knowledge" (Campbell 1988, 444). Campbell does not go into detail about what this "hypothetical" normativity amounts to, but I am sympathetic to this line of thought. If one construes knowledge along Quinean lines as a holistic product of norms and experience, then just as our knowledge *claims* are conjectural and subject to revision so the *norms* we employ to validate them can be construed as conjectural and subject to revision as well (Cf. Bradie 1997; for a somewhat different defense of the place of norms in naturalized epistemologies, see Kitcher 1992).

Future Prospects

EEM programs are saddled with the typical uncertainties of phylogenetic reconstructions. Is this or that organ or structure an adaptation and if so, for what? In addition, there are the uncertainties which result from the necessarily sparse fossil record of brain and sensory organ development. The EET programs are even more problematic. While it is plausible enough to think that the evolutionary imprint on our organs of thought influences what and how we do think, it is not at all clear that the influence is direct, significant or detectible. Selectionist epistemologies which endorse a "trial and error" methodology as an appropriate model for understanding scientific change are not analytic consequences of accepting that the brain and other ancillary organs are adaptations which have evolved primarily under the influence of natural selection. The viability of such selectionist models is an empirical question which rests on the development of adequate models. Hull's (1988) is, as he himself admits, but the first step in that direction. Cziko (1995) is a manifesto urging the development of such models (Cf. Also the evolutionary game theory modeling approach of Harms 1997). Much hard empirical work needs to be done to sustain this line of research. It is one thing to construct suggestive selectionist models of knowledge acquisition but quite another to identify the physical and psychological systems that are doing the real causal work.

Non-selectionist evolutionary epistemologies, along the lines of Ruse (1986), face a different range of difficulties. It remains to be shown that any biological considerations are sufficiently restrictive to narrow the range of potential methodologies in any meaningful way.

Nevertheless, the emergence in the latter quarter of the twentieth century of serious efforts to provide an evolutionary account of human understanding has potentially radical consequences. The application of selectionist models to the development of human knowledge, for example, creates an immediate tension. Standard traditional accounts of the emergence and growth of scientific knowledge see science as a progressive enterprise which, under the appropriate conditions of rational and free inquiry, generates a body of knowledge which progressively converges on the truth. Selectionist models of biological evolution, on the other hand, are generally construed to be non-progressive or, at most, locally so. Rather than generating convergence, biological evolution produces diversity. Popper's evolutionary epistemology attempts to embrace both but does so uneasily. Kuhn's "scientific revolutions" account draws tentatively upon a Darwinian model, but when criticized, Kuhn retreated (cf. Kuhn 1970, 172f with Lakatos and Musgrave 1970, 264). Toulmin (1972) is a noteworthy exception. On his account, concepts of rationality are purely "local" and subject themselves to evolve. The net result is the need to abandon any sense of "goal directedness" in scientific inquiry. This is a radical consequence. Pursuing the evolutionary approach to its logical conclusion raises fundamental questions about the concepts of knowledge, truth, realism, justification and rationality.

Michael Bradie
Bowling Green State University

REFERENCES

Bartley, W. W.: 1976, 'The Philosophy of Karl Popper: Part I: Biology and Evolutionary Epistemology', *Philosophia* **6**, 463-494.

Bartley, W. W.: 1987a, 'Philosophy of Biology versus Philosophy of Physics', in G. Radnitzky and W. W. Bartley (eds.), *Evolutionary Epistemology: Theory of Rationality and the Sociology of Knowledge*, Open Court, LaSalle, IL.

Bartley, W. W.: 1987b, 'Theories of Rationality', in G. Radnitzky and W. W. Bartley (eds.), *Evolutionary Epistemology: Theory of Rationality and the Sociology of Knowledge*, Open Court, LaSalle, IL.

Bradie, M.: 1986, 'Assessing Evolutionary Epistemology', *Biology & Philosophy* **4**, 401-459.

Bradie, M.: 1989, 'Evolutionary Epistemology as Naturalized Epistemology', in K. Hahlweg and C. A. Hooker (eds.), *Issues in Evolutionary Epistemology*, SUNY Press, Albany, NY.

Bradie, M.: 1994, 'Epistemology from an Evolutionary Point of View', in E. Sober (ed.), *Conceptual Issues in Evolutionary Biology*, The MIT Press, Cambridge, MA.

Bradie, M.: 1997, 'Quine as an Evolutionary Epistemologist', *Epistemologica* **XX**, 175-210.

Bunge, M.: 1983, *Treatise on Basic Philosophy, V. 5: Epistemology and Methodology I: Exploring the World*, Reidel, Dordrecht.

Campbell, D. T.: 1960, 'Blind Variation and Selective Retention in Creative Thought as in other Knowledge Processes', *Psychological Review* **67**, 380-400.

Campbell, D. T.: 1974, 'Evolutionary Epistemology', in P. A. Schilpp (ed.), *The Philosophy of Karl Popper*, Open Court, LaSalle, IL.

Campbell, D.: 1988, *Methodology and Epistemology for Social Science*, The University of Chicago Press, Chicago.

Changeaux, J.-P.: 1985, *Neuronal Man*, Pantheon, New York.

Cziko, G.: 1995, *Without Miracles: universal selection theory and the second Darwinian revolution*, The MIT Press, Cambridge, MA.

Darden L. and J. A. Cain: 1989, 'Selection Type Theories', *Philosophy of Science* **56**, 106-129.

Davidson, D.: 1973, 'On the Very Idea of a Conceptual Scheme', *Proceedings of the American Philosophical Association* **47**, 5-20.

Dennett, D.: 1978, *Brainstorms*, The MIT Press, Cambridge, MA.

Dewey, J.: 1910, *The Influence of Darwinism on Philosophy and Other Essays in Contemporary Thought*, Henry Holt & Co., New York.

Dretske, F.: 1971, 'Perception from an Epistemological Point of View', *Journal of Philosophy* **68**, 584-591.

Dretske, F.: 1985, 'Machines and the Mental', *Proceedings and Addresses if the American Philosophical Association* **59**, 23-33.

Edelman, G. M.: 1985, 'Neural Darwinism: Population Thinking and Higher Brain Function', in M. Shafto (ed.), *How We Know: The Inner Frontier of Cognitive Science*, Harper & Row, San Francisco.

Edelman, G. M.: 1985, *Neural Darwinism: The Theory of Neuronal Group Selection*, Basic Books, New York.

Harman, G.: 1982, 'Metaphysical Realism and Moral Relativism', *Journal of Philosophy* **79**, 568-575.

Harms, W.: 1997, 'Reliability and Novelty: Information Gain in Multilevel Selection Systems', *Erkenntnis* **46**, 335-363.

Hull, D.: 1973, 'A Populational Approach to Scientific Change', *Science* **182**, 1121-1124.

Hull, D.: 1975, 'Central Subjects and Historical Narratives', *History and Theory* **14**, 253-274.

Hull, D.: 1982, 'The Naked Meme', in H. C. Plotkin (ed.), *Learning, Development and Culture*, John Wiley & Sons, New York.

Hull, D.: 1988, *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*, Chicago University Press, Chicago.

Kim, J.: 1988, 'What is 'Naturalized Epistemology'?', *Philosophical Perspectives 2. Epistemology* pp. 381-405.

Kitcher, P.: 1992, 'The Naturalists Return', *The Philosophical Review* **101**, 53-114.

Kornblith, H.: 1985, *Naturalizing Epistemology*, The MIT Press: Cambridge, Mass.

Kuhn, T.: 1970, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Lakatos, I. and A. Musgrave: 1970, *Criticism And The Growth Of Knowledge*, Cambridge University Press, Cambridge.

Lorenz, K.: 1977, *Behind the Mirror*, Methuen, London.

Lorenz, K.: 1982, 'Kant's Doctrine of the *a priori* in the Light of Contemporary Biology', in H. C. Plotkin (ed.), *Learning, Development, and Culture: Essays in Evolutionary Epistemology*, John Wiley & Sons, New York.

Munz, P.: 1985, *Our Knowledge of the Growth of Knowledge: Popper or Wittgenstein?*, Routledge & Kegan Paul, London.

Munz, P.: 1993, *Philosophical Darwinism: On the Origin of Knowledge by Means of Natural Selection*, Routledge, New York.

Plotkin, H.: 1994, *Darwin Machines and the Nature of Knowledge*, Harvard University Press, Cambridge, MA.

Popper, K. R.: 1968, *The Logic of Scientific Discovery*, Harper, New York.

Popper, K. R.: 1972, *Objective Knowledge: An Evolutionary Approach*, The Clarendon Press, Oxford.

Popper, K. R.: 1976, 'Darwinism as a Metaphysical Research Programme', *Methodology and Science* **9**, 103-119.

Popper, K. R.: 1978, 'Natural Selection and the Emergence of Mind', *Dialectica* **32**, 339-355.

Popper, K. R.: 1984, 'Evolutionary Epistemology', in J. W. Pollard (ed.), *Evolutionary Theory: Paths into the Future*, John Wiley & Sons Ltd., London.

Quine, W. V. O.: 1960, *Word and Object*, The MIT Press: Cambridge, Mass.

Quine, W. V. O.: 1969, *Ontological Relativity and Other Essays*, Columbia University Press, New York.

Rescher, N: 1977, *Methodological Pragmatism*, Basil Blackwell, Oxford.

Ruse, M.: 1986, *Taking Darwin Seriously : A Naturalistic Approach to Philosophy*, Basil Blackwell, Inc., New York.

Schilpp, P. A.: 1974, *The Philosophy of Karl Popper*, Open Court, LaSalle, IL.

Stroud, B.: 1981, 'The Significance of Naturalized Epistemology', in P. A. French, J. T. G. Uehling and H. K. Wettstein (eds.), *Midwest Studies in Philosophy VI*, University of Minnesota Press, Minneapolis.

Stroud, B.: 1984, *The Significance of Philosophical Skepticism*, Oxford University Press, Oxford.

Toulmin, S.: 1967, 'The Evolutionary Development of Natural Science', *American Scientist* **55**, 456-467.

Toulmin, S.: 1972, *Human Understanding: The Collective Use and Evolution of Concepts*, Princeton University Press, Princeton, NJ.

Toulmin, S.: 1974, 'Rationality and Scientific Discovery', in K. Schaffner and R. Cohen (eds.), *Boston Studies in the Philosophy of Science* **XX**, Reidel, Dordrecht.

Toulmin, S.: 1981, 'Evolution, Adaptation, and Human Understanding', in M. B. Brewer and B. E. Collins (eds.), *Scientific Inquiry and the Social Sciences: A Volume in Honor of Donald T. Campbell*, Jossey-Bass, San Francisco.

Vollmer, G.: 1975, *Evolutionare Erkenntnistheorie*, S. Hirzel, Frankfurt.

HARVEY SIEGEL

RELATIVISM

Epistemological[1] relativism may be defined as the view that knowledge (and/or truth or justification[2]) is relative – to time, to place, to society, to culture, to historical epoch, to conceptual scheme or framework, or to personal training or conviction – in that what counts as knowledge (or as true or justified) depends upon the value of one or more of these variables. Knowledge is relative in this way, according to the relativist, because different cultures, societies, epochs, etc. accept different sets of background principles, criteria, and/or standards[3] of evaluation for knowledge-claims, and there is no neutral way of choosing between these alternative sets of standards. So the relativist's basic thesis is that a claim's status as knowledge (and/or the truth or rational justifiability of such knowledge-claims) is relative to the standards used in evaluating such claims; and (further) that such alternative standards cannot themselves be neutrally evaluated in terms of some fair, encompassing meta-standard.[4] (The character of such 'neutrality' is addressed below.)

The doctrine of relativism is usually traced to Protagoras, who is portrayed in Plato's *Theaetetus* as holding that "man is the measure of all things" ('homo mensura'), and that any given thing "is to me such as it appears to me, and is to you such as it appears to you." (Plato 1961, 152a) Plato's Socrates characterizes Protagorean relativism as consisting in the view that "what seems true to anyone is true for him to whom it seems so." (Plato 1961, 170a) This view is a form of relativism in the sense just explained, since for the Protagorean there is no standard higher than the individual – with her own specific location in time, place, culture, framework, etc. – with reference to which claims to truth (and so knowledge) can be adjudicated. But relativism is best understood as a more general doctrine than the Protagorean version of it, which places the source of relativism at the level of standards rather than (as for the Protagorean) at the level of personal opinion or perception, and as such aptly characterizes more recent, influential versions of relativism.

ARGUMENTS CONTRA

Opponents of relativism have made many criticisms of the doctrine; by far the most fundamental is the charge that relativism is *self-referentially incoherent* or *self-refuting*, in that defending the doctrine requires one to give it up. There are several versions of the incoherence charge. The most powerful (for others, see Siegel 1987) is that relativism precludes the possibility of determining the truth, justificatory status, or, more generally, the epistemic merit of contentious claims and theses –

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 747-780.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

including itself – since according to relativism no claim or thesis can fail any test of epistemic adequacy or be judged unjustified or false.

Take Protagorean relativism as an example. If "what *seems* true [or justified] to anyone *is* true [or justified] for him to whom it seems so" (emphases added), then no sincere claim can fail to be true or be justifiably judged to be false. But if there is no possibility that a (sincerely held) claim or doctrine can be false, the very distinction between truth and falsity is given up; a 'false' belief is reduced simply to one which is not believed. While Protagorean relativism is in the first instance a doctrine about the relativity of truth, it is readily extended to matters of epistemic appraisal generally (as the bracketed insertions in the just-quoted expression of Protagorean relativism are meant to illustrate), and understood as asserting the relativity of standards of rightness[5] and justification as well as those of truth. If read in this way, it follows from this form of relativism that there is no possibility that a belief sincerely judged by a person to be right or justified can be wrong or unjustified. The end result is that the very notions of truth, rightness and justifiedness are undermined. But if this is so, relativism itself cannot be true, right or justified.

Relativism is thus (according to this argument) incoherent in that, if it is true (or right or justified), the very notion of truth (or of rightness or justifiedness) is undermined, in which case relativism cannot itself be true (or right or justified). This undermining results because the relativism of standards alleged by the relativist renders it impossible to distinguish truth (or rightness or justifiedness) from its (their) contrary (-ies). The *assertion and defense* of relativism requires one to presuppose neutral standards in accordance with which contentious claims and doctrines can be assessed; but relativism denies the possibility of evaluation in accordance with such neutral standards. Thus the doctrine of relativism cannot be coherently defended – it can be defended only by being given up.[6] Relativism is thus *impotent* – incapable of defending itself – and falls to this fundamental reflexive difficulty. Defending relativism non-relativistically is logically impossible, in that any such defense must appeal to that to which the relativist cannot appeal except by giving up relativism; while 'defending' relativism relativistically is not *defending* it, i.e., providing any reason for thinking it to be in any way epistemically superior to non-relativism, at all. (Siegel 1987, ch. 1)

To put this fundamental difficulty facing the relativist in a somewhat different way: insofar as she is taking issue with her non-relativist philosophical opponent, the relativist wants both *(a)* to offer a general, non-relative view of knowledge (and/or truth or justification), and assert that that general view – i.e., that knowledge is relative – is epistemically superior and preferable to its rivals; and also *(b)* to deny that such a general, non-relative view is possible or defensible. But the relativist cannot defend the view of knowledge offered in *(a)*, according to which relativism is epistemically superior to non-relativism, in a way consistent with her own commitment to relativism. On the other hand, 'defending' relativism in a way which does not assert its epistemic superiority is not to defend it at all; neither is it to engage seriously the cluster of issues which divide the relativist from her non-relativist philosophical opponent. Embracing *(b)* – i.e., denying that a general, non-relative view of knowledge (including the relativist view) is possible or defensible – similarly precludes the relativist from seriously engaging the issues to which her relativism is a response. Moreover, defending *(b)* requires a commitment to *(a)*, which commitment the commitment to *(b)* itself precludes.

In short: the relativist needs to embrace *both (a)*, in order to see her position both as a rival to, and, further, as epistemically superior to, the position of her non-relativist opponent; and *(b)*, in order to honor the fundamental requirements of relativism. But the mutual embrace of *(a)* and *(b)* is logically incoherent. For the embrace of *(a)* forces the rejection of *(b)*: if relativism is the epistemically superior view of knowledge (i.e., *(a)*), then one general view of knowledge is both possible and defensible as epistemically superior to its rivals (contrary to *(b)*). Similarly, the embrace of *(b)* forces the rejection of *(a)*: if no general, non-relative view of knowledge is possible or defensible (i.e., *(b)*), then it cannot be that relativism is epistemically superior to its rivals (contrary to *(a)*). Here again the argument strongly suggests that the assertion and defense of relativism is incoherent.[7]

This incoherence charge is by far the most difficult problem facing the relativist. It is worth noting that attempts to overcome the problem by appealing to the notion of *relative truth* appear not to succeed. Many versions of relativism rely on such a notion, but it is very difficult to make sense of it. An assertion that a proposition is 'true for me' (or 'true for members of my culture') is more readily understood as a claim concerning what I (or members of my culture, scheme, etc.) *believe* than it is as a claim ascribing to that proposition some special sort of truth. Constructing a conception of relative truth such that '*p* is relatively true' (or '*p* is true for *S*,' or '*p* is true for members of culture *C*') amounts to something stronger than '*S* believes that *p*' (or 'Members of culture *C* believe that *p*'), but weaker than '*p* is true (*simpliciter*),' has proved to be quite difficult, and is arguably beyond the conceptual resources available to the relativist. (Siegel 1987, 9-18)[8]

Moreover, even if a viable conception of relative truth could be developed, versions of relativism based on it would apparently still fall to the incoherence argument rehearsed above. In particular, a defense of relativism which rests on the notion of relative truth appears doomed to failure insofar as it seeks either to defend the notion of relative truth as superior to its 'absolutist' contrary, or to defend any particular relative truth *p* as in any way epistemically superior to equally relatively true *not-p* or arbitrary relative truth *q*. For any such defense would presuppose neutral, fair standards by appeal to which such epistemic superiority might be established, and such standards are precisely those to which the relativist, by virtue of her own commitment to that doctrine, cannot appeal.

Furthermore (as above), to decline to offer a defense – "you have your conception of truth, I have mine, and there is no question of one being 'better' than the other," or "you have your relative truths, I have mine, and there is no question of any relative truth being 'epistemically superior' to any other" – is to fail to acknowledge (or take seriously) the philosophical issues that divide the relativist from her non-relativist opponent. For if there is *no* sense, according to the relativist, in which her general epistemological view, her conception of truth, and the particular relative truths she embraces – in particular, her embrace of the relative truth of relativism itself – are epistemically superior to their alternatives, it is hard to understand the dispute between relativists and non-relativists *as a philosophical dispute*. In this case, the relativist seems to be saying "I'm a relativist, you're not, but your view is just as good (epistemically) as mine." If the relativist does say this

– if she declines to defend her view on the grounds that she does not regard it as epistemically superior to non-relativism – it is unclear why she should be regarded *as* a relativist at all; let alone why the non-relativist should be bothered by such a seemingly inert 'challenge.' Here we see again the problem of impotence, which arises with the relativist's declining to defend relativism just as surely as it results from her inability to do so.

Thus relying on the notion of 'relative truth' seems not to help the relativist here; indeed, the centuries-old preoccupation with the viability of that notion seems to be mainly irrelevant to the question of the viability of relativism when the latter is understood as a general epistemological doctrine. Whether the relativist's conception of truth is relative or non-relative, the assertion and defense of relativism appears to remain self-refuting, and so incoherent. (Siegel 1987, 18-20)[9]

ARGUMENTS PRO

Despite these ancient and powerful responses to relativism, the last several decades have witnessed a resurgence of the doctrine. Contemporary versions of relativism occur in a wide variety of philosophical contexts and enjoy an equally wide variety of philosophical pedigrees. Chief among them are versions of relativism spawned by Wittgensteinian considerations concerning language use, conceptual schemes or frameworks, and 'forms of life'; the 'strong programme' in the sociology of knowledge; a variety of quite different positions which might be grouped together under the heading of 'contemporary neo-Pragmatism'; and, perhaps most surprisingly, highly influential work in the philosophy of science. I briefly review some of these developments below. First, I consider two more general arguments for relativism, which play important roles in many more specific arguments for it: that which claims the impossibility of a *neutral* perspective sufficient to avoid relativism; and, relatedly, that which denies the possibility of transcending one's (relative) perspective, such *transcendence* being allegedly required to avoid relativism.

a) Is 'Neutral' Judgment Possible?

As just rehearsed, the argument that relativism is incoherent relies at key junctures on the possibility and accessibility of *neutral* standards in accordance with which knowledge-claims can be adjudicated. But relativists often reject the possibility of such standards, since relativism, as defined above, results (according to the relativist) in part because there are no neutral standards available by which the claims or criteria of rival perspectives can be fairly evaluated. That is, if you and I have a dispute – concerning a given claim's status as knowledge, or its truth or justificatory status, or the standards to which we should appeal in deciding such matters – the relativist's contention that such disputes can be resolved only relative to our respective standards (and not 'absolutely') rests on her contention that there are no 'meta-' or higher-order standards available to which we can appeal which will fairly or non-question-beggingly resolve our dispute.

Thus consider the famous dispute between Galileo and the Church concerning the existence of moons orbiting Jupiter. Not only did the two parties disagree as to the truth of the relevant claim – Galileo affirmed the existence of the moons, while his opponents denied it – they also disagreed about the relevant standards (telescopic observation? naked eye observation? Scripture? Aristotle?) to which appeal should be made in order to resolve their disagreement. The relativist here claims that such disputes admit of no non-relative resolution, precisely because there is no neutral, non-question-begging way to resolve the dispute concerning (meta-)standards. Any proposed meta-standard which favors regarding naked eye observation, Scripture, or the writings of Aristotle as the relevant standard by which to evaluate 'the moons exist' will be judged by Galileo as unfairly favoring his opponents, since he thinks he has good reasons to reject the epistemic authority of all these proposed standards; likewise, any proposed meta-standard that favors Galileo's preferred standard, telescopic observation, will be judged as unfair by his opponents, who claim to have good reasons to reject that proposed standard. In this way, the absence of neutral (meta-)standards seems to make the case for relativism.

However, it does not. The 'no neutrality, therefore relativism' argument just rehearsed has an ambiguity at its heart which undermines its ability to support relativism. Let us grant that there is no standard which is neutral *generally*, i.e., neutral with respect to all possible disputes. There may nevertheless be standards which, while not neutral in that sense, are neutral in the weaker sense that they do not unfairly prejudice any particular, live (at a time) dispute. So, for example, both Galileo and his opponents recognized *logic* (or, more broadly, *'reason'*) as a standard to which either disputant may fairly appeal. Both sides also agreed that, were Galileo able adequately to explain the workings of his newly invented telescope (something he could not do at the time of the dispute), that explanation would undermine his opponents' rejection of the proposed Galilean standard of telescopic observation – thus acknowledging *adequate explanation* as a relevant meta-standard for evaluating first-order standards (i.e., those relevant to the resolution of first-order disputes).[10]

Consequently, there is no reason to think that there were not – let alone could not be – neutral (meta-)standards available, in terms of which both the first-order dispute between Galileo and his opponents concerning the existence of the moons, and the second-order dispute between them concerning the appropriateness of the various proposed standards for judging first-order disputes, might be evaluated and, at least in principle, resolved. Of course the two meta-standards noted, logic (or 'reason') and explanatory adequacy, are not neutral with respect to all possible disputes. In particular, they might fail to be neutral with respect to disputes concerning the character and force of logic, and to disputes concerning the character of explanation and its possible tie to truth (although establishing this would require considerably more extensive discussion). Still, while not neutral *simpliciter*, they are in the relevant sense neutral in the Galileo case insofar as both sides both explicitly accept them and rely upon them in the execution of their respective cases. If in the end one side measures up less well against them than the other, that is a result which that side will not like, but such a result in itself is no reason to think such standards unfair, biased, or otherwise objectionable.

The neutrality required to avoid relativism is thus not some sort of *universal* neutrality – neutrality with respect to *every* possible dispute or *all* conceivable conceptual schemes – but only neutrality with respect to the issue at hand. Such neutrality, further, does not require that standards cannot discriminate among better or worse competing views, but rather simply that such discrimination must be fair to competing views, i.e., cannot be prejudicial toward or irrelevantly biased against one or another of them. There is no reason to think that *this* weaker sort of neutrality cannot, in principle, be had.

Moreover, to say that the standards just mentioned are in this weaker sense 'neutral' is not at all to say that resolving disputes in terms of them will always be easy. The Galileo case exemplifies how difficult such resolution can be, even absent worries about relativism. Still, as Popper says, one should not "exaggerate... a difficulty into an impossibility." (1970, 56-7) The difficulty of resolving genuinely hard cases does not yet give aid or comfort to the relativist.[11]

While in the Galileo case it is clear that the two sides explicitly accepted the meta-standards mentioned, it should be emphasized that such explicit acceptance is not required for this reply to the 'no neutrality, therefore relativism' argument to succeed. For if one of the parties were to have rejected one of the meta-standards, it might well nevertheless be the case that that party, in the case now being imagined, *should* have accepted it (or a related higher-order standard), and moreover that the dispute is rightly regarded as legitimately resolvable by reference to such a standard.

A disputant's rejection of a proposed standard, e.g., as biased or prejudiced against her, may well be legitimate, but it must be established as legitimate by argument – that is, she must provide reasons for thinking that the proposed standard in fact biases or prejudices the outcome of the dispute in an unacceptable way. For if such reasons cannot be produced, the rejection of the proposed standard not only will appear to her opponent to be, but will indeed be, arbitrary. On the other hand, if such reasons can be produced, then some standards – namely, those which sanction those reasons as epistemically forceful – will be presupposed by the party offering them, and will be offered as fair, non-prejudicial (meta-)standards in accordance with which the dispute on the table, concerning the objectionable non-neutrality of the proposed first-order standards, can be fairly resolved. This is a consequence, as noted above, of both sides regarding their dispute *as a genuine dispute*. (In fact, it is a straightforward application of the incoherence arguments against relativism rehearsed above.)

It goes without saying that the non-relativist should acknowledge that all proposed standards (and meta-standards) are open to challenge, and therefore that a disputant who challenges a standard proposed by her opponent is completely within her rights in doing so. But in order for such a challenge to succeed, the challenger must presuppose some (other) standard, one which is neutral with respect to the challenge at hand. Otherwise there would be no reason to regard the originally proposed standard as problematic. For on what basis could a given standard be challenged, absent some meta-standard in accordance with which that first-order standard is (according to the challenger) problematic?

Indeed, a relevant meta-standard which any such challenge must presuppose seems clearly enough to be that of neutrality itself: the problem with proposing a standard not embraced by one's protagonist in a dispute (say, Galileo's opponents proposing to Galileo that the standard of logic is appropriately appealed to in the resolution of their dispute, in the imagined case in which Galileo rejects it), if such a proposal is in fact problematic, is that appeal to that standard prejudices the resolution of the dispute in an unacceptable way. That is, appeal to such a standard is unacceptable (if it is) because its impact upon the dispute is such that any resolution flowing from it would objectionably privilege one of the disputants – it would fail to treat them neutrally or fairly. Thus successfully challenging the appeal to such a standard appears to require acceptance of a general (meta-)standard of neutrality, or fairness. Consequently, the relativist cannot challenge all such appeals to standards as unacceptably non-neutral except by presupposing that particular (meta-)standard herself.

Of course not all proposed standards will be acceptable. But the (meta-)standard of neutrality seems unproblematic, especially since (as just argued) any disputant wishing to challenge an opponent's proposed standards must accept that one; even though, by contrast, other standards – say, that disputants make their cases in 'formal' terms, or that those cases must comport with Scripture – will be obviously problematic (e.g., prejudicial) in some cases. The point is not that an appeal to standards one's opponent does not antecedently accept is always legitimate; it is not. The point, rather, is that challenging a proposed standard itself requires appeal to some standard or other – either a (meta-)standard of neutrality, or some other standard itself presumed to meet that of neutrality – if it is hoped that such a challenge might be successful. For without such an appeal it would be impossible to distinguish successful from unsuccessful challenges of standards. (That is, without such an appeal, the very notion of '*successful* challenge' would be undermined.) Consequently, while the relativist may well be right to protest appeals to standards which are not independently agreed to by the parties to some particular dispute, she can do so only against the background of other standards which she takes to be neutrally and fairly applied to the dispute in question.

I conclude that the relativist cannot consistently defend the premise of the 'no neutrality, therefore relativism' argument. Not only is the premise false, in that, so long as 'neutral' is understood appropriately, there not only can be but are neutral standards to which parties in particular disputes can legitimately appeal; in addition, the relativist must herself presuppose the falsity of the premise when reserving the right to criticize any proposed standard as unacceptably non-neutral. Moreover, as already noted, the assertion and defense of the 'no neutrality' argument for relativism requires appeal to the sort of standard to which the relativist cannot, by her own lights, consistently appeal. Thus relativism appears not to be established by this argument.

I immediately concede that I have not here considered any actual (i.e., in the literature) relativistic denials of neutrality. This is because my aim in this section has been to discuss the general problem, broadly conceived. As we will soon see, the 'no neutrality, therefore relativism' argument plays a key role in many of the specific cases for relativism to be discussed below; more specific appeals to it will be considered there. Still, the basic point is clear: while we may agree that neutrality *simpliciter* is not to be had (at least by creatures like us), the absence of this strong form of neutrality has no tendency to establish relativism; by the same token, we have as yet no reason to think that the weaker form of neutrality required for the avoidance of relativism in any given case cannot be had. Furthermore, the making of this case for relativism itself embroils the relativist in self-referential difficulties, since making it requires appeal to neutral standards of just the sort which the relativist abjures.[12] So this general argument for relativism, based on the impossibility of neutral judgment, does not succeed in establishing that doctrine.

A related general argument for relativism concerns not the impossibility of neutrality, but the impossibility of *transcendence*. I turn to it next.

b) Is It Possible to 'Transcend' One's Perspective?

It is widely acknowledged in contemporary discussion that one can never completely escape one's perspective, framework, or conceptual scheme and achieve a 'God's eye view' or a 'view from nowhere' (Nagel 1986); that all cognitive activity is inevitably conducted from some ongoing perspective or point of view. A typical expression of this thesis is Quine's (1960, 275-6):

> The philosopher's task differs from the others', then, in detail; but in no such drastic way as those suppose who imagine for the philosopher a vantage point outside the conceptual scheme that he takes in charge. There is no such cosmic exile. He cannot study and revise the fundamental conceptual scheme of science and common sense without having some conceptual scheme, whether the same or another no less in need of philosophical scrutiny, in which to work.

Philosophers generally grant Quine's point: there is no 'cosmic exile' from all conceptual schemes; one cannot cognize except from within the confines of some scheme or other. But from the relatively uncontroversial claim that we cannot escape all perspectives and achieve a 'view from nowhere,' it seems a short step to the relativistic conclusion that what we can know, or what can be true or justified, is itself relative to the frameworks which inevitably limit our judgment; that, since there is no 'perspectiveless' judgment, there is no possibility of achieving a perspective which would allow us to compare and evaluate (except in a question-begging way) either judgments issued from different perspectives, or alternative perspectives themselves. That is, the uncontroversial claim that all judgments inevitably occur in the context of some perspective or other might be thought to entail that all judgments are therefore *bound* or *determined* by such *inescapable* perspectives – and so that what a given epistemic agent is able to know, or regard as true or justified, is problematically *limited* by her perspective or framework in such a way, or to such an extent, that relativism inevitably results. Is relativism correctly derived in this way?

It is not – or so I will argue. The alleged entailment just mentioned fails; even though we cannot attain a 'perspectiveless perspective,' in the relevant sense we *can* nevertheless 'transcend' our frameworks and perspectives. Here, as in the discussion of neutrality above, we must distinguish between transcending or escaping any given perspective from transcending *all* such perspectives. Once this distinction is drawn, the 'no transcendence, therefore relativism' argument collapses.

Are we limited by our perspectives, such that we cannot achieve any critical perspective on them? Are we really 'trapped' within our perspectives in this way? Common sense and every day experience indicate the contrary. Perhaps the most obvious range of counter-examples involves the cognitive activities of children.

Children of a certain age, for example, can count and have a reasonable grasp of whole numbers, but have no understanding of fractions or decimals, i.e., parts of whole numbers. If asked 'is there a number between 1 and 2?,' they will answer in the negative, and will be unable to comprehend any suggestion to the contrary. But, given normal psychological development, within a few years such children will answer affirmatively; they will have no problem recognizing that, e.g., 1.5 is a number between 1 and 2, and more generally, that there are non-whole numbers. This seems a perfectly straightforward case of the modification of a perspective or framework (or of the abandonment of one framework for another) which belies the claim that we are trapped in, bound by, or limited to our frameworks.[13] (Scientific examples can equally easily be given, e.g., of the recognition of the existence of things too small to see with the naked eye, or of the interanimation of space and time and of the large scale non-Euclidean geometry of the universe.)

Very different sorts of examples can also be given. Consider, for example, the 'male sexist pig' who has no awareness or understanding of women other than as (sex) objects, but who in the course of his experience comes to realize (if only dimly) that he does treat women as objects, that many women want not to be so treated, and that there might well be something objectionable about treating women in that way. Suppose that this benighted male comes eventually to a full(er) awareness of the injustice of his earlier treatment of women; he comes to believe that it is wrong to treat women as objects and, over a considerable period of time and with the help of many women (and perhaps some courses in the Women's Studies Department), he develops a radically different and more respectful view of women and (hallelujah!) treats them accordingly. (Surely many men have had their consciousnesses raised to some extent in this way in recent decades.) Here again it seems that our subject has had his perspective altered and, indeed, improved; that is, he has 'transcended' his old sexist perspective for another.

In these examples not only have perspectives altered; the cognizers considered all regard their later perspectives as *improvements*; i.e., as better than, superior to, their earlier ones. If asked, these cognizers will be able to offer reasons which purport to justify those judgments of superiority. Those reasons, and the judgment that they are good ones which offer justification for the superiority of those later perspectives, are of course made from the perspective of those later perspectives or frameworks; they are not outside of all frameworks or issued from a perspectiveless perspective. Thus is acknowledged the uncontroversial premise of the argument under consideration. But the conclusion is undermined by the several counter-examples offered: epistemic agents always judge from some perspective or other, but there is no reason to think that they are trapped in or bound by their perspectives such that they cannot subject them to critical scrutiny. In this sense, we *can* 'transcend' our perspectives; and this sense is sufficient to defeat the argument for relativism we have been considering. As Popper puts the point:

> I do admit that at any moment we are prisoners caught in the framework of our theories; our expectations; our past experiences; our language. But we are prisoners in a Pickwickian sense: if we try, we can break out of our frameworks at any time. Admittedly, we shall find ourselves again in a framework, but it will be a better and roomier one; and we can at any moment break out of it again.

> The central point is that a critical discussion and a comparison of the various frameworks is always possible. (1970, 56)

Here Popper clearly draws the crucial distinction which undermines this path to relativism. While the Quinean point that we inevitably judge from some framework or other, that we cannot judge from a perspectiveless perspective, must be granted, it does not follow that our judgments are necessarily tainted by the fact that they are made from some framework or other. On the contrary, we can and regularly do 'transcend' our frameworks from the perspective of other, 'roomier' ones, in which can fit both our earlier one and relevant rivals to it – and in this way fair, non-relative evaluations of both our judgments and the frameworks/perspectives from which they are made are possible.[14]

Of course the 'framework relativist' may reject these alleged examples of transcendence, and in this way seek to preserve the argument we have been considering. This raises in a pointed way the question: what *are* 'frameworks,' or 'conceptual schemes' or 'perspectives,' such that our judgments and our ability to know is bound by them in a way which precludes transcendence? I have thus far understood these locutions in an intuitive and rather uncritical way, since it seems clear that the examples given – do/do not recognize non-whole numbers, do/do not recognize the existence of objects too small to see with the naked eye, do/do not recognize women other than as (sex) objects, etc. – are sufficiently general that such differences constitute differences in conceptual framework or scheme if anything does. An equally plausible example of alternative schemes are the Galilean and Aristotelian schemes discussed above. Understood so generously that all these examples are indeed examples of alternative frameworks or schemes, the argument for relativism based upon that generous understanding of these terms seems clearly deficient. Attempts to resuscitate the argument minimally require a more careful explication of these terms than I have given them here – and, it must be said, than defenders of 'framework' relativism have typically given them. Further, they require attention to Davidson's (1973) famous argument against the possibility of such alternative schemes (and hence of a relativism based upon them).[15] Absent such efforts, the 'no transcendence, therefore relativism' argument seems clearly to fail.

As in the discussion of neutrality above, here too I have treated the difficulty raised by 'transcendence' in very general terms and have refrained from citing specific versions of the 'no transcendence, therefore relativism' argument. In the following sections we will see that several influential arguments for relativism utilize one or both of the arguments just discussed.

c) Kuhn and Relativism in the Philosophy of Science

Much philosophical water has passed under the bridge during the nearly four decades since Kuhn's highly influential *The Structure of Scientific Revolutions* (1962) first appeared.[16] I will not review the book's basic claims, initial reception, or lasting impact on the philosophy of science at length here; rather, I concentrate on its contribution to contemporary defenses of relativism.[17]

Both the 'no neutrality' and the 'no transcendence' arguments for relativism are present in Kuhn's apparently relativistic position – indeed, they seem to merge together into one argument in Kuhn's hands. For Kuhn famously (though this attribution to Kuhn is not entirely uncontroversial) argues both that there can be no neutral choice between rival, competing paradigms, and that paradigms are inescapable in the sense that it is not possible rationally to transcend them.

While it is significant that Kuhn himself denied being a relativist and claimed to "categorically reject" that label (1970, 234), many of his readers interpreted his discussions of 'paradigms,' 'incommensurability,' 'revolutionary science,' 'scientific revolutions,' 'gestalt switches,' and the like relativistically. Kuhn's view is especially suggestive of relativism in its apparent contention that paradigms bring with them their own criteria of paradigm evaluation – criteria which are internal to paradigms, but must nevertheless be appealed to in evaluating rival paradigms during episodes of revolutionary science. (Notice how this view appeals to the premises of both the 'no neutrality' and the 'no transcendence' arguments for relativism.) Kuhn's early critics, e.g., Shapere (1964), Scheffler (1967), and Kordig (1971) seized on this point in arguing against Kuhnian versions of relativism. They argued, among other things, both that paradigm-neutral criteria of paradigm evaluation were available, and that Kuhn's view, when interpreted relativistically, succumbed to versions of the incoherence charge.

Whether or not Kuhn's view is rightly interpreted in this way remains a controversial question: Kuhn (and his sympathetic interpreters[18]) continued, until his recent death, to insist that his view is not relativist, while some of his critics persist in finding relativism entailed by his more general philosophy of science. This question of interpretation is, for present purposes, less important than the relatively uncontroversial fact that Kuhn does not see his views concerning science as providing a basis from which to argue for relativism. Kuhn rejects relativism, and denies that his philosophy of science entails it; some critics uphold that entailment, but use it not as an argument for relativism but rather as an argument against Kuhnian philosophy of science. Both sides, in other words, agree in their rejection of relativism; they disagree not on the viability of relativism but rather[19] on whether Kuhn's views in fact lends support to that doctrine. In other words, neither Kuhn nor his critics find support for relativism in Kuhn's philosophy of science.[20]

There is of course more to recent philosophy of science than Kuhn. Themes and theses developed both in Kuhn's work and in the work of Quine, Popper, Feyerabend[21], and many others – incommensurability, holism, the non-cumulativity of scientific knowledge, the theory-ladenness of observation, the underdetermination of theory by data, the indeterminacy of reference, the failure of traditional accounts of scientific progress and of objective testing of scientific theories, the conventionality of methodological rules and the role of convention in observation statements, a holistic view of meaning, etc. – all have been used in arguing for relativism. They have been rehearsed to good effect by Laudan in his dialogues on *Science and Relativism* (1990), who engagingly labels the relativist thesis "that remarkable thesis of cognitive egalitarianism" (69), and who elaborates the ways in which these arguments for relativism face both specific difficulties and the general self-refutation problem described above. (See also Laudan 1988.) I do not have the space to address these several topics in the philosophy of science here; the reader may pursue them in the article by Humphreys in this *Handbook* or in a host of other recent works in the philosophy of science, e.g., Papineau 1996.[22]

d) The Sociology of Science and the 'Strong Programme'

The sociology of scientific knowledge concerns itself with the sociological processes through which such knowledge is generated or produced, the processes through which it is 'legitimated' and accepted within a particular community, and other sociological processes and phenomena which play a role in the collective human effort to know. Traditionally, this sort of sociological investigation into the production, acceptance, legitimation and dissemination of knowledge has been taken by sociologists and epistemologists alike to be distinct from genuine epistemological inquiry, for the most that can be expected from the former sort of inquiry is a descriptive, *causal* account of how some particular community *C* produced and came to accept some knowledge-claim *p* or theory *T*, while the truth and/or justificatory status of *p* and *T* cannot be settled by such causal accounts: *C*'s *regarding p* as true or justified, however caused, is one thing; *p*'s *being* true or justified quite another. In this way a sharp division between sociological and epistemological inquiry concerning science and its claims to knowledge has traditionally been drawn, a division which cedes to sociology the task of describing and explaining scientific beliefs and attitudes at the sociological level, and to epistemology the task of evaluating such beliefs and, more generally, dealing with the normative assessment of candidate knowledge-claims.[23] Indeed, it has seemed to many that the 'sociology of *knowledge*' is a misnomer, in that inquiry conducted under that banner happily ignores any distinction between genuine knowledge and its counterfeits, and is better called the sociology, not of knowledge, but of *belief*.

Leaving the question of what such inquiry should be called to one side, advocates of the 'Strong Programme' explicitly reject, for the purposes of their inquiries, any such distinction between genuine knowledge and spurious impostors to that title, and explicitly accept that, for them, knowledge is nothing more than belief. David Bloor, perhaps the most visible leader of the strong programme, writes: "Knowledge for the sociologist is whatever men take to be knowledge. It consists of those beliefs which men confidently hold to and live by." (Bloor 1976, p. 2) As Barry Barnes and Bloor, in their widely cited defense of relativism, put it: "We refer to any collectively accepted system of belief as 'knowledge'." (1982, p. 22 n5) Their preference for this "terminological convention" (ibid.) concerning 'knowledge' – in contrast to the more usual 'convention,' which takes for granted that, since one of the central tasks of epistemology is to say what knowledge is, for purposes of epistemological theorizing it is of central importance to distinguish between genuine knowledge and spurious contenders for that title, however widely believed – has the unfortunate consequence that much of the debate between proponents and opponents of the relativism of the 'strong programmers' seems to be ineffectual, due to these very different understandings of 'knowledge.' Nevertheless, in view of the wide-ranging influence of the strong programme in the broad area of science studies, the centrality of relativism in the overall perspective of that programme, and the fundamental status of Barnes and Bloor's argument for relativism in that perspective, it behooves us to consider that argument here.

Central to their case for relativism is their claim that *relativism is required for science*: "Far from being a threat to the scientific understanding of forms of knowledge, relativism is required by it. Our claim is that relativism is essential to all those disciplines such as anthropology, sociology, the history of institutions and ideas, and even cognitive psychology, which account for the diversity of systems of knowledge, their distribution and the manner of their change." (1982, 21-2) There are two things to notice about this proclamation. First, it must be remembered that by 'knowledge' Barnes and Bloor mean *belief*; their claim is that social scientists studying alternative systems of belief and the dynamics of belief change at the social level, if that study is to be scientific, must study both systems thought by the sociologist to be normatively praiseworthy, and systems thought to be less praiseworthy. No epistemologist who rejects relativism, and who believes that the non-relative normative evaluation of belief is possible, need disagree with this. But second, the proclamation is unclear as to the sense of 'relativism' alleged here to be 'essential' for social scientific inquiry: are Barnes and Bloor making the innocuous point that social scientists studying belief distribution and the dynamics of belief change must study belief systems of both epistemically meritorious and epistemically less meritorious normative status; or the philosophically more contentious claim that any such distinctions concerning epistemic merit are illusory? (Only the latter would qualify their view as a version of relativism of the sort we are concerned with here.)

The answer to this question is, unfortunately, less than clear. On the one hand, they endorse what I just called the 'innocuous point':

> Our equivalence postulate is that all beliefs are on a par with one another with respect to the causes of their credibility....The position we shall defend is that the incidence of all beliefs without exception calls for empirical investigation and must be accounted for by finding the specific, local causes of this credibility. This means that regardless of whether the sociologist evaluates a belief as true or rational, or as false and irrational, he must search for the causes of its credibility. In all cases he will ask, for instance, if a belief is part of the routine cognitive and technical competences handed down from generation to generation. Is it enjoined by the authorities of the society? Is it transmitted by established institutions of socialization or supported by accepted agencies of social control? Is it bound up with patterns of vested interest?....All of these questions can, and should, be answered without regard to the status of the belief as it is judged and evaluated by the sociologist's own standards. (1982, 23)

In this central passage Barnes and Bloor are clear that (a) epistemic evaluation is possible (although, as we will see in a moment, only relative to local contexts), even though the sociologist is to ignore such evaluation in her inquiries and investigate the causes of the credibility (or lack thereof) of all beliefs independently of their normative status, and (b) by 'causes of credibility' they mean those factors which cause believers to believe as they do, i.e., to regard some beliefs as credible and others not. The causes of a belief's credibility thus are not, for Barnes and Bloor (contrary to some causal theories of justification), those factors which cause beliefs to *be* justified or worthy of belief; they are rather the factors which cause beliefs to be *regarded* by believers *as* credible (although again, as we'll see in a moment, Barnes and Bloor reject this distinction). The epistemic status of all beliefs is thus left open: once the sociologist identifies the causes of community *C*'s regarding belief system *BS* as credible, her work is done. It is no concern of the sociologist to determine whether or not beliefs so regarded really are credible. So far, then, Barnes and Bloor are not committed to any philosophically controversial sort of relativism.

But on the other hand they also endorse what I called above the 'philosophically more contentious claim' committing themselves to epistemological relativism of the sort with which we are here concerned. Discussing two tribes and their local epistemic predilections, Barnes and Bloor write:

> The crucial point is that a relativist accepts that his preferences and evaluations are as context-bound as those of the tribes T1 and T2. Similarly he accepts that none of the justifications of his preferences can be formulated in absolute or context-independent terms. In the last analysis, he acknowledges that his justifications will stop at some principle or alleged matter of fact that only has local credibility....For the relativist there is no sense attached to the idea that some standards or beliefs are really rational as distinct from merely locally accepted as such. Because he thinks that there are no context-free or super-cultural norms of rationality he does not see rationally and irrationally held beliefs as making up two distinct and qualitatively different classes of thing....Hence the relativist conclusion that they are to be explained in the same way. (1982, 27-8)

Unlike the passage cited earlier, in this passage Barnes and Bloor clearly endorse an epistemologically contentious form of relativism. Let us briefly examine their case.

i) First, as this passage makes clear, Barnes and Bloor reject the distinction drawn above between beliefs which are *regarded*, perhaps erroneously, *as* justified, and beliefs which actually *are* justified: 'For the relativist there is no sense attached to the idea that some standards or beliefs are really rational as distinct from merely locally accepted as such.' This is parallel to their rejection of any distinction between genuine knowledge and a counterfeit taken by some to be genuine. Genuine knowledge, and 'really rational' beliefs, just are what people regard as such; to be *regarded as* genuine is to *be* genuine.

There are three points to make here. The first is that this 'locality claim' (let us call it) is not a consequence of the equivalence postulate concerning the causes of credibility of beliefs with which Barnes and Bloor define their brand of relativism; it is an independent dimension of their view which requires its own justification (to be considered below). The second is that their equation of genuine knowledge (and 'really rational' belief) and that which is taken to be knowledge (and rational belief) flows naturally from their initial decision to adopt the 'convention' according to which 'knowledge' is defined as belief. Insofar, their rejection of the 'is regarded as/is' distinction is of no epistemological moment, since epistemologists are concerned with a quite different conception of knowledge, and are centrally concerned to distinguish the genuine article from imposters, however sincerely they might be embraced as genuine by some cognizers.

But third, Barnes and Bloor do offer a reason for rejecting any such distinction, namely that all such judgments of genuineness will themselves be only 'local': "...a relativist accepts that his preferences and evaluations are...context-bound....Similarly he accepts that none of the justifications of his preferences can be formulated in absolute or context-independent terms. In the last analysis, he acknowledges that his justifications will stop at some principle or alleged matter of fact that only has local credibility." That is, it is not possible for any cognizer, including the sociologist, to escape her local context and judge from some 'context-free,' 'super-cultural' or context-independent perspective.

ii) This claim will sound familiar to the attentive reader. It is, in fact, nothing more than (a version of) the conclusion of the 'no transcendence' argument for relativism addressed above (and possibly of the 'no neutrality' argument as well). Barnes and Bloor's argument is in the end one of very simple form: all judgment is local – no judgments have any positive epistemic status beyond that granted them by epistemic agents in some locale, and there is no getting beyond such locales to reach a context-independent platform from which to judge – therefore relativism. I will not repeat my earlier discussions of the 'no transcendence' and 'no neutrality' arguments for relativism here. It is enough to note that Barnes and Bloor's case for relativism – insofar as it goes beyond their decision to regard 'knowledge' as belief – rests upon these general, but unsuccessful, arguments for relativism.

iii) Barnes and Bloor's 'equivalence postulate' insists that all beliefs, however appraised from whatever perspective, be dealt with in the same way by the sociologist: that is, their 'credibility' is to be *explained causally*. The sociologist's task is to identify the 'causes of credibility' of beliefs, i.e., the social forces which explain their development, acceptance, and change. This causal thesis is not something that the opponent of relativism need reject, since that opponent can simply distinguish between the causes of belief, on the one hand, and the epistemic status of belief, however caused, on the other. Barnes and Bloor would reject this distinction, since 'epistemic status' for them just means 'locally perceived epistemic status,' and the causal question in which they are interested is precisely: what social forces cause belief system *BS* to be perceived, in a given locale, as having the status it is perceived to have? But the non-relativist can happily acknowledge the scientific legitimacy of the question. The important point here is that the legitimacy of the question, and the 'equivalence postulate' more generally, offers no support to relativism; the symmetry of explanation is perfectly compatible with the non-relativity of epistemic evaluation. The social forces (feudalism, religion, poverty, etc.) which brought about the acceptance of the Aristotelian belief system in the Middle Ages is one thing, the epistemic status of that system another. Of course Barnes and Bloor reject any non-relativist reading of the latter, but, as we have seen, their reason for doing so – the 'no transcendence' argument – fails.

But there is a further question here that deserves brief comment. Many authors[24], on both sides of the debate, take the equivalence postulate to entail that the sociological explanation of belief cannot include appeal to reasons which believers judge to be *good* reasons which *justify* belief. These authors hold that explanation in terms of reasons is a kind of explanation which is fundamentally different from explanation in terms of (sociological) causes. I cannot enter into this controversy here, but it is worth noting that (a) at least on many accounts of these things, reasons *can* be causes, and, to the extent that reasons can be causes, they can be the causes of (perceived) credibility; (b) therefore, the equivalence postulate does not rule out (causal) explanation of belief in terms of reasons whose perceived epistemic strength causes belief – e.g., 'community *C* is caused by its appraisal of the evidence to regard theory *T* as highly credible'; and (c), therefore, the equivalence postulate in no way entails relativism.[25]

iv) As we have seen, for Barnes and Bloor there is nothing more to 'knowledge' than community approval.[26] The task of the sociologist of science is not to give an epistemic account of why community *C rightly* regards some theory *T* or claim *p* as knowledge (or justified), but rather to give a causal account of community *C*'s coming to so regard them.

Consider the character of such a causal account. Presumably it will have the general form: '(Particular) social forces cause the credibility of belief systems within a given community,' or, schematically, '*SF* cause the credibility of *BS* in *C*.' So suppose the sociologist proposes such a causal account of belief credibility – say, that the belief that relativism is self-referentially incoherent is caused to be credible in the community of analytic epistemologists in the second half of the twentieth century by social forces involving the elite status of private research institutions, the reward system within such institutions, etc. How do Barnes and Bloor regard such accounts? As relativists, they seem to have no choice but to regard them relativistically: within community of sociologists *C** – say, the one located in Edinburgh and environs in the last quarter of the twentieth century – social forces cause the belief in question to be highly credible; whereas within community *C*** – say, the one located around Merton in the United States in the third quarter of that century – that belief is caused by social forces to be less credible. In both communities credibility is just 'credibility-as-perceived-in that-community'; to be *regarded as* credible is to *be* credible. Barnes and Bloor are clear that they accept this consequence of their views: the sociologist enjoys no special exemption from the 'equivalence postulate'; the credibility of her beliefs, like all scientific and other beliefs, is to be explained causally.

So far none of this poses any difficulty for Barnes and Bloor. But consider now the case in which two different communities of sociologists account for the credibility of a belief system in a third community, i.e., in which *C** and *C*** offer alternative accounts of the social forces which cause a belief (system) to be credible in a third community *C*. Let *C** and *C*** be the communities of sociologists just identified; let *C* be the community of analytic epistemologists in the United States and Western Europe in the third quarter of the twentieth century;[27] let *BS* be that system of beliefs concerning knowledge, truth, justification, etc., which includes the belief that relativism is self-referentially incoherent; let *SF** be the social forces cited by *C** as those which cause the credibility of *BS* in *C* (for example, social and economic forces involving the power structure, reward system, and student selection procedures of prestigious universities during the time period in question); let *SF*** be the social forces cited by *C*** as those which cause the credibility of *BS* in *C* (for example, social forces which encourage respect for conservative values such as '(perceived) common sense'); finally, let *CC** be the account of the causes of credibility of *BS* in *C* offered by *C**, and let *CC*** be the account of the causes of credibility of *BS* in *C* offered by *C***. The question is: how are we to think about these alternative accounts *CC** and *CC***? Barnes and Bloor regard the evaluation of these alternatives as a *scientific* matter: the sociologist of knowledge is, after all, a scientist. But they also regard all such judgments as relative: the scientific worth of these accounts will be judged variously – or rather, will be caused to be credible to varying degrees – by scientists in differing communities. But this raises the question: why do Barnes and Bloor place so much importance on the *scientific* character of sociological accounts of the causes of credibility of belief systems, if all such accounts will themselves have only local credibility?

To sharpen this problem: suppose Barnes and Bloor favor some particular *CC**, and their sociological opponents (the 'weak programmers') favor an incompatible *CC***, of the credibility of some *BS* in some *C*. As relativists, Barnes and Bloor seem forced to acknowledge that their preferred account *CC** itself has only local credibility – i.e., it is caused to be credible in the community of strong programmers – while the account they reject, *CC***, is equally locally credible in the rival community of weak programmers. Is this sensibly regarded as a *scientific* account of scientific knowledge? Since judgments of the causes of credibility, and of the scientific merits of competing accounts of those causes, are themselves relative to locale, it seems that Barnes and Bloor's relativism is at odds with their desire for scientific respectability.

v) This last point brings us, finally, back to the problem of incoherence. Barnes and Bloor appear not to have overcome this problem. First, as just noted, their yearning for a *scientific* sociology of science does not sit well with their endorsement of relativism, since the former requires a non-relativistic notion of causality, and a non-relativistic account of the specific causes of credibility of any particular belief system, which the latter precludes.

Second, their argument for relativism itself requires the rejection of that conclusion. Barnes and Bloor claim to show, in their discussion, that "the balance of argument favours a relativist theory of knowledge." (1982, p. 21) By this it is clear that they do not mean that their argument supports relativism only from the perspective of their own community of sociologists, but rather that it supports it generally, and should be found persuasive even by those outside that community (e.g., philosophers who endorse "rationalism"). (ibid.) Insofar as they see themselves as providing a justification of relativism which has epistemic force beyond their local community of sociologists, and as providing a case for thinking that 'rationalism' is mistaken – as they clearly do see themselves as doing – their relativism contravenes these claims. For if their arguments are successful, and their claims correct (or justified), the epistemic status of these arguments and claims extend beyond the bounds of their local community, thus undermining their relativism. If, on the other hand, their relativism remains, then their claim to have arguments for it whose force extends beyond their community is undermined, since their relativism, according to which epistemic judgments are necessarily local and context-bound, explicitly rejects any such possibility. Either way, their relativism is incompatible with their claim to be able to justify it in terms of 'the balance of argument.' This combination remains incoherent: the latter depends upon a non-relative sense of 'argument' or 'evidence' which the former precludes.

Of course Barnes and Bloor could bite the bullet here and retreat to the view that the balance of argument does not favor relativism *tout court*, but does so only for those already on the inside of their community – that that balance favors relativism only locally, i.e., relative to their community. In this case, their argument would be presented as having no tendency or ability to establish the error of 'rationalist' ways to those in rationalist communities, let alone to fair-minded students of the issue generally. But if their case is indeed taken by them to be limited in this way, why bother making that case in the first place? Here we see again relativism's impotence.

Given the quite familiar way in which Barnes and Bloor face the incoherence problem, their attempt to deflect it requires brief comment. They eschew two alternative 'equivalence postulates' – that all "general conceptions of the natural order" are either equally false, or equally true – because they both "run into technical difficulties" involving incoherence. (1982, 22) In favoring their chosen 'equivalence postulate' concerning the 'causes of credibility' of beliefs, with which we have been concerned throughout this section, Barnes and Bloor believe themselves to have avoided these 'technical difficulties.' (Space precludes speculation concerning the causes of the credibility (for them) of *that* belief.) I have just argued that, on the contrary, those technical difficulties have *not* been overcome – mainly because, independently of their chosen equivalence postulate, they hold that all judgments of truth, justification, etc., are equally *local* and admit of no higher-order assessment – that is, they endorse the problematic 'no transcendence' argument for relativism – and this is sufficient to give life to the 'technical difficulties' involving incoherence.

In any case, and despite their claim that their choice of equivalence postulate allows them to avoid such technical difficulties, Barnes and Bloor also reject the incoherence charge *überhaupt*, on the authority of Mary Hesse: "The claim that relativism is 'self-refuting' is thoroughly discussed and thoroughly demolished in Mary Hesse [1980]." (1982, 23 n6) Has Hesse 'demolished' this claim? Alas not. While Hesse's interesting discussion may plausibly be seen to establish, as has already been noted above, that the causal explanation of a community's finding a belief system to be credible (or not) is independent of that system's epistemic status (from the sociologist's point of view), it does not in the least establish relativism or defeat the self-refutation charge. Rather, Hesse argues only that if "we *shift* our concept of 'knowledge'...so that...knowledge is now taken to be what is accepted as such in our culture," then "the alleged [self-]refutation becomes an equivocation." (1980, 42, emphasis in original). But this (a) follows Barnes and Bloor in understanding the strong programme as one involving the sociological study of belief rather than knowledge, thus trivializing it (as discussed above); (b) dismisses the problem of relativism by means of stipulative redefinition, but does not resolve it; (c) fails to face the problem raised by judgments for and against such a shift being themselves relative to locales; and (d) rests, as does Barnes and Bloor's discussion, on the 'no transcendence' argument for relativism criticized above. Hesse rejects any appeal to "transcendent rationality" (56) which is somehow beyond the pale of sociological explanation, but so do (or at least should) non-relativists: that rejection follows from the equivalence postulate, but, as we have seen, from that postulate relativism does not follow. The bottom line here is that Hesse's discussion does not in the least 'demolish' the self-refutation charge.

To summarize: 1. Given their refusal to distinguish between knowledge and belief, Barnes and Bloor's arguments concerning the 'equivalence postulate' establish at most the relativity of belief. This sort of relativism is uncontroversial, indeed trivial. 2. The equivalence postulate concerning the causes of credibility does not entail relativism; only 'locality' – a quite independent thesis – entails this conclusion. 3. The argument for this thesis relies on the unsuccessful 'no transcendence' argument for relativism, and so fails. 4. A non-relative notion of causality appears to be required for the scientific study of belief that the strong programme recommends. 5. Finally, despite their heroic attempts to deflect it, the self-refutation/incoherence problem remains as much a problem for Barnes and Bloor as for other advocates of relativism.

None of this is to deny that science is a social activity, that scientists have interests other than the 'purely cognitive,' or that the sociological study of science is an eminently worthwhile undertaking – it is; they do; and it is. The question concerns not the viability or worth of the sociological investigation of science, but only the tendency of such investigation to support relativism. If my arguments in this section succeed, it does not.[28]

AMBIVALENCE CONCERNING RELATIVISM?: MACINTYRE, PUTNAM, AND RORTY

The work of three highly visible and influential philosophers – Alasdair MacIntyre, Hilary Putnam, and Richard Rorty – is intimately involved in the relativism controversy. That they regard the issue as central to their larger philosophical projects is indicated by the fact that their respective Presidential Addresses to the Eastern Division of the American Philosophical Association all involve relativism.[29] Interestingly, all three explicitly reject relativism; yet all three are frequently interpreted as relativists by their commentators and critics. While I cannot here enter into a full discussion of these philosophers' positive views, a word about their places in the relativism controversy is I hope in order.

a) MacIntyre

MacIntyre's 1987 contains one of the best anti-anti-relativism lines I know of: "...relativism, like skepticism, is one of those doctrines that have by now been refuted a number of times too often. Nothing is perhaps a surer sign that a doctrine embodies some not-to-be-neglected truth than that in the course of the history of philosophy it should have been refuted again and again. Genuinely refutable doctrines only need to be refuted once." (385)

In his efforts "to capture the truth in relativism" (387), MacIntyre explicitly rejects the doctrine as defined above on the grounds that it cannot meet the Socratic challenges to it (386-7). Nevertheless, he claims to find in relativism an important truth: "[W]hen we learn the languages of certain radically different cultures, it is in the course of discovering what is untranslatable in them, and why, that we learn not only how to occupy alternative viewpoints, but in terms of those viewpoints to frame questions to which under certain conditions a version of relativism is the inescapable answer." (404) This is so, MacIntyre argues, because any language which affords us the ability to frame such questions itself precludes us from finding non-relative reasons for rationally preferring one such viewpoint to another. Why? Because, MacIntyre argues, in any such language there will be, "for the relevant kinds of controversial subject matter, all too many heterogeneous and incompatible schemes of rational justification. And every attempt to advance sufficient reasons for choosing any one such scheme over its rivals must always turn out to presuppose the prior adoption of that scheme itself or some other. For without such a prior pre-rational commitment, no reason will count as a good reason." (405) Consequently, MacIntyre suggests, it will not be possible to find "any genuinely neutral and independent standard of rational justification." (405)

MacIntyre's argument for this 'inescapable' (in the "certain conditions" which he specifies, 387-405) version of relativism sounds very much like another instance of the confluence of the 'no neutrality' and 'no transcendence' arguments we have seen before: any proposed 'good reason' presupposes a 'prior pre-rational commitment' to some standard which sanctions that reason as a good one; that standard cannot itself be advanced on the basis of good reasons, since one's commitment to it is 'pre-rational.' That is to say, there is no possibility of good reasons which are in the relevant sense 'genuinely neutral'; and there is no possibility that one's 'pre-rational commitment' can be 'transcended' and given rational support. Insofar as the version of relativism MacIntyre defends here depends upon these arguments, it is, as we have seen, problematic.

But MacIntyre's attitude towards this version of relativism is somewhat unclear, since he claims that it can be 'transcended.' (405) Such transcendence involves the recognition, from within one's particular conceptual framework, that that very framework faces problems it cannot solve, but that an alternative framework is preferable to it in that the latter framework can both resolve the problems facing the former, and explain the former's failure to resolve them as well. Indeed, according to MacIntyre, this recognition, from within one's framework, that another can be superior, reveals "a central characteristic of theoretical and practical rationality" (408) which itself identifies a framework-neutral (in the sense explained above) standard of framework adequacy; and this standard, which involves the requirements of "[r]ationality ...*qua* rationality" (408), consitutes a clear rejection of relativism. The historicism which MacIntyre articulates and defends throughout his work rests upon what seems clearly to be, in MacIntyre's hands, this ahistorical and non-relativistic standard. (Siegel 1997, pp. 216-17 n31.) So despite his excellent pro-relativist-sounding line cited above, and his claim to have captured 'the truth in relativism,' MacIntyre cannot be counted among the friends of relativism in the sense we are considering it here.[30]

b) Putnam

Putnam (1981, 54-55, 119-124, 157-162; 1982, 7-14; 1983, 288; 1990, 125-126, 139-141; and many other places as well) rejects relativism even more emphatically than MacIntyre, arguing with energy and passion that it falls, as Plato claimed (and MacIntyre agrees), to the now standard self-refutation/incoherence arguments against it. Indeed, in several of these works Putnam has offered interesting and original variations on the incoherence theme. (See esp. the cited passages in his 1981 and 1982.)

Nevertheless, though apparently staunchly anti-relativist, Putnam has sometimes been seen as a relativist because his 'internalism' – which rejects any "God's eye point of view" (1981, p. 50), and insists both that fundamental metaphysical and epistemological questions "only make... sense to ask *within* a theory or description" (1981, p. 49, emphasis in original) and that "'objects' do not exist independently of conceptual schemes" (1981, p. 52) – has seemed to some to involve some sort of objectionable relativism, since these passages suggest that Putnam embraces some combination of the 'no neutrality' and 'no transcendence' arguments for relativism discussed above.[31]

I will not try here to determine whether Putnam's internalism amounts to or entails a problematic form of relativism. It does not matter for present purposes how that question is resolved, although it is worth emphasizing that resolving it in a way that makes Putnam a relativist depends upon attributing to him endorsement of one or both of the problematic arguments for relativism discussed above. The important points in this context are simply that Putnam, like MacIntyre, (a) explicitly rejects relativism, (b) explicitly endorses as decisive the standard self-refutation/ incoherence arguments against it, (c) if a relativist at all, is one in virtue of his embrace of arguments for relativism which, as argued above, do not succeed, and so (d) offers no aid or comfort to relativism.

c) Rorty

Like both MacIntyre and Putnam, Rorty famously rejects relativism, when understood as "the view that every belief is as good as every other." (1989, 37; see also 1982a, 166) Indeed, Rorty agrees that "relativism is self-refuting" (1991, 202), on the basis of the familiar self-refutation/incoherence arguments canvassed above. (1982a, 167) He ridicules this sort of relativism – which is the sort, verbal quibbles aside, with which we have been concerned throughout – as one which "[n]o one holds": "Except for the occasional cooperative freshman, one cannot find anybody who says that two incompatible opinions on an important topic are equally good." (1982a, 166) He articulates here and elsewhere his positive view, called 'pragmatism' and characterized in terms of 'solidarity' rather than 'objectivity,' as his preferred alternative to relativism.[32]

However, Rorty equally famously accepts claims which seem clearly to commit him to a version of relativism suspiciously similar to the sort he rejects. For example, he holds that his pragmatist can, "in the process of playing vocabularies and cultures off against each other,...produce new and better ways of talking and acting"; but that what is in this way produced is "not better by reference to a previously known standard, but just better in the sense that they come to *seem* [to this pragmatist] clearly better than their predecessors." (1982, p. xxxvii, emphasis in original) (Here we hear echoes of the 'no neutrality' argument for relativism rehearsed above.) Moreover, as this same passage suggests, Rorty consistently rejects any appeal to non-'ethnocentric' standards or criteria, in accordance with which disputes about truth or warrant might be rationally resolved, in favor of "criterionless muddling through" (1989, 43; see pp. 40-43), which, in view of the discussion above of the need for some such standards if relativism is to be avoided, again suggests relativism. (Here the 'no transcendence' argument seems clearly in play.) In the same vein, Rorty rejects the possibility of non-relativist debate concerning the merits of old vs. new vocabularies, on the basis of the (Heideggerian *and* Davidsonian!) point that there is no higher order vantage point available from which to referee such debates (Rorty 1989a, 50 ff.). (Here again Rorty invokes the 'no neutrality' and 'no transcendence' arguments.) He says, further, that he "view[s] warrant as a sociological matter, to be ascertained by observing the reception of *S*'s statement by her peers" (1993, 449), which seems clearly enough to relativize warrant, in distinctly Protagorean-sounding fashion, to the judgment of one's fellow community members: if one's statement is judged by one's peers to be warranted, it *is* warranted. (Here both the 'no neutrality' and 'no transcendence' arguments are suggested.) Finally, as MacIntyre shrewdly observes, Rorty's rejection of relativism seems clearly incompatible with his efforts, in the name of pragmatism, to render "is true" as "seems true to such and such persons, namely *us*" (MacIntyre 1987, 386, emphasis in original; for Rorty's discussion of his 'pragmatic' conception of truth, see, e.g., his 1982, xxiii-xxix, or his 1989, 37-39), since that rendering, and Rorty's pragmatist view of truth more generally, is itself open to those same self-refutation arguments: "[T]he premises from which Plato derived Socrates' refutation of Protagoras' version of relativism also entailed the necessary failure of any reinterpretive reduction of 'is true' to 'seems true to such and such persons.' From these premises the one conclusion is not available without the other." (MacIntyre 1987, 387)

What are we to make of all this? I will not here attempt to sort out Rorty's status as a relativist. We may accept his own proclamations that he is not; or we may try to show that, despite those proclamations, his general views of rationality, truth, criteria, our locatedness in history, etc., commit him to a version of relativism which he claims to eschew. This interpretive task is not essential to the present inquiry. Rather, the important points to notice are that (a) Rorty accepts the self-refutation arguments as decisive against relativism of the sort with which we are concerned; (b) insofar as Rorty accepts relativism, such acceptance seems clearly enough to be based upon his acceptance of some combination of the 'no neutrality' and 'no transcendence' arguments criticized above; and (c) relativism, consequently, derives no aid, comfort or support from Rorty's varied musings on the topic.

It is worth reiterating that although MacIntyre, Putnam, and Rorty all advocate positive views not always easily distinguished from relativism (MacIntyre's 'historicism,' Putnam's 'internalism,' Rorty's 'pragmatism'), all three firmly reject relativism. Their determination to grapple in this serious way with the relativism issue, trying hard to distinguish versions of relativism which fall to the standard self-refutation arguments from versions which capture important philosophical insights, is I think one explanation for the widespread interest in the work of these three distinguished figures. It is also worth noting how the three have tried to clarify their own views by contrasting them with the views of the other two – sometimes by criticizing one or both of the other two for embracing a problematic form of relativism – thus establishing a complex network of cross-criticism which is worth sustained study.[33]

All three of these thinkers are clearly involved in the effort to find a way to acknowledge and avoid the defects of overly strong forms of 'absolutism,' while at the same time avoiding the defects of relativism. Whether or not their efforts succeed – that is, whether or not Rorty's pragmatism, Putnam's internalism, or MacIntyre's historicism prove to be successful articulations of non-relativistic epistemologies – I leave for the reader to judge. The important thing to note, for present purposes, is that, as these three important contemporary thinkers explicitly claim, relativism must be avoided, on pain of self-refutation and incoherence. To the extent that any of their positive stories are committed to relativism, as many of their critics allege, those stories are, even from their authors' points of view, defective. Consequently, the work of Putnam, Rorty and MacIntyre does not help the relativist cause, however relativistically that work is interpreted and understood.[34]

IF NOT RELATIVISM, WHAT?: THE SHAPE OF A DEFENSIBLE 'ABSOLUTISM'

To briefly summarize the case against relativism presented thus far: I have suggested that the more specific arguments for relativism considered above all rely on either the 'no neutrality' or the 'no transcendence' (or both) arguments for relativism; I have also tried to indicate why these arguments – specific and general – are problematic.[35] Whether or not they can be repaired sufficiently to overcome those problems I leave the reader to judge. Moreover, it is worth repeating that, even if one or more of these arguments for relativism can be adequately repaired, it will still face the incoherence problem considered earlier. How can the relativist regard one of these arguments, or indeed any argument, as rationally compelling – or supportive of its conclusion to any degree – given her denial of non-relative standards of evaluation, appeal to which is required in order to establish such rational compulsion or support? In endorsing one or another of these arguments *as* rationally compelling or supportive, such that it *ought* to be found (at least to some degree) persuasive by fair-minded students of the issue, the relativist seems forced to give up her commitment to relativism, according to which no arguments or standards have probative force beyond the bounds of the communities which endorse them. On the other hand, to acknowledge that these arguments have force only for such communities, the relativist explicitly acknowledges that she has no reason to offer which should persuade her opponent to give up her non-relativist position and switch to the relativist's camp, or which should persuade the fair-minded student of the issue to join that camp. Thus, whatever the ultimate fate of the arguments for relativism we have considered, the relativist still faces the hoary and deep problem of incoherence.

Given the apparent intractability of this fundamental problem, how should we understand the continuing philosophical appeal of the doctrine? As is so often the case in philosophy, relativism benefits from the problems facing its main alternative. The contemporary resurgence and continuing appeal of relativism is at least in part due to the difficulty of formulating a defensible conception of 'absolutism' (understood simply as the contradictory of relativism). In addition to the arguments for relativism just reviewed, many relativists argue for relativism on the grounds that any non-relativistic alternative will require repugnant epistemological commitments, e.g., to certainty, privileged frameworks, or dogmatism. And it must in fairness be granted that in the long and complex history of epistemological consideration of relativism in the Western tradition, anti-relativists, from Plato to Descartes and beyond, have indeed often supposed that avoiding relativism requires the embrace of one or another of these unpalatable alternatives.

Although this is no doubt something of an historical oversimplification, it is only relatively recently (historically speaking) that epistemologists have recognized that one can reject both relativism and certainty (and dogmatism, privileged frameworks, and other epistemological evils historically associated with absolutism), and opt instead for a *fallibilistic* absolutism. While both doctrines reject the idea – characteristic of historically important (e.g., Cartesian) absolutist epistemologies – that knowledge requires certainty, fallibilism differs from relativism in that the former holds (while the latter denies) that, in the sense explained above, non-relative evaluations of knowledge-claims can be made.[36]

Still, while saying this is easy enough, it must nevertheless be admitted that a fully developed absolutism remains to be articulated and defended. The challenge to opponents of relativism is to develop a non-relativistic, 'absolutist' epistemology, which includes an acceptable account of rationality and rational justification (as the discussions above of Kuhn, Barnes and Bloor, MacIntyre, Putnam and Rorty suggest),[37] which is fallibilistic and non-dogmatic, which rejects any notion of a privileged framework in which knowledge-claims must be couched, and which is self-referentially coherent. This is obviously a philosophically demanding task, which involves many of the most fundamental issues of epistemology.[38]

Given the difficulty of formulating a satisfactory version of absolutism, and the understandable and quite justified unwillingness of relativists to embrace unsatisfactory versions of it – i.e., those which endorse certainty, privileged frameworks, dogmatism, and other epistemically noxious views historically associated with it – it should come as no surprise that activity on both sides of the relativism/absolutism controversy remains high. A further explanation for the continued intense interest in this issue is that there is much at stake: it (along with rationality, with which it is entwined) can be seen as the most basic epistemological issue of all, since, whichever side is correct, the outcome of the dispute has enormous implications for epistemology generally. For how we are to understand the full range of fundamental epistemological issues (e.g., those treated at length in this *Handbook*), and what counts as success in resolving them, depends to a significant extent upon the resolution of the relativism/absolutism issue.

Take, for instance, truth. Philosophers defend and criticize a range of theories here: correspondence, coherence, pragmatic, redundancy, etc.[39] If absolutism in some form is correct, disputes concerning alternative theories of truth are appropriately understood as disputes concerning truth *simpliciter*, i.e., independently of (not relativized to) persons, cultures, communities, conceptual schemes, frameworks, historical epochs, etc. On the other hand, if relativism is correct, theorists of truth cannot be happily understood as offering *general* theories of truth, since what counts as the correct or most adequate theory of truth will be relative to one or another of these relativizing variables or contexts. How epistemologists understand epistemological controversy concerning truth, consequently, depends upon how the relativism issue is ultimately resolved. The same point can be made with respect to all other matters of epistemological moment. In this sense, the relativism issue is as fundamental an issue as there is in epistemology.[40]

Given the fundamental nature of the issue, and the formidable difficulties facing those on either side of it, it is safe to predict that the controversy will not be put to rest any time soon.[41]

Harvey Siegel
University of Miami

NOTES

[1] There are many sorts of relativism other than the epistemological sort – ontological, moral, axiological, cultural, etc. – which are not addressed here. I typically do not use the modifier 'epistemological' in what follows, but the reader should assume, unless it is explicitly indicated otherwise, that all mentions of 'relativism' below are to epistemological relativism. For discussion of other varieties of relativism, see (e.g.) Krausz and Meiland 1982 and Harré and Krausz 1996.

[2] The two main versions of epistemological relativism may thus (following Knorpp 1998) be labeled *alethic* relativism, i.e. relativism with respect to truth, and *justificatory* relativism, i.e. relativism with respect to justification. It is perhaps worth noting that the relativity of *belief* – the third of the three 'standard' conditions of knowledge, along with truth and justification – is at most a trivial aspect of epistemological relativism, although it is rightly seen as central to *cultural* (or anthropological (or sociological)) relativism. Indeed, the most basic way to distinguish between these latter two sorts of relativism is to note that the latter requires only the uncontroversial relativity of belief and cultural practice, while the former requires the philosophically more contentious relativity of truth and/or justification as well – and hence, at least on most accounts of it, of knowledge itself. For this reason I will not in what follows discuss the cultural relativism associated with Sumner, Benedict, and other cultural anthropologists. Relativism inspired by the sociology of knowledge is discussed below.

[3] Authors use these terms – 'principles,' 'criteria' and 'standards' – somewhat idiosyncratically, to refer to that which is (allegedly) the ultimate source of relativism. Although the terms can be distinguished from one another in various ways, in what follows I use them more or less interchangeably in order to honor the preferences of the authors discussed.

[4] For a somewhat more technical definition, see Siegel 1987, p. 6. Similar definitions are offered by Krausz and Meiland (1982, 'Introduction,' p. 8), Krausz (1989, 'Introduction,' p. 1), and Bayley (1992, 'Introduction,' p. 2). Authors are generally agreed that relativism is a matter of the relativity of evaluative *standards* or *criteria* governing judgments of knowledge, truth, and justificatory status. Many definitions are cited and discussed in Siegel 1987, e.g. those of Brown (p. 10), Doppelt (p. 90), Field (p. 26), Goodman (p. 150), Popper (p. 33), and Weinert (p. 33); virtually all of them are clearly in keeping with the definition proposed in the text.

It is important to distinguish relativism from related but distinct epistemological positions with which it is frequently confused, e.g., skepticism, fallibilism, pluralism, nihilism, etc. For an instructive guide to the relevant distinctions, see Knorpp 1998.

[5] I use this term in deference to Nelson Goodman's (1978) unique form of relativism; I regret that I cannot consider Goodman's version of relativism further here. For discussion of Goodmanian (relativistic) 'rightness,' and Goodman's 'radical relativism within rigorous restraints' more generally, see Siegel 1987, ch. 7.

[6] As Levin (1992, 72) engagingly puts it, in attempting to defend relativism, the relativist commits "dialectical suicide."

[7] The relativist can respond by denying that she is engaged in the project of defending relativism, and asserting that relativists have other purposes in mind when arguing for relativism. For discussion and references, see Siegel 1987, 21-23. She can also respond by holding that the argument purporting to demonstrate the incoherence of relativism just rehearsed in fact begs the question against the relativist by presupposing an 'absolute' conception of truth. For discussion, see the following four paragraphs in the text; also Siegel 1987, 23-25.

[8] Joseph Margolis 1991 defends what he regards as a form of alethic relativism. But he grants (pp. 9-12 and *passim*) that relativism, as defined above (which he calls 'relationism'), falls to the standard incoherence arguments against it. The view he defends appears to be not that truth is relative, but that we ought to reject ***tertium non datur*** and so embrace more than the usual two truth values. I will not discuss Margolis' idiosyncratic treatment further here.

[9] Hales 1997 is an interesting attempt to show how appropriate logical machinery can be utilized to avoid the self-refutation charge and establish "a consistent relativism," which holds not that "everything is relative" but that "everything true is relatively true." (pp. 33-4) This paper aims to establish the consistency of alethic relativism; it addresses mainly metaphysical and semantic problems rather than the epistemological ones I have been belaboring thus far, and does not (as Hales notes) speak directly to the epistemic status of either relative truths or the arguments offered in defense of the consistency of this version of relativism. Hales' version of relativism, according to which truths are relative to 'perspectives,' raises deep questions which are discussed below in terms of the 'no neutrality' and 'no transcendence' arguments for relativism but which are not discussed in detail by Hales. It is, further, unclear that that version escapes the 'impotence' problem just discussed. Finally, I must point out that Hales' definition of 'relativism' (and of 'absolutism'), on analogy with modal terms, is sufficiently non-standard that it is unclear how the relativism whose consistency Hales claims to establish is related to relativism as defined above. For these reasons I do not pursue his discussion further here, though I happily acknowledge the originality of his approach and the contribution to the metaphysical and semantic issues addressed in his paper which it makes.

[10] This is obviously a very superficial account of 'The Galileo Affair,' offered here for illustrative purposes only. For a more serious treatment, see, e.g., Finocchiaro 1989.

[11] It might perhaps be thought that such aid or comfort might flow from the recognition that, while there may well be neutral standards which the parties to a given dispute will (or should) acknowledge as relevant to the rational resolution of that dispute, it is nevertheless the case that such resolution is often *underdetermined* by the available evidence, even granting such shared standards. It must I think be granted that shared standards will often be insufficient to resolve such disputes; even granting shared standards, the resolution of such disputes will be underdetermined by the total available evidence. But this sort of underdetermination is not sufficient to secure relativism, since many disputes will not be so underdetermined. To move from underdetermination to relativism, one must argue both that all disputes are in fact underdetermined, and, moreover, that they are necessarily so. But the first is false – e.g., 'eye color in humans is genetically determined' is not underdetermined by the evidence – and the second (as well as the first) is unavailable to the relativist, given relativism's impotence (discussed above). In order to secure relativism on the basis of underdetermination, the relativist must claim to have (non-relativistic) good reason to believe that all disputes will inevitably be underdetermined by all relevant evidence, but it is unclear how the relativist can claim this in a way consistent with her relativism. In other words, the phenomenon of underdetermination can (and should) be acknowledged; such acknowledgement is perfectly consistent with the rejection of relativism. The anti-relativist need not (and should not) say that the existence of shared standards is in and of itself sufficient to preclude underdetermination in all cases.

[12] That is, this argument, if it did appear to be epistemically forceful, would still have to face the incoherence charge already discussed: how could a relativist regard the argument (or any other) as epistemically forceful, given her rejection of the possibility of non-relativistic epistemic forcefulness? I won't pursue this point further here, but the fundamental problem of incoherence plagues this and all other cases for relativism.

[13] Children typically attain 'a reasonable grasp of whole numbers' by age three or four. Grasp of fractions and decimals usually involves a process which extends over several years and is in part a function of what is taught, when. The classic work in this area is Gelman and

Gallistel 1978; it (including their account of what counts as a 'reasonable grasp' of numbers) is summarized briefly and lucidly in Moshman, Glover and Bruning 1987, 420-423. Thanks to David Moshman for helpful advice on matters concerning psychological development.

[14] For critical discussion of Popper's view, and of 'framework relativism' more generally, see Siegel 1987, ch. 2; for consideration of this issue in the context of arguments for/against naturalized epistemology, see Siegel 1995, esp. pp. 50-1; for more general discussion of the possibility of 'transcendence,' see Siegel 1997.

This path to relativism also arises in the context of stage theories of psychological (or 'conceptual' or 'foundational') development. For discussion, see the exchanges between van Haaften 1990, 1993 and Siegel 1993, and van Haaften and Snik 1997 and Scheffler 1997.

[15] Davidson's argument is criticized, though not in a way which lends support to relativism, in Siegel 1987, 38-42.

[16] 'Highly influential' is something of an understatement. At the time of this writing (early 1998) approximately one million copies of Kuhn's book have been sold; translations into twenty-five languages have been authorized. It 'revolutionized' the philosophy of science, dramatically invigorated the (at the time) rather quiet scholarly study of the history and historiography of science, virtually created the intense contemporary interest in the sociology of science, and continues to be studied not only in these fields but across the humanities, the social sciences and the natural sciences. It seems safe to say that, in light of its dramatic impact upon the philosophy, history and sociology of science, its influence across a much wider range of scholarly disciplines, and the degree to which its basic concepts (e.g. 'paradigm') have seeped into every day discourse, ***Structure*** will prove to be among a very small number of the most influential philosophy books of the second half of the twentieth century.

[17] The secondary literature on Kuhn's philosophy of science is vast; I can only nod at it here. Siegel 1987, chs. 3-5 critically discusses Kuhnian philosophy of science, especially as it bears upon issues concerning relativism and the (ir)rationality of science. The papers in Gutting 1980 and Horwich 1993 (including Kuhn's 'Afterwords,' 311-341) include important analyses, clarifications and mostly sympathetic criticisms of Kuhn's work; the former provides some sense of the wide range of disciplines Kuhn's work has influenced. Hoyningen-Huene 1993 provides an impressively researched and imaginatively conceived (although in some respects philosophically controversial) systematic interpretation of Kuhn's philosophy of science, one which Kuhn himself warmly endorses in his 'Foreword.'

[18] See esp. Hoyningen-Huene 1993. It is unfortunate that while Hoyningen-Huene's thorough discussion touches on relativism at many points, the topic is not systematically discussed in the book, and 'relativism' does not appear in the index.

[19] Of course there are also important disagreements on many other aspects of Kuhn's views. For a very broad and detailed consideration of the enormous critical response to Kuhn, see Hoyningen-Huene 1993.

[20] There are of course some who both praise relativism and find support for it in Kuhn's philosophy of science; this view is not uncommon among sociologists of science, especially those sympathetic to the 'Strong Programme' (discussed below).

[21] I will not consider Feyerabend's complex version of relativism here; for discussion, see Siegel 1989.

[22] It is perhaps worth noting that Quine, in whose writings may be found important early statements of many of these allegedly relativism-supporting themes (in particular, those of holism, indeterminacy and underdetermination), has in recent years qualified/weakened his view to such an extent that it can no longer be seen (if it ever could) as lending any support to relativism. (1991, 268-272; 1992, 13-16, 93-102) Further, underdetermination of a sort strong enough to yield relativism is effectively challenged in Laudan and Leplin 1991 (and in Laudan 1990, 54-6). The route to relativism which passes through the undeterdetermination

thesis – roughly, 'For any evidence set *E* which supports theory *T*, *E* will equally strongly support rival, empirically equivalent but incompatible theories *T*′, *T*″, etc.; hence all such rivals to *T* are as well supported by *E* as *T* itself is; hence there can be no epistemic grounds for preferring one of these rivals to another; hence relativism' – contains insuperable roadblocks.

[23.] It is worth noting that one of the protagonists of this section, Barry Barnes, has explicitly endorsed this distinction, and agreed that the sociologist's project is distinct from the epistemologist's: "The sociologist is concerned with the naturalistic understanding of what people take to be knowledge, and not with the evaluative assessment of what deserves to be so taken; his orientation is normally distinct from that of the philosopher or epistemologist." (Barnes 1977, 1) This acknowledgement does not sit well with his paper with Bloor (1982) discussed below, since that paper emphatically rejects this distinction, and advocates epistemological relativism on the basis of the sociologist's concern with 'what people take to be knowledge' – indeed, far from these being two distinct projects, the epistemological point, according to that paper, follows directly from the sociological ones.

This traditional distinction has come under attack in ways other than those to be addressed here. In particular, *causal* accounts of knowledge and justification, and *naturalistic* accounts more generally, have played important roles in recent decades. I regret that I can consider only matters related to relativism in what follows; for further discussion of causal/naturalistic theories in epistemology, see the articles by Bloor, Bradie, Lammenranta, Schmitt, and Shope in this *Handbook*. Criticism of naturalized epistemology/philosophy of science is offered in Siegel 1995, 1996, 1996a, and 1998.

[24.] See references to Barnes, Bloor, Hollis, Lakatos, Laudan, Lukes, and Mannheim at Barnes and Bloor 1982, 26; and several papers in Brown 1984, esp. those by Bloor, Brown, Gutting, and Laudan.

[25.] On the last point see Gutting 1984, 106; Elster 1982, 147; and McCarthy 1989. A simple example: a given student's belief that the distance between the sun and the Earth is approximately 93,000,000 miles (on average) might be caused by her conducting an experiment to measure the distance, combined with her instructor's approval of her experimental procedure; it might instead be caused by her having learned and memorized that figure in her childhood, and her unconscious 'fudging' of the experiment to produce that result. Both of these are possible causal explanations of her belief. That they differ in epistemic status (presumably, the first would go some way toward justifying her belief, while the second would not) is irrelevant to the question of their (in)correctness as causal explanations of the belief. Both explanations of the belief's credibility are causal; but this leaves open the epistemic status of the belief. Thus the equivalence postulate offers no support to relativism. Barnes and Bloor would of course disagree, and reject the distinction between epistemic status and perceived epistemic status. But this rejection depends upon their version of the defective 'no transcendence' argument for relativism.

[26.] This brings to mind Kuhn's famous remark that, with respect to paradigm choice, "there is no standard higher than the assent of the relevant community." (1962, 94) Here is one clear instance of Kuhn's influence on the strong programme in particular, and on post-Kuhnian sociology of science more generally.

[27.] For the record, Barnes and Bloor do talk about "the received culture of epistemologists" (1982, p. 39); there is nothing unfair in characterizing their view in such a way that specific academic groups – e.g., epistemologists, sociologists, and even particular 'schools' within these groups – constitute their own local communities which can be investigated sociologically in order to determine the causes of the credibility of their belief systems.

[28.] I must acknowledge that my discussion is open to the charge of being out of date: while Barnes and Bloor's 1982 is perhaps the classic defense of relativism from the perspective of the sociology of science, there are many more recent discussions available which defend, presuppose, or critically discuss this route to relativism. The work of Collins, Edge, Gooding, Knorr-Cetina, Latour, Lynch, Mulkay, Pickering, Pinch, Shapin, and Woolgar spring immediately to mind; there are course many other important practitioners in this remarkably active field. I regret that I cannot treat this work here. For references and further philosophical critique of its tendency to support relativism, see Slezak 1994, 1994a and 1994b; for additional references and sociological critique, see Cole 1992.

[29.] Rorty's (1982a) and MacIntyre's (1987) Presidential Addresses centrally concern relativism. Putnam's Presidential Address (1977) concerns relativism only tangentially; it is mainly concerned to criticize 'metaphysical realism' and to articulate and defend his 'internal realism.' But since it is that doctrine of Putnam's that is often taken to be relativistic, it seems fair to regard his Address as also highly relevant to the relativism controversy. It is worth noting that a fourth highly influential figure, Donald Davidson, also devoted his Presidential Address to the Eastern Division (1973) to the issue of relativism. (I discuss Davidson's Address in my 1987, 38-42.)

[30.] This tension between MacIntyre's embrace of a thoroughgoing historicism and his insistence on his proposed traditional-neutral, ahistorical standard in accordance with which rival traditions can themselves be fairly evaluated occurs elsewhere in his work as well. See, e.g., MacIntyre 1988, 7-10, 346-403. Here the ahistorical character of his standard is put as follows: "It is in respect of their adequacy or inadequacy in their responses to epistemological crises that traditions are vindicated or fail to be vindicated" (366); judgments concerning such (in)adequacy and (non)vindication are not, on MacIntyre's positive view, tradition-relative, but are rather a function of features of '[r]ationality... *qua* rationality.' Whether or not he succeeds in resolving this tension I leave the reader to judge. But MacIntyre himself is clear that in his view his historicism offers no aid or comfort to relativism, which he clearly means to reject (e.g., 366-7).

[31.] See Siegel 1987, 176 n64 for brief further references. Putnam's more recent discussions of 'conceptual relativity' (1990, x-xi) might also contribute to the perception that his view is in the end relativist.

[32.] Rorty defends this alternative mainly in terms of his historicist rejection of the possibility of achieving an ahistorical standpoint – of stepping "outside our skins" (1982, xix) – thus bringing to mind the 'no neutrality' and especially the 'no transcendence' arguments for relativism examined earlier. I cannot here enter into a full scale discussion of Rortian pragmatism. For brief criticism and further references, see Siegel 1997, 174-5. To Rorty's suggestion that 'no one holds' a relativistic view, I can only recommend the wide range of relativist literature cited above.

[33.] A few of many examples: MacIntyre criticizes Rorty in his 1987, 387, and 1990, 710-711. Putnam criticizes Rorty in his 1981, 216, in 1982, 9-12, in 1990, ix, 19-29, in 1992, 67-71, and elsewhere. Rorty criticizes Putnam in 1989, p. 39, at many points in 1991, and in 1993, *passim*. (Putnam and Rorty are interestingly compared in Forster 1992.) Rorty criticizes MacIntyre in 1991a, 158-163. No doubt important light will be shed on this vexing cluster of issues, which has relativism at its center, in future studies of these three-way disputes. It seems an obvious topic for future doctoral dissertations.

[34.] Richard Bernstein's influential 1983 also deserves brief comment. In it Bernstein bequeaths to the debate the notion of "Cartesian anxiety." (16-20 and *passim*) Those who have this anxiety "quest for some fixed point, some stable rock upon which we can secure our lives against the vicissitudes that constantly threaten us"; this quest is motivated by "[t]he specter that hovers in the background of this journey [, which] is not just radical epistemological skepticism but the dread of madness and chaos where nothing is fixed, where we can neither touch bottom nor support ourselves on the surface.... *Either* there is some support for our being, a fixed foundation for our knowledge, *or* we cannot escape the forces of darkness that envelop us with madness, with intellectual and moral chaos." (18, emphases in original) Bernstein urges us to reject (or overcome) Cartesian Anxiety, and, in so doing, to get 'beyond objectivism and relativism.'

It must be granted that Bernstein's discussion usefully relates the literature we have been considering to the hermeneutical tradition, in particular the work of Heidegger, Dilthey, and especially Gadamer. His discussion is wide-ranging and often insightful. But it should be noted, first, that there is a looseness to his characterizations of the positions he discusses which renders his account of the issue somewhat unhelpful. For example, Bernstein's 'objectivism' is not equivalent to 'absolutism.' In the passage just cited, he explicitly conflates 'objectivism' with 'foundationalism' and 'relativism' with 'radical epistemological skepticism,' and implicitly conflates 'absolutism' with 'objectivism.' In the course of the book these conflations, and others, are quite clear. Bernstein's conception of 'relativism,' moreover, is far more general than the epistemological relativism being treated here.

Second, it is unclear whether 'Cartesian Anxiety' helps much. To be Cartesianly anxious is to have a certain dread, fear, or anxiety concerning the possible absence of a certain foundation. But being anxious in this way is being in a particular psychological state; while pointing out that one who wishes to avoid relativism is in that state might help to explain that person's tendency to embrace 'objectivism,' it seemingly has no tendency to discredit that view (or to support or discredit either its relativist contrary, or whatever it is we get to when we get 'beyond' both objectivism and relativism). To reason in this way – 'she is (Cartesianly) anxious, therefore her objectivism is misguided' – is straightforwardly to commit the freshman-level fallacy of psychologizing, i.e., of evaluating the epistemic status of a belief in terms of the psychological state or motivation of the believer. A person's anxiety concerning relativism, however genuine, has no tendency to undermine either her arguments against it or her arguments for a non-relativist alternative. Whether or not Bernstein is himself guilty of this fallacy I leave to the reader to judge. (I think it is clear that others who appeal to Cartesian Anxiety in their arguments against 'foundationalism' (which, again, is not equivalent to 'absolutism') do indeed commit it.)

Third, Bernstein's 'Cartesianly anxious' philosopher appears to be bothered by her inability to achieve an overly strong sort of neutrality or transcendence. But (as we have seen) the non-relativist need not and should not aspire to such overly strong forms of these – weaker forms of neutrality and transcendence are both available and sufficient to block relativism. In this respect, Bernstein's discussion trades on the equivocations central to the 'no neutrality' and 'no transcendence' arguments discussed above. For these reasons, I do not think that Bernstein's book, despite its undeniable strengths, significantly advances our understanding of the issues being considered here.

[35.] While I have not the space to consider them in detail, I want to point out that a further cluster of relativistic positions – those associated with the later Wittgenstein and his treatment of 'forms of life,' with the related idea of (inescapable) 'conceptual schemes' and the impossibility of judging such schemes 'from the outside,' and with the anthropological work of Evans-Pritchard on the Azande and the huge philosophical literature it provoked – also rely on the 'no transcendence' argument for relativism, and typically on the 'no neutrality' argument as well. Insofar as they do so rely (I regret being unable here to establish that they do in detail), they are likewise deficient. For the arguments see especially Winch 1958 and Winch 1964 (and references to Wittgenstein therein); for discussion for and against see the essays in Wilson 1970 and in Hollis and Lukes 1982. (See also Siegel 1987, 178 n37.)

[36.] Crucial to the articulation and defense of fallibilism is the work of Peirce 1931-58; for lucid discussion and references see Scheffler 1974, pp. 42-57. Popper's *conjectural* account of knowledge (1963, 1972) has also been influential in the fostering of a widespread acceptance of fallibilism.

[37] For further discussion of the interanimation of relativism and rationality, see Siegel 1987, ch. 8.

[38] Of value in clarifying the character of an acceptable form of absolutism are Harré and Krausz' (1996) distinctions among three different sorts of absolutism – 'universalist,' 'foundationalist,' and 'objectivist' – which they use with some effectiveness (despite their somewhat imprecise handling of those three key terms) to delineate alternative varieties of relativism in terms of their rejection of one or more of these varieties of absolutism, and to evaluate the many positions thus delineated.

[39] I am ignoring here *relativistic* theories of truth, which have been discussed above and which are not central players in mainstream philosophical treatments of truth. For further discussion and references, see the essay by David in this *Handbook*.

[40] I note in passing that working epistemologists, at least in the analytic tradition, generally presume absolutist understandings of the issues – perception, memory, induction, rational belief change, etc. – on which they work. This seems clear from the way in which epistemology is generally presented and practiced, from introductory textbooks to sophisticated, 'cutting edge' journals. In this sense absolutism is the 'default' position in epistemology. No doubt there are compelling historical explanations of this, in terms of the influence of the Greeks in establishing the agenda of the Western philosophical tradition. That absolutism is the default position of most practicing epistemologists is not, of course, in itself a powerful argument for the correctness of that view. For further discussion of the relation between the relativism/absolutism issue and other epistemological issues, see Siegel 1987, ch. 8, esp. pp. 165-6.

[41] There are of course many serious works on relativism other than those mentioned above. I regret that I cannot consider them here, but I have listed some of them in the following bibliography. Many of the individual articles in the several anthologies listed especially merit attention. Thanks to Ilkka Niiniluoto for suggestions on an earlier draft.

REFERENCES

Barnes, B.: 1977, *Interests and the Growth of Knowledge*, Routledge & Kegan Paul, London.

Barnes, B. and D. Bloor.: 1982, 'Relativism, Rationalism and the Sociology of Knowledge', in Hollis and Lukes (eds.), pp. 21-47.

Bayley, J. E. (ed.): 1992, *Aspects of Relativism: Moral, Cognitive, and Literary*, University Press of America, Lanham, MD.

Baynes, K., J. Bohman, and T. McCarthy (eds.): 1987, *After Philosophy: End or Transformation?*, The MIT Press, Cambridge, MA.

Bernstein, R. J.: 1983, *Beyond Objectivism and Relativism: Science, Hermeneutics, and Praxis*, University of Pennsylvania Press, Philadelphia.

Bloor, D.: 1976, *Knowledge and Social Imagery*, Routledge & Kegan Paul, London.

Brown, J. R.: 1984, *Scientific Rationality: The Sociological Turn*, D. Reidel, Dordrecht.

Cole, S.: 1992, *Making Science: Between Nature and Society*, Harvard University Press, Cambridge, MA.

Davidson, D.: 1973, 'On the Very Idea of a Conceptual Scheme', *Proceedings of the American Philosophical Association* **47**, 5-20.

Davson-Galle, P.: 1998, *The Possibility of Relative Truth*, Ashgate, Aldershot.

Elster, J.: 1982, 'Belief, Bias and Ideology', in Hollis and Lukes 1982, pp. 123-148.

Feyerabend, P.: 1975, *Against Method: Outline of an Anarchist Theory of Knowledge*, Verso, London.

Finocchiaro, M. A. (ed.): 1998, *The Galileo Affair: A Documentary History*, University of California Press, Berkeley.

Forster, P. D.: 1992, 'What Is at Stake Between Putnam and Rorty?', *Philosophy and Phenomenological Research* **52**, 585-603.

Gelman, R. and C. R. Gallistel: 1978, *The Child's Understanding of Number*, Harvard University Press, Cambridge, MA.

Goodman, N.: 1978, *Ways of Worldmaking*, Hackett Publishing Company, Indianapolis, IN.

Gutting, G. (ed.): 1980, *Paradigms and Revolutions: Applications and Appraisals of Thomas Kuhn's Philosophy of Science*, University of Notre Dame Press, Notre Dame, IN.

Gutting, G.: 1984, 'The Strong Program: A Dialogue', in Brown 1984, pp. 95-111.

Hales, S. D.: 1997, 'A Consistent Relativism', *Mind* **106**, 33-52.

Harré, R. and M. Krausz: 1996, *Varieties of Relativism*, Blackwell Publishers Ltd., Oxford.

Harris, J. F.: 1992, *Against Relativism: A Philosophical Defense of Method*, Open Court, LaSalle, IL.

Hesse, M.: 1980, 'The Strong Thesis of Sociology of Science', ch. 2 of Hesse, *Revolutions and Reconstructions in the Philosophy of Science*, Indiana University Press, Bloomington, pp.29-60.

Hollis, M., and Lukes, S. (eds.): 1982, *Rationality and Relativism*, The MIT Press, Cambridge, MA.

Horwich, P. (ed.): 1993, *World Changes: Thomas Kuhn and the Nature of Science*, The MIT Press, Cambridge, MA.

Hoyningen-Huene, P.: 1993, *Reconstructing Scientific Revolutions: Thomas S. Kuhn's Philosophy of Science*, The University of Chicago Press, Chicago, trans. A. T. Levine.

Knorpp, W. M., Jr.: 1998, 'What Relativism Isn't', *Philosophy* **73**, 277-300.

Kordig, C. R.: 1971, *The Justification of Scientific Change*, D. Reidel, Dordrecht.

Krausz, M., (ed.): 1989, *Relativism: Interpretation and Confrontation*, University of Notre Dame Press, Notre Dame, IN.

Krausz, M., and J. W. Meiland (eds.): 1982, *Relativism: Cognitive and Moral*, University of Notre Dame Press, Notre Dame, IN.

Kuhn, T. S.: 1962, *The Structure of Scientific Revolutions*, University of Chcago Press, Chicago. (Second edition, enlarged, 1970.)

Kuhn, T. S.: 1970, 'Reflections on My Critics', in Lakatos and Musgrave 1970, pp. 231-278.

Lakatos, I., and A. Musgrave (eds.): 1970, *Criticism and the Growth of Knowledge*, Cambridge University Press, Cambridge.

Laudan, L.: 1988, 'Are All Theories Equally Good? A Dialogue', in Nola 1988, pp. 117-139.

Laudan, L.: 1990, *Science and Relativism: Some Key Controversies in the Philosophy of Science*, University of Chicago Press, Chicago.

Laudan, L., and J. Leplin: 1991, 'Empirical Equivalence and Underdetermination', *The Journal of Philosophy* **88**, 449-472.

Levin, M.: 1992, 'Reality Relativism', in Bayley 1992, pp. 69-85.

MacIntyre, A.: 1987, 'Relativism, Power, and Philosophy', in Baynes, Bohman, and McCarthy 1987, pp. 385-411; originally published in 1985, *Proceedings and Addresses of the American Philosophical Association* **59**, 5-22.

MacIntyre, A.: 1988, *Whose Justice? Which Rationality?*, University of Notre Dame Press, Notre Dame, IN.

MacIntyre, A.: 1990, 'Review of Richard Rorty, Contingency, Irony, and Solidarity', *Journal of Philosophy* **87**, 708-711.

Margolis, J.: 1991, *The Truth About Relativism*, Blackwell, Oxford.

McCarthy, T.: 1989, 'Contra Relativism: A Thought Experiment', in Krausz 1989, pp. 256-271.

Melchert, N.: 1994, *Who's To Say?: A Dialogue on Relativism*, Hackett, Indianapolis.

Monist **67**(3): 1984, 'Is Relativism Defensible?'

Moshman, D., J. A. Glover, and R. H. Bruning: 1987, *Developmental Psychology: A Topical Approach*, Little, Brown and Company, Boston.

Nagel, T.: 1986, *The View From Nowhere*, Oxford University Press, Oxford.

Nola, R.: 1988, *Relativism and Realism in Science*, Kluwer, Dordrecht.

Norris, C.: 1997, *Against Relativism: Philosophy of Science, Deconstruction and Critical Theory*,
Blackwell, Oxford.

O'Gorman, F.: 1989, *Rationality and Relativity: The Quest for Objective Knowledge*, Avebury, Aldershot.

Papineau, D. (ed.): 1996, *The Philosophy of Science*, Oxford University Press, Oxford.

Peirce, C. S.: 1931-58, *Collected Papers*, volumes I-VI edited by C. Hartshorne and P. Weiss, volumes VII-VIII edited by A. W. Burks, Harvard University Press, Cambridge, MA.

Plato: 1961, *Theaetetus*, trans. F. M. Cornford, in E. Hamilton and H. Cairns (eds.), *The Collected Dialogues of Plato*, Bollingen Series, Pantheon Books, Random House, New York, pp. 845-919.

Popper, K. R.: 1963, *Conjectures and Refutations*, Routledge & Kegan Paul, London.

Popper, K. R.: 1970, 'Normal Science and Its Dangers,' in Lakatos and Musgrave 1970, pp. 51-58.

Popper, K. R.: 1972, *Objective Knowledge: An Evolutionary Approach*, Clarendon Press of Oxford University Press, Oxford.

Putnam, H.: 1977, 'Realism and Reason', *Proceedings of the American Philosophical Association* **50**, 483-498.

Putnam, H.: 1981, *Reason, Truth and History*, Cambridge University Press, Cambridge.

Putnam, H.: 1982, 'Why Reason Can't Be Naturalized', *Synthese* **52**, 3-23.

Putnam, H.: 1983, *Realism and Reason: Philosophical Papers*, Volume 3, Cambridge University Press, Cambridge.

Putnam, H.: 1990, *Realism with a Human Face* (edited and introduced by J. Conant), Harvard University Press, Cambridge, MA.

Putnam, H.: 1992, *Renewing Philosophy*, Harvard University Press, Cambridge, MA.

Quine, W. V.: 1960, *Word and Object*, MIT Press, Cambridge, MA.

Quine, W. V.: 1991, 'Two Dogmas in Retrospect', *Canadian Journal of Philosophy* **21**, 265-274.

Quine, W. V.: 1992, *Pursuit of Truth* (revised edition), Harvard University Press, Cambridge, MA.

Rorty, R.: 1979, *Philosophy and the Mirror of Nature*, Princeton University Press, Princeton, NJ.

Rorty, R.: 1982, *Consequences of Pragmatism*, University of Minnesota Press, Minneapolis, MN.

Rorty, R.: 1982a, 'Pragmatism, Relativism, and Irrationalism', in Rorty 1982, 160-175. Originally published in 1980, *Proceedings of the American Philosophical Association* **53**, 719-738.

Rorty, R.: 1989, 'Solidarity or Objectivity?', in Krausz 1989, pp. 35-50.

Rorty, R.: 1989a, *Contingency, Irony, and Solidarity*, Cambridge University Press, Cambridge.

Rorty, R.: 1991, *Objectivity, Relativism, and Truth: Philosophical Papers*, Volume 1, Cambridge University Press, Cambridge.

Rorty, R.: 1991a, *Essays on Heidegger and Others: Philosophical Papers*, Volume 2, Cambridge University Press, Cambridge.

Rorty, R.: 1993, 'Putnam and the Relativist Menace', *The Journal of Philosophy* **90**, 443-461.

Scheffler, I.: 1967, *Science and Subjectivity*, Bobbs-Merrill, Indianapolis. (Second edition, Hackett Publishing Company, Indianapolis, 1982.)

Scheffler, I.: 1974, *Four Pragmatists: A Critical Introduction to Peirce, James, Mead, and Dewey*, Humanities Press, New York.

Scheffler, I.: 1997, 'Reply to van Haaften and Snik', in Siegel 1997a, pp. 260-261.

Shapere, D.: 1964, 'The Structure of Scientific Revolutions', *The Philosophical Review* 73, 383-94.

Siegel, H.: 1987, *Relativism Refuted: A Critique of Contemporary Epistemological Relativism*, D. Reidel Publishing Company, Dordrecht.

Siegel, H.: 1989, 'Farewell to Feyerabend', *Inquiry* **32**, 343-69.

Siegel, H.: 1992, 'Relativism', in J. Dancy and E. Sosa (eds.), *A Companion to Epistemology*, Basil Blackwell, Oxford, pp. 428-430.

Siegel, H.: 1993, 'Justifying Conceptual Development Claims: Response to van Haaften', *Journal of Philosophy of Education* **27**, 79-85.

Siegel, H.: 1995, 'Naturalized Epistemology and "First Philosophy"', *Metaphilosophy* **26**, 46-62.

Siegel, H.: 1996, 'Naturalism and the Abandonment of Normativity', in W. O'Donohue and R. Kitchener (eds.), The Philosophy of Psychology, Sage, London, pp. 4-18.

Siegel, H.: 1996a, 'Instrumental Rationality and Naturalized Philosophy of Science', *Philosophy of Science* **63**, Supplement (PSA Proceedings, Part 1), S116-S124.

Siegel, H.: 1997, *Rationality: Redeemed? Further Dialogues on an Educational Ideal*, Routledge, New York.

Siegel, H. (ed.): 1997a, *Reason and Education: Essays in Honor of Israel Scheffler*, Kluwer, Dordrecht.

Siegel, H.: 1998, 'Naturalism and Normativity: Hooker's Ragged Reconciliation', *Studies in History and Philosophy of Science* **29**, 639-652.

Slezak, P.: 1994, 'Sociology of Scientific Knowledge and Scientific Education: Part I', *Science & Education* **3**, 265-294.

Slezak, P.: 1994a, 'Sociology of Scientific Knowledge and Scientific Education. Part II: Laboratory Life Under the Microscope', *Science & Education* **3**, 329-355.

Slezak, P.: 1994b, 'The Social Construction of Social Constructionism', *Inquiry* **37**, 139-157.

van Haaften, W.: 1990, 'The Justification of Conceptual Development Claims', *Journal of Philosophy of Education* **24**, 51-69.

van Haaften, W.: 1993, 'Conceptual Development and Relativism: Reply to Siegel', *Journal of Philosophy of Education* **27**, 87-100.

van Haaften, W., and G. Snik.: 1997, 'Critical Thinking and Foundational Development', in Siegel 1997a, pp. 19-41.

Wilson, B. R. (ed.): 1970, *Rationality*, Basil Blackwell & Mott, Ltd., London.

Winch, P: 1958, *The Idea of a Social Science and its Relation to Philosophy*, Routledge & Kegan Paul, London.

Winch, P.: 1964, 'Understanding a Primitive Society', *American Philosophical Quarterly* **1**, 307-324.

Zellner, H.: 1995, 'Is Relativism Self-Defeating?', *Journal of Philosophical Research* **20**, 287-295.

JAN WOLEŃSKI

ANALYTIC VS. SYNTHETIC AND A PRIORI VS. A POSTERIORI

1 INTRODUCTION

The division of human cognitive faculties into those based on reason and those based on experience belongs to the standard epistemological vocabulary. The controversy between empiricism and rationalism, which is one of the most important in epistemology, is organized around these categories. Both parties occur in their genetic and methodological versions. Within the former version, we have nativism (genetic rationalism) and genetic empiricism, but apriorism (methodological rationalism) and aposteriorism (methodological rationalism) are connected with the latter.[1] This chapter deals mainly with apriorism and aposteriorism, although their connections with the genetic issue will be also noted. The distinctions pointed out in the title are usually regarded as helpful in explaining how apriorism and empiricism are related. In particular, since both views appear in radical or moderate versions, it is important to see where the borderline between them should be drawn.

When we speak about methodological aspects of cognition, we take into account various things on which apriorism and aposteriorism debate. Typically, justifications, concepts and sentences (propositions, beliefs, statements, judgements, etc.) play a central role in those discussions. Roughly speaking, justifications are processes, activities or simply acts of a kind, sentences receive support from justifications or they do not, and, finally, some concepts, at least in Kantian tradition, make knowledge possible.[2] Thus, labels 'based on reason' and 'based on experience' may be directed, disjunctively or jointly, to justifications, sentences or concepts . Philosophers attribute various properties to activities performed by reason (for instance, deductive inferences) and their results (truths of reason). They are declared to be independent of experience, universal, necessary, certain and infallible. Similarly, philosophers declare that activities and statements based on experience are revisable, probable, contingent, uncertain and fallible. However, some philosophers argue that no essential difference between performances of reason and those of experience occurs at all. Also, it happens that reason is dethronized, being merely considered as an auxiliary device for knowledge principally organized by senses. On the other hand, we encounter attempts at attributing certainty, necessity or universality to results of experience.

Thus, at a very preliminary stage, we face a fairly complicated picture of relations between reason and experience, their differences and dependencies. Apart from questions concerning features of knowledge based on reason and those based on experience, we encounter, for example, the question: Is sensory knowledge possible without participation of reason? The reverse question also gained much attention. It was Kant who said that reason without senses is empty, but senses

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 781-839.
© 2004 *Kluwer Academic Publishers.*

without reason are blind. Clearly, both these questions are relevant for genetic empiricism and nativism as well. Yet we have problems suggested by ways of knowing realized in particular fields. Is logic analytic, that is, is it based on meaning relations that hold between constituents of sentences? Is mathematics a priori? Is physics a posteriori? Are there a priori ethical principles? In fact, formal sciences, that is, logic and mathematics always served as the pattern of the empire of reason, and natural science as provided the model how experience works.

Perhaps it will suffice to show in a very condensed way how the distinctions 'analytic vs. synthetic'(AS for brevity) and 'a priori vs. a posteriori' (AP for brevity) can contribute to epistemology.[3] Of course, AS and AP are interesting in themselves because they are related to the basic problems of logic, semantics, philosophy of mathematics or philosophy of science. Our distinctions operate on various levels (see Moser 1987, Boghossian 1997, Bealer 1999). AP is mainly epistemological, but AS is explicated in semantics in the broad sense, that is, as including syntax, semantics proper (the theory of referential relations involved in language) and pragmatics. On the other hand, both distinctions are closely related, and it is shown by their respectable history. A quite popular view sees AS and AP as extensionally identical (it is the so called linguistic theory of the a priori in a preliminary characterization), but there are several objections against this position. To some extent, the recent discussion about AP and AS can be largely reduced to an exchange between proponents of the linguistic theory and its critics.

I will begin with AS as perhaps elaborated in the most complete way. I will review different conceptions of analytic sentences and objections directed against them. Then, I will pass to AP which combined with AS is central for epistemology.[4] And finally, one terminological question. It is customary, that 'analytic a priori' sentence' is an abbreviation for 'analytically (a priori) true sentence'. It is a simplification due to the fact that our qualifications, i.e., 'analytic' and 'a priori' are also applicable for false sentences. Thus, if one says that '*A* is an analytic sentence if and only if, e.g., *A* is true in such and such circumstances', a related definition of synthetic sentences cannot be constructed as: *A* is a synthetic sentence if and only if *A* is not analytic. Rather, we should say: *A* is synthetic if and only if *A* is neither analytically true nor analytically false. However. I will follow the traditional way of speaking in most cases for its convenience. Moreover, it is not difficult to derive correct definitions of analytic falsities and synthetic statements from particular proposals concerning analytic truths. Although I will report various views about AS and AP in order to give a comprehensive survey of related discussions, I find it difficult to abstain from expressing my own position. Roughly speaking, it consists in a defence of moderate empiricism, that is, the view that the analytic and the a priori can be coherently embedded into empiristic epistemology. I hope that this attitude does not obscure that other solution are also arguable.

2. ANALYTIC VS. SYNTHETIC

Disregarding anticipations, the great philosophical career of the concept of analyticity began with Kant, although he was conscious that he had predecessors.[5] For Kant, the linguistic form '*S* is *P*' represents the most general structure of sentences. Two principal passages introducing the concept of analyticity in Kant's *Critique of Pure Reason* run as follows (1 replace letters *A* and *B* used by Kant by *S* and *P* respectively):[6]

> (Kl) "In all judgements in which the relation of a subject to the predicate is thought (I take into consideration affirmative judgement only, the subsequent application to negative judgements being easily made), this relation is possible in two different ways. Either the predicate *P* belongs to the subject *S*. as something which is (covertly) contained in this concept *S* or *P* lies outside the concept *A*, although it does stand in connection with it. In the one case I entitle the judgement analytic, in the other synthetic. Analytic judgements (affirmative) are therefore those in which the connection of the predicate with the subject is thought through identity; those in which this connection is thought without identity should be entitled synthetic. The former, as adding nothing, through the predicate to the concept of the subject, but merely breaking it up into those constituent concepts, although confusedly, can also be entitled explicative. The latter, on the other hand, add to the concept of the subject a predicate which has not been in any wise thought in it, and which no analysis could possibly extract from it; and they may therefore be entitled ampliative. If I say, for instance, 'All bodies are extended', this is an analytic judgement. For I do not require to go beyond the concept which I connect with 'body' in order to find extension as bound up with it. To meet with this predicate, I have merely to analyse the concept, that is, to become conscious to myself of the manifold which I always think in that concept. The judgement is then analytic. But when I say, 'All bodies are heavy', the predicate is something quite different from anything that I think in the mere concept of body in general; and the addition of such a predicate therefore yields a synthetic judgement." (Kant 1781, 48-49). (K2) "For, *if the judgement is analytic*, whether negative or affirmative, its truth can always be adequately known with accordance in the principle of contradiction. [...]. *The principle of contradiction* must therefore be recognized as being completely sufficient *principle of all analytic knowledge*." (Kant 1781, 190; all italics in quotation are given by particular authors).

A slightly different explanation given by Kant in his (1800, 117-118) is this:

> "*Analytic* propositions one calls those propositions whose certainty rests on *identity* of concepts (of the predicate with the notion of the subject). Propositions whose truth is not grounded on identity of concepts must be called *synthetic*,
>
> The identity of concepts in analytic judgements can be either *explicite (explicita)* or *nonexplicit (implicita)*, In the former case analytic propositions are *tautological*.
>
> [...] Tautological propositions are *virtualiter* empty or *void of consequences*. Such is, for example, the tautological proposition, *Man is man* [...].
>
> Implicitly identical propositions, on the contrary, are not void of consequences, for they clarify the predicate which lay undeveloped *(implicite)* in the concept of the subject through *development (explicatio)*"

There are various problems connected with Kant's understanding of analyticity.[7] The first question concerns the relation between (Kl) and (K2). Are they equivalent or not? According to (Kl), a sentence is analytic if and only if its predicate *P* is contained in its subject *S*. Let this containment be symbolized by $S \geq P$.[8] Hence, a synthetic sentence is characterized by the connection $S <$ P. However, it is clear that the case in which *S* and *P* exclude each other is omitted. Kant clearly sees it, when he says that (Kl) concerns affirmative analytic sentences. A more explicit statement of this point is in his (1783, 14):

> "[...] All analytic judgments depend wholly upon the law of contradiction [...] For the predicate of an affirmative analytical judgement is already contained in the concept of the subject of which it cannot be denied without a contradiction. Such is the nature of the judgement "All bodies are extended," and "No bodies are unextended (that is, simple)".

It is confirmed by the quoted fragment of *Critique* in which Kant makes a reservation that *P* lies(?) outside *S*, but it is related to the subject. At first sight, (Kl) and (K2) are equivalent with respect to affirmative propositions. However, if we take Kant's remark that *P* is outside *S* in synthetic propositions literally, the problem with negative propositions arises. Consider the sentence (a) '*S* is not-*S*'' which is analytic and grammatically affirmative. However, its subject and predicate exclude each other and the latter is entirely outside the former. Thus, there are affirmative analytic sentences which are not based on the containment of the predicate in the subject, but grounded in the exclusion holding between them. Apparently, one can say that (a) is equivalent to (b) 'No *S* is *S*' that is, to a negative sentence. But it does not account for the identity of nominal constituents of (b). Since (K2), as Kant explicitly notes, is applicable to affirmative as well as negative analytic statements, it presumably determines a wider class of analytic sentences than (Kl), although Kant did not probably realize this consequence.[9] Another problem arising with Kant's definition concerns generality of his treatment of '*S* is *P*' as the most general logical form. It is a common objection against Kant that his treatment is not sufficiently general, relative to possible linguistic forms of sentences. This objection does not seem very serious because one can claim that the analyticity of compound sentences is always reducible to their constituents, and that every constituent has the subject-predicative form.

Hitherto, I interpreted Kant *via* logical tools. However, there is something else, perhaps even more important, in Kant's definitions of analyticity than their formal-logical aspects.[10] He uses, particularly in (Kl), psychological and epistemological language. We read that connections between *S* and *P* are thinkable or that we exceed or not the logical subject of sentences. Due to the generally antipsychological attitude of the 20th century philosophy, Kant is often accused of psychologism. I will not argue for or against this accusation. On the other hand, it is important to see that Kant's explanations go immediately into the foundations of cognitive activities. When Kant speaks about the explicative role of analytic sentences and synthetic statements as extending our knowledge, he obviously assumes that, behind formal linguistic structures, there are also definite cognitive faculties responsible for acts of analysis or synthesis as well as their results, that is, analytic and synthetic sentences. This bridge enabled Kant to map epistemology onto AS (and AP). Kant also linked analyticity and syntheticity with particular fields. Logic is analytic in Kant's view, but mathematics and theoretical physics are synthetic. Of course, it is not everything in Kant's view: further qualifications require an appeal to AP.

The next important step in the history of analyticity was made by Bolzano. Although, for historical reasons, Bolzano did not influence philosophy to a great extent, he is worthy of being be mentioned in the present context for his anticipation of semantic treatment of analyticity. Bolzano understands propositions as consisting of ideas. He points out that replacing ideas by other ideas sometimes leads to truth and sometimes to falsity. However, there is also another case:

> “(1) [...] But suppose that there is a *single* idea in it which can be arbitrarily varied without disturbing its truth or falsity, i. e. if all propositions produced by substituting for this idea any other idea we pleased are either true altogether or false altogether, presupposing only that they have denotation. This property of the proposition is already sufficiently worthy of attention to differentiate it from all those propositions for which this is not the case. I permit myself, then, to call propositions of this kind, borrowing an expression from Kant, *analytic*. All the rest however, i. e. in which there is not a single idea that can be arbitrarily varied without affecting their truth and falsity, I call *synthetic propositions*. So, for example, I shall call the propositions, “A morally evil man deserves no respect” and “a morally evil man nevertheless enjoys eternal happiness,” a pair of analytic propositions. In both of them there is a certain idea, namely the idea of a man, for which you may substitute any idea you please, c. g. angel, being, etc., in such a way that the first (if only has denotation) is true and the second false in every case. On the other hand, I could not point to a single idea in the proposition, “God is omniscient” and “a triangle has two right angles,” which could be arbitrarily varied with the result that the former would remain constantly true and the latter constantly false. Consequently for me, these would be examples of synthetic propositions.
>
> (2) We have some very general examples of analytic propositions which are also true in the following propositions [in the original we have ‘proposition’, but it is obviously a misprint; in German original, we have *an folgenden Sätzte* the same remark concerns the next sentence – J. W.]: *A* is *A*, *A* which is *B* is *A*, *A* which is *B* is *B*, Every object is either *B* or not *B*. etc. The propositions of the first type, or those included under the form, *A* is *A* or *A* has (the property) a, we are used to identifying by a name of their own as *identical* or *tautological* propositions.
>
> (3) The examples of analytic propositions I have just cited are differentiated from those given in (1) by the fact that nothing is necessary for judging the analytic nature of the former besides logical knowledge, because the concepts that make up the invariant part of these propositions all belong to logic. But judging the truth or falsity of propositions like those in (1) requires quite another kind of knowledge, because concepts alien to logic exert an influence in them. To be sure, this distinction has its ambiguity, because the domain of concepts belonging to logic is not so sharply demarcated that no dispute could ever arise over it. At times it could be useful to pay attention to this distinction, and so we could call propositions such as those in (2) *logically* analytic or analytic in the *narrower* sense and those in (1), on the other hand, analytic in the *broader* sense.” (B. Bolzano 1839, 192-193)[11]

In spite of Bolzano’s worries that the stock of logical concepts is fuzzy, his idea of logical analyticity is clear: *A* is logically analytic if and only if its truth value is invariant under substitutions (more precisely: replacements) of non-logical constituent. On the other hand, Bolzano’s account of analytic sentences in the broader sense is much more vague and puzzles commentators.[12] In particular, it is not clear how far invariance goes in the case of such propositions, in particular, whether it is restricted at most or at least to one single idea. Moreover, one may wonder why the sentence (a) ‘A morally evil man deserves no respect’ is analytic, but the sentence (b) ‘A triangle has two right angles’ is not. It seems that (a) is a general analytic principle relatively to ‘man’, that is, it is invariant under substitutions of ‘man’ by something else, but not relatively to ‘morally evil’ and ‘respect’. However, it does not explain the difference of roles of ‘man’ as contrasted with ‘morally evil’ and ‘respect’. There must be something else except substitutions that generates analyticity in the broader sense. Definitions? Conceptual connections? Well, but (b) can also be reduced to definitions and conceptual dependencies. I will not enter into possible interpretations of Bolzano. What is certainly important is that he observed various kinds of analyticity.

Frege contributed to our problem in a double way. First, like Kant, he associated the problem of analytic sentences with the nature of logic and mathematics. Second, he gave a definition of analyticity:

> "It, in carrying out this process, we come only on general logical laws and on definitions, then the truth is an analytic one, bearing in mind that we must take account also of all propositions upon which the admissibility of any of the definitions depends. If, however, it is impossible to give the proof without making use of truths which are not of a general logical nature, but belong to the sphere of some special science, then the proposition is a synthetic one." (Frege 1884, 4)

Frege's conception of analyticity is justification-oriented. A sentence A is analytic if and only if it is provable with the help of laws of logic and definitions, but a sentence is synthetic if such a proof is impossible. Three points should be noted. First, Frege assumes in advance that laws of logic are analytic. Secondly, there is an unclarity concerning the meaning of the term 'definition'. One interpretation is that any (of course, formally correct) definition can be used in analytic proof, but another and weaker one consists in admitting only those definitions which are formulated in logical terms. For instance, the definition '$A \vee B = \neg A \Rightarrow B$' is an example of a purely logical definition, but 'Bachelor is an unmarried man' belongs to a wider category of definitions.[13] This question is legitimized by the fragment of Frege which says that synthetic proofs make appeal to "truths which are not of a logical nature", and, as we will see, it is not without significance. The quoted passage also contains a hint how to explain analyticity of logical laws. Since non-logical truths belong to special fields, we can conclude *a contrario* that logical principles have universal application. Frege endorsed this view in other of his writings (see Frege 1885). In order to complete Frege's account let me remind that logic and mathematics, except geometry, were identical for him. Geometry does not belong to logic for its appeal to particular spatial relations, i. e., it is not fully general.[14]

The next important step in the development of the concept of analyticity was made by Russell and Wittgenstein. Russell extended Frege's logicism to entire mathematics, including geometry. He also shared Frege's view that mathematics is analytic.[15] However, the crystallic purity of Frege's construction was damaged by the Russell paradox. The new situation required introducing new constructions (the theory of logical types) with certain artificialities like the axiom of reducibility, mysterious from the point of view of their analytic character. Also, certain mathematical axioms, namely the axiom of infinity and the multiplication axiom (equivalent to the axiom of choice) raised some doubts as analytic sentences. Russell's first solution consisted in "if-thenism", that is, a view that we accept implications of the type $A \Rightarrow B$, where A is an axiom (e.g., the axiom of infinity) and B a derived theorem. However, this solution dissatisfied Russell because it tacitly assumed non-analytic character of antecedents. Russell (1919) characterizes logical truths as tautologies, although without any general definition of this concept:

> "The law of contradiction is merely one among logical propositions; it has no special pre-eminence [...]. Nevertheless, the characteristic we are in search of is the one which was felt, intended to be defined, by those who said that it consisted in deducibility from the law of contradiction. This characteristic, [...] for the moment, we may call *tautology* [...]." (203)

"The complete asserted proposition of logic will all be such that some propositional function is *always* true." (204)

"It is clear that the definition of "logic" or "mathematics" must be sought by trying to give a new definition of the old notion of "analytic" propositions. Although we can no longer be satisfied to define logical propositions as those that follow from the law of contradiction, we can and must still admit that they are a wholly different class of propositions from those that we come to know empirically. They all have the characteristic which, a moment ago, we agreed to call "tautology". This, combined with the fact that they can be expressed wholly in terms of variables and logical constants (a logical constant being something which remains constant in a proposition even when *all* its constituents are changed) – will give the definition of logic or pure mathematics. For the moment, I do not know how to define "tautology". [..] It would be easy to offer a definition which might seem satisfactory for a while; but I know of none that I feel to be satisfactory, in spite of feeling thoroughly familiar with the characteristic of which a definition is wanted." (204-205)

In a footnote, Russell expressed his debt to Wittgenstein:

"The importance of "tautology" for a definition of mathematics was pointed to me by my former pupil Ludwig Wittgenstein." (205)

However, it is disputable whether Wittgenstein accepted that mathematics consists of tautologies, although he defended the tautological character of logic.[16] There is a sample of Wittgenstein's view taken from his 1922:

"4.46 Among the possible groups of truth-conditions there are two extreme cases.

In one case the proposition is true for all the truth possibilities of the elementary propositions. We say that the truth-conditions are *tautological,*

In the second case the proposition is false for all truth-possibilities. The truth-conditions are *self-contradictory.*

In the first case we call the proposition a tautology, in the second a contradiction.

4.461 The proposition shows what it says, the tautology and the contradiction that they say nothing.

The tautology has no truth-conditions, for it is unconditionally true; and the contradiction is on no condition true.

Tautology and contradiction are without sense.

(Like the point from which two arrows go out in opposite directions.) (I know, *e.g.* nothing about the weather, when I know that it rains or does not rain.)

4.4611 Tautology and contradiction are, however, not nonsensical; they are part of the symbolism, in the same way that "0" is part of the symbolism of Arithmetic.

[...]

6.1 The propositions of logic are tautologies.

6.11 The propositions of logic say nothing. (They are analytic propositions.)

6.112 The correct explanation of logical propositions must give them a peculiar position among all propositions.

6.113 It is the characteristic mark of logical propositions that they are true; and this fact contains in itself the whole philosophy of logic. And so also it is one of the most important facts that the truth of falsehood of non-logical propositions can *not* be recognized from the propositions alone.

6.12 The fact that the propositions of logic are tautologies *shows* the formal – logical – properties of language, of the world.[...]"

Literally taken, Russell and Wittgenstein agree that all tautologies are analytical propositions. However, the converse implication is not clear. Nothing can be derived from Wittgenstein about that. It seems that Russell also maintained that all analytic propositions are tautologies (see Russell 1948, 154-155). Nevertheless, for its historical influence, particularly for the early Vienna Circle, we can simplify the Russell-Wittgenstein view about analyticity to the equality: analytic propositions = tautologies.[17]

Numerous definitions of analyticity were proposed after Russell and Wittgenstein. There is a list of suggestions including also older ideas:[18]

(1) *A* is analytic if and only if *A* is true in all possible worlds.
(2) *A* is analytic if and only if *A* could not possibly be false.
(3) *A* = '*S* is *P*' is true if and only if *A* attributes to *P* no more content than is already contained in *P*.
(4) *A* is analytic if and only if its negation is a contradiction.
(5) *A* is analytic if and only if *A* is true by virtue of meanings and independently of facts.
(6) *A* is analytic if and only if either *A* is logically true or it can be reduced to a logical truth by replacing synonyms by synonyms.
(7) *A* is analytic if and only if *A* is true under every state-description.
(8) *A* is analytic if *A* can be reduced to a logical truth by definition.
(9) *A* is analytic in **L** if and only if *S* is true according to semantic rules of **L.**
(10) *A* is analytic if and only if *A* is true invariantly of substitutions for its extralogical constituents.
(11) *A* is analytic in **L** if and only if the sentence '*A* is necessary' is true in **L.**
(12) *A* is analytic if and only if *A* is true by virtue of a relation between intensional meanings.
(13) *A* is analytic if and only if *A* is a tautology.

Historically speaking, (1) and (2) go back to Leibniz, (3) and (4) are Kantian (see (Kl) and (K2) above), (5) is Ayer's (1946, 105) and Lewis' (1946, 39) precizations of the traditional account, (6) is Quine's (1951) reconstruction in his criticism of the concept of analyticity, (7) is Carnap's (1950, 83), (8) is Fregean, (9) is in Carnap (1947, 10), (10) is Bolzano's definition of logical truth, later revived by Quine (1936, 73-74), (11) and (12) are taken from Lewis (1946, 89-90), and (13) is Russellian-Wittgensteinian (see above). Of course, there are several other possibilities of interpreting particular authors, but it is not very fruitful enterprise to enter into further hermeneutical moves. Perhaps the most important point concerns relations between items listed in (1)-(13). The matter is not simple, even in the case of proposals obvious at the first sight. One can claim that (1), (2), (4), (10), (11) and (13) are mutually equivalent. However, it requires at least: (a) assuming in advance that (1) and (13) are equivalent, and (b) interpreting 'necessity' as 'logical necessity'. Other cases are much more complicated and depend on understanding such concepts as, e.g., semantical rule, meaning, definition, synonymity, etc.

Now I pass to some recent suggestions (formulated in the 20th century) how analyticity could be defined, including also some comments in most cases (others are to be found in works quoted in notes).
(14) Schlick (1918, 74)

A is analytic if and only if it ascribes to a subject a predicate that is already contained in the concept of the subject.

"Contained" in this definition means that "the predicate is part of the definition of the subject."

(15) Dubislav (1926, 23)

Given a system of assumptions and kinds of justifications *X*, *A* is analytic if and only if its truth or falsity is demonstrable merely by use of elements of *X*.

Dubislav regards this definition as an improvement of ideas of Bolzano and Frege. Although Dubislav's definition does not decide about the nature of justification, examples given by him suggest that analyticals are results of deduction.

(16) Carnap (1934, 39)

A is analytic if and only if *A* is a consequence of the null class of sentences. Formally: $A \in \mathrm{AN} \Rightarrow A \in Cn\varnothing$.

This definition concerns only the so-called Language I in Carnap's *Logical Syntax of Language.* This language covers arithmetic of natural numbers, but is too weak for the whole mathematics and theoretical physics. However, in order to overcome the first Gödel incompleteness theorem (roughly speaking, if S is a consistent formal system sufficient for first-order arithmetic of natural numbers, then S is incomplete, that is, it produces undecidable sentences), the consequence operation *Cn* in (16) is infinitistic as based on ω-rule, enabling to derive the sentence $\forall xPx$ from the infinite class P(1),P(2),P(3),.... The definition of analyticity for Language II is much more complicated and I skip it (see Kleene (1939) for a simplification of Carnap's original proposal). Carnap usually identified logical truth and analytic truth, but not analyticity and analytic truth, although (16) is an exception in this respect. He defined logical (analytic) determinacy as logical truth or logical falsity, and synthetic propositions as logically not determinate.[19]

(17) Carnap (1942, 60)

A is analytically true (L-true) if and only if *A* is true on logical grounds.

Carnap equates here analytic truth and logical truth. In fact, he construes the concept of L-truth by stating several postulates.

(18) Scholz (1944, 195-200)

A is analytic if and only if *A* is analytically justified.

The concept of analytic justification is crucial here. Scholz considers analyticity in scientific axiomatic theories which are divided by him into logical, mathematical and physical ones. Assume that J is a justification base. Now, *A* is analytic with respect to J if and only if every element of J is analytic. *A* is synthetic if and only if it is not analytic, that is, if at least one element of J is synthetically justified. Theorems of logic are analytic. It is guaranteed by intuition which maps logical truths into all possible worlds. Scholz points out that arithmetic of natural numbers is synthetic for its non-reducibility to pure logic.

(19) Gödel (1944, 138-139)

(a) *A* is analytic if and only if *A* is either a special case of the law of identity in virtue of explicit or implicit definitions of its constituents or *A* is disprovable as a negation of this law;

(b) *A* is analytic if and only if *A* holds in virtue of meanings of its constituents.

Gödel remarks that arithmetic is demonstrably non-analytic in the sense (a) because, otherwise, its analyticity would imply its decidability, contrary to the results of Church and Turing proving that arithmetic (and first-order logic) are undecidable. On the other hand, one can defend analyticity of mathematics in the sense (b). Moreover, according to Gödel, the sense (b) is more fundamental than (a), and this difference justifies to call the class generated by (b) analytic, but the class generated by (a) tautological. It resembles Kant's view (see above).

(20) Ajdukiewicz (1947, 174)

A is analytic in **L** if and only if *A* is axiomatically accepted in **L** or logically follows from the axioms of **L.**

'Axiomatically accepted' means 'accepted on the base of axiomatic rules of language'. This concept was introduced in Ajdukiewicz (1934).

(21) Kokoszyńska (1947, 37, 39)

A is analytic in **L** if and only if *A* is an analytic theorem or the negation of an analytic theorem in **L.**

There are different ways of accepting sentences. We assert some sentences according to empirical circumstances. Other procedures refer to structural forms of expressions. Call rules of the latter kind 'discursive'. Now, *A* is a necessary theorem of **L** if and only if *A* is accepted *via* discursive rules of **L.** Further, *A* is an analytic theorem of **L** if and only if .4 is a necessary theorem of **L** or *A* is derivable from necessary theorems of **L.** Kokoszynsńka claims that her proposals fit ordinary language.

(22) Langford (1949, 21)

A is analytic if and only if *A* is certified solely by reference to logical principles, that is, principles of the second-order logic.

Langford uses this definition in his proof that synthetic a priori sentences exist (see section 3 above).

(23) Copi (1949, 243)

A is analytic in **L** if and only if its truth or validity follows from the syntactical or grammatical rules of **L.**

Copi says that this definition is "current" and uses it in his argument for the existence of synthetic a priori sentences.

(24) Waismann (1949-1953, 134)

A is analytic if and only if *A* can be reduced to a logical truth merely with the help of definitions, logical operators and idiomatic (linguistic) operators.

Any reduction to analytic truth must be performed by consecutive equivalencies. At first, equivalencies can be based on logic, for instance, $A \vee B = \neg A \Rightarrow B$. Another kind is provided by definitions, for example, 'A planet is an object which moves around the sun'. The definitional equivalence enables us to eliminate terms from a given context. Now consider the sentence 'There is a planet that moves around the sun if and only if there is an object such that it is a planet that moves around the sun'. According to Waissman, it is an analytic truth based on an idiomatic (contextual) equivalence. Waissman intended to generalize Frege's account of analyticity.

(25) Reichenbach (1951, 17)

> "Consider the statement "all bachelors are unmarried". This statement is not very useful. If we wish to know whether a certain man is a bachelor, we must first know whether he is unmarried; and once we know it, the statement does not tell us anything else. [...] Statements of this kind are empty; they are called *analytic,* an expression which may be translated as *self-explanatory.*"

(26) Gödel (1951, 321)

> " [...] the basic axiom, or rather, axiom schema, for the concept of set of integers says that, given a well-defined property of integers (that is, a propositional expression $\varphi(n)$ with an integer variable n), there exists the set *M* of those integers which have the property φ. Now, considering the circumstances that φ may itself contain the term "set of integers", we have here a series of rather involved axioms about the concept of set. Nevertheless, these axioms [...] cannot be reduced to anything substantially simpler, let alone to explicit tautologies. It is true that these axioms are valid owing to the meaning of the term "set" – one might say they express the very meaning of the term "set" – and therefore they might fittingly be called analytic, however, the term "tautological", that is, devoid of content, for them is entirely out of place, because even the assertion of the existence of a concept of set satisfying these axioms (or the consistency of these axioms) is so far from being empty that it cannot be proved without again using the concept of set itself, or some other abstract concept of [[a]] similar nature. [...] "analytic" here does not mean "owing to our definitions", but rather "true owing to the nature of the concepts occurring [[therein]], in contradistinction to "true owing to the properties and the behaviour of things". This concept of analytic is so far from meaning "void of content" that it is perfectly possible that an analytic proposition might be undecidable (or decidable with [[a certain]] probability)."

Like in his (1944) (see (19) above), Gödel distinguishes here two concepts of analyticity. One (truth by virtue of definitions) is related to the concept of tautology, but the second (truth by the nature of concepts) should not be understood in the way that analytic sentences are void of content. The comprehension scheme in set theory is an example of analytic truth which is not tautological.

(27) Carnap and Bar-Hillel (1953, 229)

A is analytically true if and only if *IN(A),* that is, the set of information contained in *A* is the minimum in the set of all possible information-sets. Formally: $A \in \mathrm{AN}$ iff $IN(A) = \varnothing$.

This definition describes precisely the concept of analyticity as the emptiness of content. Now, *A* is analytically false if and only if *IN(A)* contains the maximal information, that is, it logically entails an arbitrary sentence. The success of this definition strongly depends on the availability of a good explication of the concept of semantic information. Unfortunately, there is no commonly accepted account of this idea.

(28) Carnap (1955, 39)

A is analytically true if and only if *A* is true on the base of meanings of logical constants and the relations between meanings of descriptive constants established by meaning postulates (A-postulates).

Meaning postulates concern descriptive expression of a language, that is, individual constants and predicates. A-postulates pick some models as admissible. Hence, analytic truths with respect to a set *X* of A-postulates are those which are true in all A-models generated by *X*. Logical truths in the strict sense, i.e., tautologies are true in all models. Carnap identified analyticity with logical truth in a wider sense and regarded both concepts as semantic. As it is documented by his (1963, 918), it was his final position: consult, however, (30) below for analyticity in theoretical systems, although Carnap's solution of this problem is coherent with his general account of analytic sentences.

(29) Kemeny (1956, 11)

A is analytically true if and only if *A* is true in all intended models.

This definition has the following rationale. Since arithmetic is incomplete, we cannot define analyticity by provability from the Peano axioms, because there are arithmetical truths which are not provable. The definition '*A* is analytically true if and only if *A* is true in all models' does not account for the existence of arithmetical truths that are not true in all models: since the Peano axioms are true in all arithmetical models and certain truths expressible in the language of arithmetic are unprovable, the latter cannot be true in all models. However, we are interested in truths holding in the standard (intended) models, that is, models that differ at most in interpretations of extralogical constants: it is important that arithmetical primitives are logical constants for Kemeny. This intuition is just captured by (19). It is interesting that Kemeny's definition is a semantic counterpart of (16). In fact, if we add ω-rule to formal arithmetic, we obtain the system in which all standard truths of arithmetic are provable. On Kemeny's account, truth in all models is a property of universal analytic sentences which are of course also valid in intended models.

(30) Carnap (1958, 246)

I begin with explanation of the problem. Let T be an empirical theory with theoretical terms $t_1, \ldots, t_n$, and observational terms $o_1, \ldots, o_n$,. Let C_t denote the conjunction of theoretical axioms of T that contextually define theoretical constructs, and let *CR* refer to correspondence rules connecting theoretical and observational terms. Now, the conjunction $C_t \wedge CR$ represents T. Denote this conjunction by *K*. The problem is as follows. Observational terms are definable by explicit definitions that can be regarded as meaning postulates, that is, analytic sentences. On the other hand, T gives only a partial interpretation of theoretical terms. Since T represents an empirical knowledge, it has a synthetic component. But T also possesses an analytic component as it partly defines $t_1, \ldots, t_n$. How to extract the analytic component of T?

We replace $t_1, \ldots, t_n$ occurring in *K* by variables, and prefix the result by the existential quantifier binding all variables in *K*. Thus, we obtain the formula (*) $\exists x_1, \ldots, x_n K(x_1, \ldots, x_n)$ which is equivalent to *K*. The transformation of *TK* into (*) has nothing to do with observational terms. Therefore, if *TK* is a synthetic statement, (*) is of the same character too. However, the formula (**) '(*) $\Rightarrow$ *T*' is a meaning postulate for theoretical terms (and observationales too, but it is not relevant here). We can prove (a) (*) $\wedge$ (**) $\Leftrightarrow$ *K*, and (b) if *A* is a sentence without theoretical terms and it is derivable from (**), then *A* is logically true in the sense of (7). It justifies the following definition:

A is analytically true in a theoretical system T relatively to the set *MP* of its meaning postulates for $t_1, \ldots, t_n, o_1, \ldots, o_n$. if and only if $A \in CnMP$.[20]

(31) Ajdukiewicz (1958, 254-257)
(a) *A* is analytic in the semantic sense in **L** if and only if *A* is a postulate of **L** or logical consequence of L-postulates;
(b) *A* is analytic in **L** if and only if *A* is a logical truth, that is, a truth invariant under substitutions of descriptive (extralogical) constituents occurring in *A* or it is reducible to logical truth in virtue of syntactic terminological conventions of **L.**

Semantic conventions ascribe denotations to descriptive terms occurring in postulates of **L**, but syntactic conventions produce expressions which can be used interchangeably, that is,

> "[...] in such a way that if one accept a sentence involving one of those expressions one ought to accept also the sentence if that expression is replaced by the other expression. Syntactical conventions so characterized are simply rules of transformation of sentences into others." (Ajdukiewicz (1958, 260).

Ajdukiewicz also observes that justification of analytic sentences sometimes requires an appeal to experience, because it happens that we have to show that terms involved in sentences are not empty.
(32) Pap (1958, 423)
(a) *A* is broadly analytic if and only if *A* is true by virtue of meanings of constituent terms.
(b) *A* is explicitly analytic if and only if *A* is a substitution instance of a logical truth.
(c) *A* is implicitly analytic if and only if *A* is translatable into an explicit analytic truth by virtue of an adequate explicative definition.

The category of the broadly analytic is generic, with the explicit analytic and the implicit analytic as its species.
(33) Bergmann (1958, 74-76)
A is analytic if and only if *A* is a tautology, that is, either a sentential tautology or a tautology of predicate logic or meaning-tautology.

According to Bergmann, (a) analyticity is a syntactic concept, (b) arithmetic is analytic, (c) descriptive terms occur in analytic sentences at most vacuously, (d) every analytic truth is either a logical tautology or deducible from logic by specialization.
(34) Mehlberg (1958, 271)
A is analytic if and only if *A* is a consequence of every referentially true sentence.

In fact, this definition could be generalized: *A* is analytic if and only if *A* is a consequence of every sentence. The original Mehlberg's version is perhaps less general because he discussed the problem of verifiability of analytic sentences.
(35) Martin (1959, 25-26)
Martin distinguishes non-translational and translational semantics. The latter is based on a translation of the object-language into the metalanguage, but the former lacks this property. At first, analyticity is defined for translational semantics:
(a) *A* is analytic in the object language **L** if and only if the formula '*A* is true' is provable in the metalanguage by means of the logical, syntactic and semantic axioms.

Analyticity for not-translational semantics is determined by
(b) *A* is analytic in L if and only if the formula '*A* is true' is provable only from logical and syntactic axioms together with comprehension rules.

And Martin's general idea of analyticity is as follows:
(c) A is analytic of **L** if and only if A is true and every sentence B resulting from A by simultaneous replacement of its primitive predicates by other predicates (primitive or definable) in the prescribed way is also true.

The formulas (a)-(c) reproduce the main intuitions only because Martin's full definitions, particularly (c), are technically complicated. However, we must say something about comprehension and admissible operations yielding analytic truths. The rules of comprehension establish denotations of predicates, and admissible operations are simply ways of correct introduction of extralogical predicates into a formal language. (35c) looks Quinean, and it is, except as Martin himself notes (see 107-113), Quine's definition *via* truth-preserving through replacements does not appeal to a definite logical apparatus.

(36) Bennett (1959, 186)
A is analytic in an argument if and only if A is a terminal sentence involved in this argument.

Bennett assumes that there are sentences which are indispensable in argument. A sentence occurring in an argument RG is involved in this argument. Let $< E,A_l >$ refer to a pair consisting of experiential data E and a sentence A_l confirmed by E. If a person 0 performs an argument, 0 can explain his grounds of accepting a sentence A_n by saying 'if S_{n+1} is true, S_n is true too'. The last (terminal) sentence in a given RG is analytic, relatively to that argument. Bennett suggests that 'analytic' could be defined by 'analytic in most arguments', but he does not develop this idea to a sufficient degree.

(37) Putnam (1962, 59)
A is an analytic statement if and only if A is immune to revision without a change of meaning.

Putnam's justification for this definition is pragmatic. Simply, we need languages with fixed points, that is, sentences immune to revision without abandoning stipulated meanings.

(38) Pollock (1965, 153)
A is analytic if and only if the following conditions are logically provable:
(a) $\exists B(\neg A \Rightarrow (B \wedge \neg B)$, (b) $\neg A \Rightarrow A$, (c) $A \Leftrightarrow (A \vee \neg A)$, (d) $\exists B\ (B \in \mathbf{TAUT} \wedge (B \Rightarrow A))$, (e) $\exists B\ (B \in \mathbf{TAUT} \wedge (B \Leftrightarrow A))$,

Informally speaking, A is analytic if its negation entails a contradiction or it is entailed by self-negation or it is equivalent to a tautology or it implies tautology. Pollock remarks that (a)-(e) are provably equivalent (in fact, either (c) or (d) is intensionally redundant). Finally, Pollock selects (a) as the most intuitive account of analyticity, and says that analytic truth "can be ascertained simply by the analysis of meaning." (133) This construction seems to capture certain Frege's intuition, but there still is a problem when provability is dependent on extralogical meaning postulates. Everything is clear only if we assume that Pollock defined analyticity of logical tautologies.[21]

(39) Sloman (1965, 13)
A is analytic if and only if its truth-value can be discovered or demonstrated using purely logical considerations based on the understanding of logical constants and

constructions involved in *A*, and defining relations between its non-logical constituents.

This definition is used by Sloman in his analysis of relationships between 'necessary', 'a priori' and 'analytic'.

(40) Moravscik (1965, 417)

A is analytic if and only if *A* is true in virtue of the meanings of its nonindexical constituents.

This definition is a result of an analysis of definitions like (19b) fairly popular between philosophers. It is essentially based on a difference between sentences and propositions. Since every proposition can be expressed by various sentences with indexical constituents, the additional constraint is essential. However, in the terminology of this paper, sentences subjected to semantic analysis, at least as far as the matter concerns analyticity, are regarded as having fixed meanings. Hence, we assume that indexicals are always determined by context. Anyway, Moravscik is right that something must be added about indexicals.

(41) Borkowski (1966, (60-61)

(a) *A* is analytic in the semantic sense if and only if *A* is true in all non-empty domains.

(b) *A* is analytic in the syntactic sense if and only if *A* is derivable in the virtue of logical rules.

(c) *A* is analytic in the pragmatic sense if and only if *A* is asserted *via* axiomatic and deductive rules of language.

The concept of axiomatic and deductive rules is borrowed by Borkowski from Ajdukiewicz (1934). A person *0* axiomatically accepts *A* if and only if *0* accepts it unconditionally, and *0* deductively accepts *A* if and only if *0* accepts it by the virtue of deductive rules. The point (b) is a modification of Frege's definition (see above) by omitting definitions as a source of analyticity. It is possible by strengthening the rules concerning quantifiers, but details must be omitted here.

(42) Suszko (1968, 216)

A is analytic if and only if *A* is asserted in every process of transformations of theorems.

Analytic sentences cover logical tautologies and extralogical principles of thinking. It does not mean that analytic axioms are immune to revision. However, if we assume that the process of thinking is to be identified by its principles, their change breaks its identity.

(43) Pollock (1969, 5)

A is analytic if and only if it is truth in all possible worlds.

For Pollock, this definition covers sentences like 'All brothers are male'. All bachelors are unmarried' and logical tautologies (formal analytic sentences). This last category can be defined in the following way:

A is formally analytic if and only if any sentences having its form is analytic.

(44) Stenius (1972, 68)

A is analytic if and only if, according to the semantic conventions for the use of certain symbols, *A* is true no matter what is the case.

For Stenius, his definition satisfies the following intuitions: (a) the AS distinction is semantic (not epistemological), (b) analytic sentences are recognized as true by analysis of their content, (c) analytic sentences are tautological, (d) all logical truths are analytic. Stenius ascribes (a)-(d) to Kant. but he is not quite right with respect to (d) (see above).[22]

(45) Katz (1972, 173-175)

If *A* is a copulative sentence (that is, a sentence of the form '*S* is P'), then *A* is analytically true if and only if the sense of *P* is contained in the sense of *S*.

This formulation is simplified. The original definition uses the concepts of semantic marker and reading, which are too complicated for a brief treatment. There is a significant feature in Katz's approach. It is assumed that truth-conditions of sentences are given not only by direct grammatical constituents, but also by presuppositions. Katz's definition is intended for sentences of natural language.

(46) Hintikka (1973, 148-149)

(a) *X* is a valid analytic argument if and only if the conclusion is a subsistence of one of the premises.

(b) *X* is a valid analytic argument step if and only if *X* does not introduce new individuals into reasoning.

(c) *X* is a valid analytic argument step if and only if the information carried by its conclusion is not greater than the information carried by its premises.

Hintikka's approach is, like Frege's, justification oriented. The definition (a) delimits analytic method of proof, and (c) applies to tautological truth. For Hintikka, (b) is the most important. In particular, it provides a tool for analysis of the synthetic or analytic character of mathematical theorems Hintikka defends syntheticity of some mathematical truths.[23]

(47) Bunge (1974, 170)

A is analytic in **T** if and only if *A* is either a definition in **T** or *A* is model-free.

A formula is model-free if it is a tautology. Thus, definitions must be always relativized to definite theories.

(48) Hao Wang (1974, 276)

A is analytic if and only if *A* is asserted on the base of conventions concerning the use of words.

Wang admits that his definition is not precise, but he points out that empirical investigation of linguistic habits gives a sufficient evidence that we have a stock of sentences which behave in a way traditionally attributed to analytic sentences.

(49) Nowaczyk (1977, 477)

A is analytic in **L** if and only if *A* is a thesis of S(L), where S(L) is a semantic system formulated in **L**.

The concept of semantic system is based on the following intuition: S(L) is a semantic system if and only if it satisfies constraints of meaningfulness given by rules assumed in advance.

(50) Chisholm (1977, 57)

A is analytic if and only if *A* may be expressed in a sentence in which the predicate-term is analyzed out of the subject-term.

For Chisholm, this definition improves Kant's formulation and is important for proving that synthetic a priori sentences exist (see section 3 below).[24]

(51) Priest (1979, 292)
A is analytic if and only if *A* can be correctly derived from conditionals corresponding to valid inferential rules.

Priest says that rules of inference involved in his definition are better than semantic rules used by Carnap and similarly thinking philosophers.

(52) Gupta (1982, 58)
A is an analytic sentence if and only if the denial of *A* is self-contradictory.

Gupta deliberately follows Kant in the general definition, but gives a modern account of Kant's distinction of *explicita* and *implicita.* The former are *reiterata* by repeating either the same term, e.g. A square is a square' or a part of the subject, e.g. 'A square is rectangular', but the latter are *explicata* on the base of an earlier analysis. Also, *explicata* are either full, if they contain the whole result of analysis, e.g., 'If I know that *p*. then I believe that p, and *p* is true, and I have an evidence for believing that *p*. or they are partial, if they mention only a part of analysis, e.g. 'If I know that *p*, then *p'* is true.

(53) Bencivenga (1986, 17)
(a) *A* is analytic if and only if any of its maximal paraphrases is logically true.
(b) *A* is analytic if and only if its maximal paraphrase is logically true.

A is a paraphrase of *B* if it preserves the structure of B, relatively to expressive means of a given language. It is not a formal concept, and it is admitted that the structure is preserved "as much as possible". *A* is a proper substitution instance of *B* if *A* is a paraphrase of *B* and *A* is longer than *B*. *A* is a maximal paraphrase of *B* if and only if *A* is a paraphrase of *B* and no proper substitution instance of *A* is a paraphrase of *B*. Bencivenga says that (a) is a clarification of analyticity in the sense of (Kl), and (b) explains analyticity as true by the virtue of form. It is important to note that Bencivenga's semantics does not admit the rule of substitution.

(54) Martin (1987, 322)
A is analytically true in **L** if and only if it is true in every intensional interpretation of **L.**

This formulation is a simplification of the original version as far as it omits various formal details. It is important to note that Martin claims that intensional logic is necessary for a correct definition of analyticity.

(55) Moses (1992, 13)
A is analytically true for a person *0* at the time *t* if and only if at *t O,* owing just to his commitment to criteria for the use of ^-constituents, will reject all falsifiers of *A,*

As Moses explains, his definition tries to meet Quine's objections to analyticity (see below). However, Moses also points out that (53) is coherent with Quine's general epistemological position.

(56) Tappendem (1993, 244)
A is analytic if and only if *A* is logically valid or it exhibits the proper kinds of intralinguistic meaning relations between lexical items.

This definition is a variation of the idea that analyticity is a feature of logical truths or sentences which are reducible to logical truths by putting synonyms for synonyms.

(57) Martin-Löf (1994, 91)

Martin-Löf places the problem of analyticity in the framework of a special logical construction, namely intuitionistic type theory. Omitting details, Martin-Löfs definition goes as follows:

A is analytic if and only if *A* has the form 'the object *o* belongs to type *t* or *A* has the form 'o_1 and o_j, are identical in type *t*.

According to Martin-Löf, if we replace 'type' by 'category in Kant's sensed this definition is close to Kantian account. Moreover, no existential judgment is analytic.

(58) Müller (1998, 275)

A is analytic if and only if for any *B*, for all $n \geq 0$, and for every $(n + 1)$-ary sentential formula $F(C_1, ..., C_n)$, the following condition is satisfied: $F(B, C_1, ..., C_n)$ is stimulus synonymous to $F(A, C_1, ..., C_n)$.

Müller uses the concept of stimulus synonymity, that is, meaning sameness organized by behavioral criteria. Thus, users of a language react to analytic sentences as to tautologies.

(59) Detlefsen, McCarty, Bacon (1999, 5)

A is analytic if and only if *A* is true by dint of their forms or the meaning of their constituent terms.

Perhaps the most interesting point about this definition is that it was included into an elementary dictionary of logic. It shows that analyticity is on the tongue of teachers of logic.

And three more quotations.[25]

(60) Quine (1962, 55)

> "[...] I call a sentence *stimulus-analytic* for a subject if he would assent to it, or nothing, after every stimulation (within the modulus)."

The modulus is understood as a behavioural standard of use of expressions.

(61) Quine (1974, 79)

> "|...| a sentence is analytic if *everybody* learns that it is true by learning its words. Analyticity, like observability, hinges on social uniformity."

(62) Quine (1995, 45)

> "An observation categorical is *analytic* for a given speaker if the range of stimulations under which he is disposed to assent the first of the two observation sentences in the categorical already includes all the stimulations under which he is disposed to assent to the second observation sentence, so that for him the categorical is trivially true out of hand and worthless in testing scientific hypotheses."

The simplest case of an observation categorical consists in co-occurrence of two observation statements associated with an expectation in such a way that both are satisfied in all expected circumstances. For example, the sentence 'When it snows, it is cold' is an observation categorical.

I listed sixty two proposals how to define analyticity. I do not claim that all suggestions are mutually different. On the contrary, the list contains variations of a few ideas and it can be considerably reduced (see below). I decided to mention so many (although certainly not all) attempts to solve the problem of analyticity in order to collect many intuitions concerning this concept.[26] I hope that this material throws some light on the issue and it will help in further analysis. The 'analytic business' has been fairly popular among philosophers since Kant, notably within the last hundred years. However, it had to meet a very serious challenge, namely Quine's attack on AS in his 1951.[27]

For the sake of argument, Quine distinguishes two kinds of analytic sentences: (a) logical truths, and (b) cases, like 'All bachelors are unmarried'. The class (a) does not make any trouble, but the concept of analyticity is redundant here because we have the notion of logical truth which is well explained by invariance with respect to substitutions for extralogical expressions. Thus, the troubles are caused by cases of the type (b). Quine investigates several attempts (covered by (1)-(13)) undertaken to define analyticity and tries to show that all are fatally incorrect. A popular account (mentioned several times in (1)-(62)) consists in reducing analytic sentences of the type (b) to logical truths by replacing synonyms by synonyms. However, it requires a definition of synonymity, but this concept is by no means clearer than analyticity. The interchangeability *salva veritate* is too weak as a criterion, and other attempts require an appeal to the concept of meaning, etc. The same problem arises when one assumes the concept of semantic rule. Referring to necessity of analytic statements does not help either, because modal concepts are notoriously unclear. Therefore, all definitions of analyticity are burdened by *obscumm per obscurum* or *idem per idem* as employing either unclear concepts (meaning, synonymity, necessity, semantic rule) or come back to analyticity. At best, we can define 'analytic in $\mathbf{L}_i$', where $\mathbf{L}_i$, is a concrete artificial language system, but something different is intended, namely a definition of 'analytic' or even '*A* is analytic in **L**', where the variables *A* and **L** range on arbitrary sentences and languages, even formalized one. Quine's final verdict is radical:

> "It is obvious that truth in general depends on both language and extra-linguistic facts. The statement 'Brutus killed Caesar' would be false if the world has been different in certain ways, but it would be false if the world 'killed' happened rather to have the sense of 'begat'. Thus one is tempted to suppose in general that the truth of a statement is somehow analyzable into a linguistic component and a factual component. Given this supposition, it next seems reasonable that in some statements the factual component should be null; and these are analytic statements. But, for all its a priori reasonableness, a boundary between analytic and synthetic statements simply has not been drawn. That there is such a distinction to be drawn at all is an unempirical dogma of empiricists, a metaphysical article of faith." (1951, 36-37)

For Quine, the only way to analyze meanings goes *via* the concept of stimulus meaning which can be accommodated by the behavioristic perspective. Now, it is perhaps not quite strange that Quine is ready to admit a residuum of analyticity in (60)-(62).[28] Although these three attempts are not equivalent (in particular, (60) seems more restrictive than (61)), all are consistent with Quine's basic claim: there is no sharp boundary between analytic sentences and synthetic sentences. On the other hand, all efforts to delimitate analyticity as an absolute property fail and have to fail. Using a nice distinction introduced by Gewirth (1953, 397), Quine is a gradualist, though Carnap defends genericism, that is, a view that analyticity is a generic property with well-defined boundaries and species.

There is a definite background behind Quine's criticism of the AS distinction: indeterminacy of translation, methodological holism (knowledge as the whole is subjected to testing procedures), extensionality (only referential semantic properties of expressions are legitimate), antiessentialism (the concept of necessity referring to essential properties is meaningless), behaviourism (all semiotic data should be analyzed in terms of external human behaviour), antimentalism (mental states are excluded), and nominalism (abstract entities are not admitted). These views support Quine's scepticism concerning synonymity, necessity, semantic rules, etc. in their traditional setting. A good summary of the Quine style criticism of analyticity is given by Nordenstam (1972, 32):

> "[...] Quine, White and Goodman demand a definition or criterion of analyticity and/or synonimity which is (i) general, (ii) has the from t... if and only if —', (iii) avoids suspect notions, (iv) is non-circular, (v) is behaviouristic, (v) is nominalistic."

Doubtless, Quine's 1951 gained the reputation of one of the most important, provocative and influential papers in the 20th century philosophy. Friends of analyticity reacted very soon.[29] Mates (1951) raised the following points. He remarked that one would distinguish between intension and extension of 'analytic'. Now, the definition of 'analytic' as a clarification of the meaning of this word need not be confused with an account of its extension. It can happen that people understand the intension of 'analytic' but they worry about its extension. On the other hand, it is also possible that people are able to identify analytic sentences, but they have problems with their definition. In particular. Mates says (528)

> "the competency [of recognizing analytic sentences – J. W.] has perhaps been acquired somewhat in the manner in which a sightless person can learn the correct and appropriate use of language appropriate to visual phenomena."

This argument was also followed by other authors who thought (see Wang (1974)) that vagueness of a distinction was one thing, but the recognaizability of intuitively plausible instances of it had to be seen as a different matter. We can summarize this point by saying that Quine simplified the matter, when he rejected the AS distinction for its vagueness. On the other hand, Quine could be defended by noticing that the considered argument confirmed his gradualism. Nevertheless, it seems that Mates and his followers exhibited an ambiguity in Quine's position. It is not always clear whether the adjective term 'analytic' is for him empty or vague.[30]

Mates agrees with Quine that circularity is involved any definition of analyticity. On the other hand. Mates notes that circular definitions are illuminating in some cases (528-529):

> "It is interesting fact that circular definitions are very often very effective in creating understanding: i.e., it often happens that after being subjected to such definitions people are able to make the various decisions which we regard as indicative of the psychological phenomenon called "understanding". Thus, even if most of the definitions [of analyticity – J. W.] [...] are in some intelligible sense circular (which I do not grant), they may well be of help in understanding the term "analytic". Further, these definitions state interesting semantical relationships among the terms occurring in them, and this information is valuable in itself."

I think that Mates' point expressed in the last quotation can be strengthened. The philosophical understanding (perhaps, not only philosophical) very often requires simultaneous grasping of somehow related terms, very frequently constituting oppositions by their meanings. For example, materialism is unintelligible without idealism, realism without anti-realism, rationalism without empiricism, and analyticity without syntheticity. A further analysis in such a case must take into account cognate distinctions, for example, the AP distinction in the case of analyticity. It seems that Quine ascribes to philosophy too high standards of non-circularity, perhaps proper for mathematics and natural science. This was also observed by Mates with respect to Quine's claims concerning adequacy conditions for the definition of analyticity. According to Mates, Quine (and other critics of the AS distinction) postulate that the definition should satisfy usual formal constraints, but, additionally, that its definiens must be understood better than definiendum. However, if we demand (a) better understanding in the above sense, and (b) the cognitive synonimity, nothing more can be achieved than the equality 'analytic = analytic'. Moreover, Mates argues that we can imagine practical tests for synonimity and other semantic properties. Finally, Mates says that we easily encounter intuitive cases of analyticity and synonymity. Although an amount of vagueness is obvious when examples are taken from the ordinary language, "the distinctions are there and cannot be argued or doubted away." (533) Mates also suggests that, in spite of all difficulties mentioned by Quine, Carnap and other philosophers made a progress in analyzing analyticity, even if their efforts concerned simple languages: "in any case, a difficult task is not necessarily impossible." (533)

The next criticism of Quine was made by Martin (1952); see also Martin 1959, 107-113. In general, Martin defends a formal approach to semantic questions. His first objection concerns Quine's use of the phrase 'analytic in **L**' According to Martin, Quine demands a definition in which L ranges over all languages, natural and formal as well. Martin sees no justification for such a general claim. In fact, one could not expect **L** to range over all formalized language. The task must be more modest and only concern a well described formalized language. Any precise definition of analyticity in natural language is hopeless for the well known features of ordinary speech like its vagueness or changeability. Moreover, since natural language is inconsistent, we could prove that all its sentences are analytic. Martin also observes that Quine confused 'analytic in **L**' with constraints concerning an adequate definition of analyticity. Assume that we define analyticity for a given language **L**. The definition must be formulated in its metalanguage **ML.** Now, the conditions of adequacy for defining 'analytic in **L**' belong to **MML.** If this distinction is respected, then, Martin says, semantic rules can be considered in a satisfactory way:

> "Perhaps a meta-theoretic definition of 'analytic in L' is being confused with a meta-meta-theoretic definition of being an adequate definition 'analytic in *L*' At any event, 'true by semantical rules' is never intended to provide a definition of 'analytic' but only a (partial) condition under which any such definition might be regarded as adequate." (Martin 1959, 112)

Finally, Martin generally evaluates the debate. For him, there are two criteria of judging a definition of analyticity: (a) does it agree with the traditional use of 'analytic', at least whenever this adjective is clear, and (b) is the proposed definition fruitful? Martin says that efforts undertaken by Carnap and himself suggest the decisively affirmative verdict. I share a general reconstructionist position of Martin, and I will later argue that we should define analyticity for well determined cases, that is, specified theories, conceptual schemes or at least units of language with clearly indicated definitions. On the other hand, I have reservation whether the formal approach gives as decisive results as Martin is inclined to think. If one agrees that formal methods are proper in semantics, one will follow Martin's recommendations and his optimism, but outside the reconstructionist camp the situation is not so unconditionally clear. Almost thirty proposals how to define analyticity registered above, which appeared after 1959, show that the issue is still controversial. The point is that we are in the domain of philosophy where it can hardly be expected that one metatheoretical perspective will ultimately win. Returning to more concrete problems, it is obvious that Martin learned a lesson from Tarski's truth-definition. In fact, it is strange that Quine fully agreed with Tarski as far as the matter concerned the definition of truth, but applied different conditions of general adequacy to the definition of analyticity. Another interesting point regards the relation of **ML** and **MML.** It seems quite plausible to maintain that we can formulate semantic rules for **L**, even if we are not able to give a satisfactory definition of the category of semantic rules in **MML.**

Further important remarks against Quine were formulated by Kemeny (1952), (1952a); see also Kemeny 1964. He agreed with Mates (see Kemeny 1952a) in general points, and raised certain concrete objections. In particular, Kemeny defends the concept of meaning postulates. He points out that if we define analytic sentences as valid in all possible worlds, we must take into account not only meanings of logical terms, but also the content of extra-logical postulates. Kemeny applies this idea to an example considered by Quine, namely the sentence (a) 'Every green is extended', often considered as analytically true, but not reducible to logical truths. Let (a) be written as (b) $\forall x(Gx \Rightarrow Ex)$. Kemeny argues that 'everything' is the crucial word. If we accept (c) $\forall xEx$ as a meaning postulate, (a) becomes analytic. At the end of his review, Kemeny says: "(The reviewer] fails to see how the difference between a pure convention and a factual assertion is a matter of degree." (138) However, I think that Kemeny begs the question. The last quotation expresses a subjective attitude. The discussion of (a) and (b) is illuminating, but Quine could reply that he is still unconvinced that (a) is devoid of a factual content and purely linguistic.

Perhaps arguments endorsed by Grice and Strawson (1956) are the most recognized criticism of Quine. At first, they tried to identify what Quine actually intended to do in his criticism of the concept of analyticity. Grice and Strawson argued that Quine should not be understood as assuring that the AS distinction is empty. It would be at odds with the well-established philosophical tradition. It is rather that (198):

> "Quine's thesis might be better represented not as the thesis that there is *no difference at all* marked by the use of these expressions [i.e., 'analytic' and 'synthetic' – J. W.], but as the thesis that the nature of, and reasons for, the difference or differences are totally misunderstood by those who use the expressions, that the stories they tell themselves *about* the difference are full of illusion."

Grice and Strawson who are close to ordinary language philosophy believe that the philosophical tradition provides a strong presumption for the AS distinction, but they agree that Quine need not be convinced by this argument. Thus, they look for other arguments. Although they do not quote Mates 1951, their arguments are similar to those advanced by him. Grice and Strawson point out that Quine's demands addressed to the definition of analyticity are too strong. Using their own metaphilosophy, they suggest that we have at our disposal various means to explain the meaning of related terms, not necessarily definitions in the strict sense. I think that the most interesting argument raised by Grice and Strawson is this (211):

> "The point of substance (or one of them) that Quine is making [...] is that there is no absolute necessity about the adoption or use of any conceptual scheme whatever, or, more narrowly and in terms that he would reject, that there is no analytic proposition such that we *must* have linguistic forms bearing just the sense required to express this proposition. But it is one thing to admit this, and quite another thing to say that there are no necessities without any conceptual scheme we adopt or use, or, more narrowly again, that there are no linguistic forms which do express analytic propositions."

Gochet (1986, 26) renders the last point formally. We should distinguish two theses:
(Tl) There is a sentence *A* that for any conceptual scheme **SCH,** *A* is analytic in **SCH.**
(T2) For any conceptual scheme **SCH** there is a sentence *A* such that *A* is analytic in **SCH.**
It is clear that Quine requires a definition justifying (Tl). However, another route can consist in taking (T2) as a guide or combining both theses in the following way:
(T3) If *A* is an analytic sentence, there is a conceptual scheme **SCH** such that *A* is analytic in **SCH.**
(T3) seems more appropriate for constraining analyticity than (T2) because the latter rather states a condition for being a conceptual scheme than an analytic sentence. The intuitive content of (T3) is simply this: every analytic sentence possesses this qualification relatively to a certain conceptual scheme. I will come back to that problem later.

Carnap addressed to Quine's criticism in his (1963). Carnap is puzzled by Quine's treatment of meaning postulates:

> "The first of Quine's critical arguments consists in the remark that the meaning postulates are recognizable only by the label "meaning postulates" and that the sense of this label is not clear [..] I was puzzled by this remark because neither Quine nor anybody else has previously criticized the obvious fact that, e.g., the admitted forms of sentences of a formalized language are only recognizable by a label like "Sentence Forms in L" preceding a list of forms of expressions, or the fact that the axioms of a logical calculus are only recognizable by the label "Axioms". Why should the same fact be objectionable in the case of meaning postulates?" (918)

Carnap's reply echoes that of Martin. Yet, Quine might remark that syntactic forms and the set of axioms could be identified without any further ado, just by ostensive procedures. However, the intended function of meaning postulates or semantic rules is differently conceived by Carnap and other friends of the generic concept of analyticity. In particular, a Quine-style reply to Carnap might point out that meaning postulates make some sentences purely verbal and devoid of factual content. And this is what Quine abandons, not meaning postulates as decisions concerning the use of words, but their function as generating factually empty statements. Thus, we need a justification of why meaning postulates in Carnap's sense are needed for a semantic theory.

Carnap comments also Quine's claim that analyticity is legitimate only in the sense of (60)-(62):

> "It seemed to me puzzling why for semantical concepts like analyticity or synonymy the definition of a corresponding empirical, pragmatical concept is required, while for other semantical concepts like truth, the name-relation, and the like, a requirement of this kind is not made." (918)

In his typical honest way Carnap tries to interpret Quine as making a point which could be commonly accepted. Carnap says that Quine probably sees the concept of truth as having a clear and uncontroversial pre-theoretical explication, but the opposite holds for analyticity. I do not think so. Quine seems to think that intensional aspects of truth are not troublesome contrary to the case of analyticity. Moreover, Carnap is mistaken in his view that the concept of truth has nothing to do with pragmatics. As I will argue, truth and analyticity share the same fate: both are intelligible only for interpreted languages.

Nordenstam (1972) agrees with the previous authors that Quine's formal requirements for the definition of analyticity are too strong. He also raises the question of a general philosophical environment of Quine's attack against analyticity. Nordenstam says that although extensionalism, nominalism and behaviourism are valuable positions in philosophy, their use should not be exaggerated. My position is stronger. I agree that all Quinean views deserve to be respected. On the other hand, they are at least as controversial as their opposites. The same concerns Quine's more special favourites, like stimulus meaning, indeterminacy of translation or behavioural criteria of synonymity. Perhaps it is better to do philosophical analysis without any appeal to general controversies if it is possible. I believe that analyticity is a case of this kind.

Moser (1992, 17-18) accuses Quine of neglecting

> "the distinction between *what is true* just in virtue of a usage commitments and *what justifies* a usage commitment. [...] A statement can be analytically true just in virtue of a usage commitment, even if that commitment is supported by, or justified, only by certain alterable explanatory purposes a language user has. The alterability of the supporting purposes does nothing to discredit the analyticity arising from a usage commitment."

Thus, Moser suggests that a sentence may be analytically true and yet the grounds of analyticity, that is, usage commitments recur to various, basically changeable, data. However, it is not clear, at least on the base of Moser's literal formulation, whether data which support commitments are empirical, factual, experiential, etc.

Boghossian (1997, 340) observes that scepticism toward analyticity can be expressed by two different theses: (a) Since no coherent, determinate property is expressed by the predicate 'is analytic', propositions '*A* is analytic' and '*A* is synthetic' are not coherent either, and (b) since there is a coherent, determinate, but unistantiated property 'is analytic', all sentences of the form '*A* is analytic' and '*A* is synthetic' are false. Then, Boghossian remarks that is it not clear whether Quine defends (a) or (b). Another Boghossian's critical remark points out that Quine restricts his criticism only to Frege-analyticity. I do not think that Boghossian is fully right. It is rather that most of Quine's critical remarks concern reducing analytic sentences to tautologies *via* synonimity, although, as Katz (1992, 16) pointed out, Quine disregarded the fact that analyticity could be understood less or more narrowly. Perhaps Boghossian's most interesting argument concerns holism (344). Boghossian attributes meaning holism to Quine. It is a risky move because Quine's holism is rather methodological. The difference is that methodological holism is the thesis about testing hypotheses by empirical data, and meaning holism is the view that meaning relations depend on language as the whole. Well, since meaning holism is also interesting from the point of view of analyticity, let us reproduce Boghossian's argument without deciding whether it applies to Quine or not. The argument is as follows. Assume that meaning holism is correct. Therefore, it is very unlikely that synonyms occur in any language. Boghossian says that "unlikely" does not mean "impossible". Moreover, he finds meaning holism quite implausible. It is important that the argument from holism is highly conditional: if meaning holism is correct, then analyticity *via* synonymity is difficult to be expected. However, much more can be said here. Ajdukiewicz 1934 accepted a radical meaning holism, but he still maintained that some sentences were accepted by purely linguistic rules. In Ajdukiewicz 1947, we find methodological holism as well as a definition of analyticity (see (20) above). It shows **L** that holism can be perfectly coherent with analyticity.

Finally, I mention Hintikka's 1999 arguments against Quine. Hintikka's main target is Quine's thesis that every sentence is a composition of conceptual and factual content. At first, Hintikka argues that it is important to extract purely analytic (conceptual) truths, that is, logical tautologies:[31]

> "If there are no purely conceptual truths, there are no purely conceptual consequence relations. And if there are no purely conceptual consequence relations, there are not purely tautological inferences. Consequently, the classical project of purely logical axiomatization (axiomatic method) is a chimera." (2)

It is possible to give a formal version of this argument. Let **L** be a language understood here as a set of sentences. A consequence operation defined on **L** is any mapping from $2^{\mathbf{L}}$ to $2^{\mathbf{L}}$, that is, a mapping from subsets of **L** to subsets of **L**. We say that a consequence operation *Cn* is reasonable if and only if it preserves the distinguished logical value. Intuitively, if truth is the distinguished value, then *Cn* is reasonable if leads from true premises to true conclusions. Now assume that a sentence A is a consequence of the set $X \cup \{B\}$, that is, $A \in CnX \cup \{B\}$. Then, assuming the deduction theorem, $B \Rightarrow A \in CnX$. If $X = \emptyset$, the implication $B \Rightarrow A$ is a logical theorem. Of course, any mapping from $2^{\mathbf{L}}$ to $2^{\mathbf{L}}$ is a consequence operation, but not every operation of this kind generates a reasonable one. Assume that $A \in Cn\{A\}$ if and only if A materially implies B. Provided that the deduction theorem holds for this consequence operation, we obtain that every material implication belongs to logic. However, it is difficult to regard this *Cn* as reasonable for its unstable relation to truth-preserving. Clearly, the required property of *Cn* is related to a philosophically important metalogical theorem that logic does not distinguish any extralogical constant. We have also a way to justify why such a stable consequence operation is cognitively important. A hint for doing that can be derived from probability calculus. One of axioms of that theory says that probability of tautologious events is equal to 1, because dispersion of probability values were not blocked in another case.

Hintikka's second argument concerns analytical which are not logical tautologies. Let me call them non-purely analytic sentences. They can be understood as consequences of boundary conditions which determine domains (models, possible worlds) subjected to particular scientific theories. Now, sentences performing this role certainly are not true in all possible models. The key question is how to look at the information covered by non-pure analyticals. Of course, we cannot use here the mentioned fact that logic regards all extralogical constants *al pari,* because just the opposite is the case: if boundary conditions determine concrete worlds, they automatically contain a relevant information about them. Hintikka is fully conscious of the problem when he says that sometimes it is difficult to decide how mathematical apparatus is related to empirical content. However, he denies that it is an argument for Quine's view that conceptual content is inseparable from empirical one. Yet I think that Hintikka's terminology is misleading to some extent. He says:

> "A conceptual isolation is nevertheless normally possible by specifying what mathematical physicists would call boundary conditions. If we specify the conditions holding at the boundaries of the system, we can often disregard what happens outside those boundaries and thus deal with the system itself as an entire model of the relevant inquiry." (18)

I think that it is much better to relate isolation to theoretical axioms which determine boundary conditions together with the internal structure of a given world. For example, the postulate that the speed of light is constant is not only a boundary condition, but also a theoretical axiom. What we need for analysis of non-purely analytic sentences are two analogies of the theorem of non-distinguishing extralogical constants. Firstly, boundary conditions decide that all models located outside the isolated system are *al pari.* Secondly, theories dot not distinguish between extratheoretical constants, that is, related to initial conditions. Thus, all sentences performing these roles together with their logical consequences deserve to be called 'non-pure analytical'. We can generalize the approach used in analysis of reasonable *Cn.* It is obvious that reasoning about the world subjected to a theory, requires more than is given by purely logical consequence operation. When a consequence operation is applied outside logic, we need something to establish basic connections between concepts. One can say that we need a concept of analytic consequence in such situations which is wider than purely logical *Cn,* Hintikka's generalized argument is this: rejecting analyticity means that every implication is analytically acceptable, independently of logic and specific theoretical axioms. This result is highly implausible.

Since we have several objection against Quine's criticism of analyticity, further attempts to define this concept are not without a justification. I will offer an approach which follows my 1993 and 1993a. Although particular items in the variety (1)-(62) appeal to many apparently different ideas, this multiplicity can be essentially reduced. Derivability and provability are syntactic concepts, truth in all models, possible worlds or states-descriptions belongs to semantic vocabulary, but meanings, meaning postulates or intension point out elements definable within pragmatics. Accordingly, all quoted definitions can be classified as, explicitly or implicitly, syntactic ((4), (16), (19a), (22), (34), (35a-c), (38), (41b), (46a), (46b), (51), (52)) semantic ((1), (2), (7), (10), (13), (17), (27), (32b), (41a), (41), (44c), (55)), pragmatic ((3), (5), (12), (19b), (21), (24), (25), (29), (30), (31), (32a), (32c), (36), (37), (40), (41c), (42), (44), (45), (48), (49), (50), (53), (54), (55), (56), (58), (60), (61), (62) or mixed ((6), (8), (9), (11), (14), (15), (19), (23), (28), (33), (39), (47), (59)).[32]

My analysis follows suggestions of Borkowski (see (41) above), but I introduce essential changes and additions. I also draw conclusions from a very common distinction of analytic sentences in the narrow and broad sense which occurs in many mentioned contributions (see for example, (19) and (32)), and from constant difficulties with delineating the analytic from the empirical (see (31) and (40) with respect to existence assumptions or indexicals), contrary to Carnap's simple optimism that there is nothing to be problematic here. Borkowski's reason for introducing the division into syntactic and semantic analytic sentences was motivated by Gödel's first incompleteness theorem. Since there are true and unprovable arithmetical sentences, not every true proposition is derivable. Hence, there are analytic truths in the semantic sense which are not syntactically analytic. Unfortunately, this simple idea does not work. Recall that Borkowski defines semantic analytic sentences as true in all models. If this proposal is to be seriously taken, it must be restricted to first-order logic. Now, by the completeness, theorem the set of logically valid sentences coincides with the set of theses which are derivable from the empty set.[33]

Let G be a Gödelian sentence, that is, unprovable in arithmetic. Thus, $\neg G$ is also improvable. Since both **AR** $\cup$ (G) and **AR** $\cup$ ($\neg G$) are consistent, provided that **AR** is, they have models. It means that there is a model in which G is true as well as another model in which $\neg G$ is true. Hence, neither G nor $\neg G$ are true in all models and cannot be taken as examples of sentences analytic in the semantic sense. The restriction of the problem solely to the axioms of **AR** also does not help because they are both true in all arithmetical models and provable in **AR.** One can try to use Kemeny's route (see (29) below) and decide to base semantics on the concept of intended model, but this move introduces explicit pragmatic moment into considerations

I will reformulate Borkowski's proposal and link it with another division of analytic sentences, namely into absolute and relative. Due to the priority of semantics in contemporary logic, I will first define absolute semantic analytic sentences:

(63) A is an absolute semantic analytic sentence if and only if A is true in all models.

Having (63), we can propose the following definitions

(64) A is an absolute syntactic sentence if and only if (a) A is an absolute semantic analytic sentence, (b) there is a set X such that (i) $A \in$ X, and (ii) the property of X-theoremhood is decidable.

(65) A is a relative semantic analytic sentence of a theory **T** if and only if A is true in all models of **T.**

(66) A is a relative syntactic analytic sentence of a theory **T** if and only if (a) A is a relative semantic analytic sentence in **T,** (b) there is a subset X of **T** such that (i) $A \in X$, and (ii) the property of X-theoremhood is decidable.

(67) A is a pragmatic analytic sentence if and only if (a) there is a theory **T,** (b) $A \in$ **T,** (c) A is true in all intended models of **T.**[34]

Some comments are in order. Absolute analytic sentences are restricted only to first-order logic. Since (63) appeals to the well-known idea of truth in all models, possible worlds, state-descriptions, etc., it is fairly standard. Then, (64) recurs not to provability, but to other syntactic property, namely decidability. This choice has a deep historical justification, because Leibniz thought about truths of reason as calculable, resolvable or decidable. Even if Leibniz's idea is explained with the help of provability, it is clear that he identified 'to be provable' with 'to be calculable, etc.' Moreover, decidability is, as Gödel (1946) pointed out, a much more "absolute" concept in the sense that it is independent of a concrete formalism. On the contrary, provability is always relativized to the stock of axioms and rules of inference. If A is an unprovable sentence, there is a trivial way to make it provable: it is sufficient to add it as an axiom. Decidability is not subjected to such games. If provability does not coincide with truth in all models *via* the completeness theorem, true and unprovable sentences are true only in selected models, for example, intended ones. In fact, non-coincidence of truth in all models and provability gives a reason for the distinction of semantic and pragmatic analytic propositions, whereas the lack of the correspondence between truth and decidability is responsible for the distinction between semantic and syntactic analyticals. Of course, pragmatic analytic sentences correspond with non-pure analyticals (non-purely conceptual truths in Hintikka's sense).

We have the obvious and desirable connections. Absolute analytic sentences are special cases of relative ones. Every syntactic (absolute or relative) analytic sentence is a semantic analytic sentence, absolute respective relative, and every semantic analytic sentence is automatically a pragmatic one, and the reverse connections do not hold. Since intended models are always determined by pragmatic factors, although sometimes very obvious as, for instance, in the case of arithmetic, there is no reason to divide pragmatic analyticals into absolute and relative because they are always dependent on various circumstances. The qualification 'absolute' must be taken *cum grano salis* in the present context. It is clearly connected with the chosen logic. Perhaps the simplest way to explain the point goes *via* provability. Define logic as the set of the consequences of the empty set: **LOG** = $Cn\varnothing$. Everything depends then on the properties of a selected consequence operation. The choice of classical *Cn* is common, but other ways are also possible and defended, for example, by the intuitionists. Thus, one can argue that even analyticity of logic is relative to metalogical choices and holds only for the language in which a given logical system is expressed. I will return to this problem in section 3.

Logic seems to be the only candidate for which (Tl) holds: if something is analytic in theories (conceptual schemes), it is identical with logic.[35] However, all analyticals are subjected to (T3). We can define absolute analytic sentences as those that are analytic in all theories (conceptual schemes) and relative ones as those that are analytic in some theories (conceptual schemes). The problem arising from Quine's criticism of analyticity is whether something else except logic is analytic in extralogical conceptual schemes. If we admit the concept of analytic consequence (see above), that is, *Cn* based on logical rules plus other items (axioms, definitions) used for deriving conclusions from already accepted premises, the category of the analytic covers surely something more than pure logic. It is also clear why the definition of analyticity requires a relativization to "well-determined cases" because, otherwise, the concept of analytic consequence is too vague. Yet we have a problem of analyticity in ordinary language. Strictly speaking, (65)-(67) are not applicable to ordinary language because it is not a theory. On the other hand, as I already noted, we can operate on language-units with definitions. Such units may be interpreted as ordinary conceptual schemes that pick up intended applications. For instance, the classical example 'All bachelors are unmarried men' selects intended uses of the word 'bachelor' and, thereby, intended situations in which this sentence and its consequences are true. Thus, there is no obstacle to consider the sentence All bachelors are unmarried men' as pragmatically analytic. It is certainly a relative analytical as we can imagine a quite different application of this word, regardless of its ambiguity in ordinary speech.

Now we see the point of problems raised by Quine: it lays in pragmatic factors as grounds of analyticity in extralogical cases. Clearly, it is very unlikely that any analytic sentence in the pragmatic sense holds for every conceptual scheme. Thus, there remains, as it was to be expected, only (T3) as valid for pragmatic analytic sentences. Yet, it is not clear how to explain the source of pragmatic analyticity. One way consists in selecting analytic constants responsible for analyticity in extralogical theories. Certain examples are perhaps beyond major doubts, for example 'natural number' or 'set', but other raise considerable difficulties of how they could be extracted and analyzed. Conventionalism offers a different solution: pragmatic analyticity is a result of a convention. Both ways are compatible and, roughly speaking, display the fact that analyticity in mathematics was always felt as different than analyticity in factual science or ordinary language. If it is a right intuition, mathematical analyticity is "more absolute" than physical which must be always relativized to isolation condition. Anyway, the foregoing considerations suggest that we can be gradualist, that is, maintain that analyticity is relative to several factors and, at the same time, defend AS as justifiable. Gradualism does not need to be regarded as the view that analytic and synthetic factors are mixed in all sentences. By gradualism, I rather mean the thesis that we have degrees of analyticity. Prepositional calculus provides the strongest kind of analyticity in which absolute semantic and syntactic analytic sentences perfectly coincide. First-order logic lacks this coincidence due to its undecidability. Mathematical analyticity is always relative, semantically and syntactically, due to the limitative theorems of Gödel (incompleteness) and Church (undecidability). Incompleteness causes not only semantic relativity, but also introduces pragmatic factors. Since axioms of arithmetic of natural numbers are true in its standard model as well as its non-standard interpretations, there is no other way to defend analyticity of the intuitive number theory than by pointing out that all its theorems are true in the standard, i.e. intended model. Then, we have mathematized theories of natural science with analyticity pragmatic in its character and defined by (30). Then, cases like 'All bachelor are unmarried' come in which pragmatic factors are due to definitions (stipulative or not) producing or reproducing synonymity. Finally, sentences about colours, for instance, 'Any green surface cannot be yellow at the same time' have the lowest degree of analyticity, if any.[36] Independently of various details, the above analysis strongly suggests that Carnap's favourite idea that analytic and logical truth are coextensive must be abandoned.[37]

So far I did not give any definition of synthetic sentences. A simple negative formulation '*A* is synthetic if and only if *A* is not analytic' does not help very much considering the complexity of the concept of analyticity produced by the division of analytic sentences into several categories. Thus, syntheticity should be defined step by step according to (63)-(67) If absolute analytic sentences are taken as a point reference, all the other are synthetic. In other words: every extralogical sentence is synthetic relatively to logic. However, even inside logic we have degrees. Propositional calculus is, for its decidability, more analytic that first-order logic.[38] Similarly, decidable extralogical theories have a greater degree of analyticity than undecidable. In this sense, first-order logic introduces certain amount of syntheticity in comparison to propositional logic, and full number theory with respect, for example, to the Presburger arithmetic in which addition functions as the sole binary operation. If **T** is a mathematical incomplete theory, its unprovable sentences, that is, pragmatic analyticals are synthetic relatively to **T**-provable sentences. Further, specific axioms of physical theories are synthetic in comparison with their logical and mathematical bases. Moreover, if **T** and **T'** are mutually inconsistent, although internally consistent, their elements can be mutually synthetic even if they are T or T'-analytic. Finally, if we pass to language units with definitions establishing synonyms, those (extralogical) sentences which have nothing to do with given synonymity can be regarded as synthetic. Thus, although the AS distinction is, *pace* Quine, very relative, it is still defensible.[39]

In order to convince the reader how this issue is important even for fairly elementary philosophical debates, I consider a concrete example in which the concept of analyticity is essentially employed:

> "Campbell says that 'if one is expanding knowledge beyond what one knows, one has no choice but to explore without the benefit of wisdom' (i.e. blindly). This, Campbell admits, makes evolutionary epistemology close to being a tautology [...]. Evolutionary epistemology does assert the analytic claim that when expanding one's knowledge beyond what one knows, one must proceed to something that is not already known but, more interestingly, it also makes the synthetic claim that when expanding one's knowledge, one must proceed by blind variation and selective retention. This claim is synthetic because it can be empirically falsified. The central claim of evolutionary epistemology is *synthetic*, not *analytic*. If the central claim were analytic, then all non-evolutionary epistemologies would be logically contradictory, which they are not. Campbell is right that evolutionary epistemology does have the analytic feature he mentions, but he is wrong to think that this is a *distinguishing* feature, since any plausible epistemology has the same analytic feature." (Stein 1992, 123).

I will not discuss whether the main claim of evolutionary epistemology is analytic or not. My main concern is Campbell's statement, that if the central claim of evolutionary epistemology, (C) for brevity, were analytic, then all non-evolutionary epistemologies would be logically contradictory, which they are not. Take for granted that non-evolutionary epistemologies are non-contradictory, that is, internally consistent. Is this assumption at odds with the thesis that (C) is analytic? In order to say 'Yes', one must make two additional assumptions: (a) all central claims of all non-evolutionary epistemologies are contrary or contradictory to (C), and (b) (C) is a logical truth. Now, the assumption (b) is certainly false, independently of the fate of (a). If (C) is analytic at all, it is a pragmatic analytical holding in evolutionary epistemology as a specific conceptual scheme. It is perfectly consistent with analyticity of central claims of non-evolutionary epistemologies, provided that these other claims are also interpreted as pragmatically analytic, related to definite conceptual schemes. In this sense. Stein's final remark about the status of (C) and similar statements of other epistemological theories is correct, but without a closer analysis of the concept of analyticity is not quite intelligible.

3. A PRIORI VS. A POSTERIORI

The great career of AP began, like in the case of AS, with Kant, although this distinction appeared in philosophy much earlier.[40] It goes back to Aristotle and his distinction of *proteron te fizei* (priority in nature; the cause is prior to its effects) and *proteron pros hemas* (priority in knowledge; *A* is prior to *B* in knowledge if we must know *A* in order to know *B*) . The expressions 'a priori' and 'a posteriori' appeared in the Middle Ages and their principal application concerned inferences. An inference proceeding from causes to effects was a priori, but that starting with effects and going back to causes followed the a posteriori way. For Descartes, a priori truths are self-evident, universal principles serving as the ultimate base for deductions producing knowledge. An important step was made by Leibniz who regarded a priori inferences as independent of experience. It concurred to some extent with Hume's view that relations between ideas are devoid of factual content, although terms a priori and a posteriori did not occur in works of British empiricists.

The pre-Cartesian and Cartesian perspective about the a priori was changed by Kant. He introduces his understanding of a priori knowledge in the following way:

> "This, then, is a question which at least calls for closer examination, and does not allow of any off-hand answer:- whether there is any knowledge that is thus independent of experience and even of all impressions of the senses. Such knowledge is entitled *a priori,* and distinguished from the *empirical,* which has its sources *a posteriori,* that is, in experience. The expression *'a priori'* does not, however, indicate with sufficient precision the full meaning of our question. For it has been customary to say, even of much knowledge that is derived from empirical sources, that we have it or are capable of having it *a priori,* meaning thereby that we do not derive it immediately from experience, but from a universal rule – a rule which is itself, however, borrowed by us from experience. Thus we would say of a man who undermined the foundation of his house, that he might have known *a priori* that it would fall, that is, that he need not have waited for the experience of its actual failing. But still he could not know completely *a priori.* For he had first to learn through experience that bodies are heavy, and therefore fall when their supports are withdrawn. We shall understand *by a priori* knowledge, not knowledge independent of this or that experience, but knowledge absolutely independent of all experience. Opposed to it is empirical knowledge, which is knowledge possible only *a posteriori,* that is, through experience. *A priori* modes of knowledge are entitled pure when there is no admixture of anything empirical. Thus, for instance, the proposition 'every alteration has its caused while an *a priori* proposition, is not a pure proposition, because alteration is a concept which can be derived from experience." (Kant 1781, 42-43).

As it is clear from the quoted passage, Kant was perfectly conscious of the old meaning of the label 'a priori' as referring to reasoning from causes to effects, but he radically departed from this use. Also, contrary to his rationalistic predecessors, like Descartes or Leibniz, Kant did not draw a sharp boundary between reason and experience; the latter is equated by Kant with empirical knowledge accumulated by sensory activities.[41] The opening passage of *Critique of Pure Reason* is as follows:

> "There can be no doubt that all our knowledge begins with experience [...] In the order of time [...] we have no knowledge antecedent to experience, and with experience all our knowledge begins. But though all our knowledge begins with experience, it does not follow that it all arises out of experience." (41)

In general, a priori knowledge provides possibility-makers for any experience, and it is *de iure* (legislative), because it decides about the very possibility of cognition starting with sensory activities. Whatever this legislative function might mean, the following question arises: How a priori concepts are applicable to objects? Since application of a concept to an object finds its realization in judgements expressed by sentences, a priori notions generate a special kind of sentences, namely synthetic a priori. Kant had no doubts concerning the existence of a priori syntheticals, but he wanted to explain how they were possible. Hence, the question 'How are synthetic a priori propositions possible?' became central in Kant's philosophy, and he answered it by pointing out the role of a priori in synthesis of concepts into propositions.

It is very important to see that a priori (and a posteriori as well) sentences must be taken as sentences-cum-way-of-their-knowability. Thus, if '*S* is *P*' is a synthetic a priori sentence, it is synthetic by the relation between its constituent, and a priori for the way of its knowability, namely independently of all experience. In turn, this entails that a posteriori propositions which constitute, on Kant's view, empirical knowledge are contingent and justifiable by experience, where experience is understood as sensory experience. Kant merged AS and AP. It gives four categories of propositions: analytic a priori, analytic a posteriori, synthetic a priori and synthetic a posteriori. Kant maintained that all analytic sentences are a priori. This category, like that of the synthetic a posteriori, did not present to him any difficult problem, and he justified the possibility of synthetic a priori sentences in the way which I indicated above. Here, I add that Kant supplemented apriorism by nativism, that is, the view that categories responsible for the a priori are innate in our cognitive faculties. He had to take this direction, because if the a priori is before any experience, its genesis is either mysterious or we need nativism as a working hypothesis. Examples of synthetic a priori propositions are to be found in arithmetic (for instance, 5 + 7 = 12), geometry (for instance, space is 3-dimensional) and theoretical physics (for example, the law of gravitation). Also the principle of causality is synthetic a priori. On the other hand, logic is analytic on Kant's view.

Kant's account of the a priori and his examples became standard and every analysis of this notion begins with his account, its interpretation, support, modification or criticism. Although I cannot enter here into the history of the reception of Kant's ideas about the a priori (see Milmed 1961 for many discussions and interpretations), I must at least mention (without details and bibliographical data) some criticisms of his views. Perhaps the strongest objections came from the philosophy of mathematics. The development of Non-Euclidean geometries was interpreted as a disproof of Kant's view that space is three-dimensional. Conventionalists, like Henri Poincare, point out that since various consistent geometries are possible, their choice is a matter of decision. Frege and Russell maintained that Kant overlooked analytic character of mathematics, although, according to Frege, geometry is synthetic a priori. For Russell and Carnap, pure mathematics is logic, but applied mathematics is always an a posteriori empirical science and, consequently, the synthetic a priori appears neither in mathematics nor in physics.[42] Moore and Russell argue that human capacities are contingent in themselves, and that this fact devastates Kant's explanation how a priori truth are knowable. According to Reichenbach, Kant's theory of the a priori is inconsistent with physics, particularly with the relativity theory.

The category of a priori was extremely important for Husserl. A priori knowledge organizes *mathesis universalis*. Perhaps the best explanation of this task can be found in the following fragment of Husserl 1929:

> "As a theory of science concerned with principles, logic intends to bring out *"pure" universalities, "apriori" universalities* [...] Constantly investigating the pure possibilities of a cognitive life, as such, and those of the conditional formations, as such, attained therein, logic intends to bring to light the essential forms of genuine cognition and genuine science in all their fundamental types, as well as the essential presuppositions by which genuine condition and genuine science are restricted and the essential forms of the true methods, the ones that lead to genuine cognition and genuine science. [...].
>
> The universality of logic, as concerned with principles, is not simply an apriori or eidetic universality; rather it is, more particularly, a *formal* universality. [...].

In a certain sense every eidetic condition is a product of "pure reason" – *pure from all empeiria* ([...] indicated, from another side, by the word apriori): but not every eidetic condition is pure in a *second sense,* the one pertaining to *form as a principle.* An apriori proposition about all *sounds* as such, about sounds meant with "pure" universality, is pure only in the first senses it is, as we may say for certain reasons, a *"congintent" Apriori.* It has in the eidos *sound* a materially determinate core, which goes beyond the realm of the universality of "principles" in the most radical sense, and restricts it to the "contingent" province of ideally possible sounds. *'Pure" reason is not only above everything empirically factual, but also above every sphere of hyletic, materially determinate essences.* It is the title for the self-contained system of pure principles that precede, every hyletic, materially determinate, Apriori [...]

[...] contingent [Apriori] is not an Apriori of pure reason; or, as we may also say, introducing and old word that tended blindly in the same direction, it is not an "innate" Apriori.

If we restrict ourselves to judicative reason, then as pure reason, as the complete system of this *formal Apriori in the most fundamental sense,* it designates at the same time the highest and widest conceivable theme of logic, of "theory of science. Consequently we may say that logic is *self-explication of pure reason* itself or, ideally, the science in which pure theoretical reason accomplishes a complete investigation of its own sense and perfectly Objectivates itself in a system of principle". (28-31)

Some of Husserl's theses sound Kantian (see Kern 1964, 135-145) for comparison of both philosophers about the a priori). In fact, Husserl, like Kant, contrasted formal and transcendental logic, and, like Kant, was much more interested in the latter. Although Husserl characterizes the formal a priori in the most fundamental sense as manifestation of pure reason itself, he departs from Kant when he tries to connect activities of reason with formal and material ontologies. Also, although Husserl used the word "innate", he avoided any deeper involvement into full-blooded nativism. Anyway, Husserl's idea of the a priori links epistemological and ontological issues: pure reason in itself and essential connections in the world. Frege introduced the a priori in the following way:

"[...] distinctions between a priori, and a posteriori, synthetic and analytic, concern, as I see it, not the content of the judgement but the justification for making the judgment. Where there is no such justification, the possibility of drawing the distinction vanishes. An a priori error is thus a complete a nonsense, as, say, a blue concept. When a proposition is called a posteriori or analytic in my sense, this is not a judgement about the conditions, psychological, physiological and physical, which have made it possible to form the content of the proposition in our consciousness; nor it is a judgement about the way in which some other man has come, perhaps erroneously, to believe it true; rather, it is judgement about the ultimate ground upon which rests the justification for holding it to be true. [...] For truth to be a posteriori, it must be impossible to construct a proof of it without including an appeal to facts, i.e., to truths which cannot be proved and are not general, since they contain assertions about particular objects. But if, on the contrary, its proof can be derived exclusively from general laws, which neither need nor admit of proof, then the truth is a priori." (Frege 1884, 3-4)

It is not clear what is the status of "general laws, which themselves neither need nor admit of proof". Take logic. It is a priori for its irreducible generality and analytic for its provability by logical rules definitions. The same concerns arithmetic due to its reducibility to logic. But what about geometry? According to Frege, it is a priori, because recurs to general principles, but synthetic as it appeals to spatial intuition. However, it is not obvious that laws based on spatial, that is, particular, intuition do not admit of a proof. I only note this problem without further comments and possible interpretations of Frege.[43] As far as the matter concerns historical filliations, there is an important difference between Kant and Frege concerning AS and AP. For Frege, both distinctions are justification-oriented, but for Kant, AS is certainly related to the content of propositions. On the other hand, we can try to Frege Kant (about a priori) and replace 'knowability' by ' justification'. It is a sound move that leads to taking the a priori as covering sentences-cum-a-priori-justification. More strictly: a sentence is a priori if its justification does not appeal to any sensory experience. I will assume this understanding in my further analysis, and, of course, I retain the concept of analyticity as defined *via* the content of sentences.

The apriority of logic was beyond questioning for Wittgenstein in *Tractatus:*

5.133 All inference takes place a priori.

[...]

5.4731 |...| That logic is a priori consists in the fact that that we *cannot* think illogically.

5.552 The "experience" we need to understand logic is not that such and such is the case, but that something *is'*, but that is *no* experience. Logic *precedes* every experience – that something is *so*. It is before the How, not before the What.

5.61 Logic fills the world: the limits of the world are also its limits.
We cannot therefore say in logic: This and this there is in the worried, that there is not.

6.3211 [...] as always, the a priori certain proves to be something purely logical [...]."

However, *Tractatus* leads to a puzzle concerning a priori. Consider the following statements from *Tractatus:*

2.225 There is no picture which is a priori true.

3.04 An a priori true thought would be one whose possibility guaranteed
its truth.

3.05 We could only know a priori that a thought is true if its truth was to be
recognized from the thought itself (without an object of comparison). [...]

5.55 We must now answer a priori the question as to all possible forms of the elementary propositions.
The elementary propositions consists of names. Since we cannot give the number of names with different meanings, we cannot give the composition of the elementary proposition. [...]

5.634 [...] No part of our experience is a priori. Everything we see could also be otherwise. Everything we can describe at all could also be otherwise. There is no order of things a priori. I...]

6.31 The so-called law of induction cannot in any case be a logical law, for it is obviously a significant proposition.- And therefore it cannot be a law a priori either.

6.32 The law of causality is not a law but the form of a law.

6.33 We do not *believe* a priori in a law of conservation, but we *know* a priori the possibility of a logical form.

6.34 All propositions, such as the law of causation, the law of continuity in nature, the law of least expediture in nature, etc., etc., all these are a priori intuitions of possible forms of the propositions in science.

6.35 Although the spots in our picture are geometrical figures, geometry can obviously say nothing about their actual form and position. But the network is *purely* geometrical, and its properties can be given a priori.
Laws, like the law of causation, etc., treat of the network and not of what the network describes."

The puzzle consists in a tension between Wittgenstein's account of logic and some theses included into the second sample derived from *Tractatus*. For, on the one hand, what Wittgenstein says about logic, suggests that it entirely fills the a priori territory, but on the other hand, 6.34 opens the door for a priori intuitions which are hardly reducible to logic. Thus, putting this in Kantian terminology, 2.225, 3.04 and 3.05 exclude the synthetic a priori, but 6.34 admits it. Moreover, 6.35 is sound only if geometry is a part of logic. In turn, it invites logicistic interpretation of Wittgenstein's philosophy of mathematics (see note 17). It is, however, a minor problem in this context, and let me return to general questions of Wittgensteinian account of a priori. One can say that Wittgenstein's idea of logic is close to Kant. It is true, but not logic is the key issue. The main problem is that we find in *Tractatus* no justification for a priori intuitions, other than purely logical. I do not intend to solve this trouble which, at least, points out how difficult the problem of a priori is, especially if one, like Frege and Wittgenstein, contrary to Kant, does not appeal to nativism or, like Husserl, to eidetic intuition.[44] Anyway, I will regard Wittgenstein (in *Tractatus)* as a predecessor of the linguistic account of a priori (see Introduction at the beginning of this paper). Even if there are other interpretative possibilities, this choice is historically justified because of Wittgenstein's role in the development of the Vienna Circle where the linguistic theory became standard.

Now we have sufficient data in order to show how AS and AP are relevant for epistemology.[45] As I already noticed, if AS and AP are combined together, we formally obtain four categories of sentences: analytic a priori, analytic a posteriori, synthetic a priori and synthetic a posteriori. Since, following Kant, analytic sentences are typically counted as a priori, the four categories are reducible to three ones: analytic, synthetic a priori and synthetic a posteriori. Thus, presumably, the a priori consists of the analytic and the synthetic a priori, and the a posteriori is equal to the synthetic a posteriori. Now, radical apriorism (Parmenides, Plato, Descartes, Leibniz) admits only the a priori as constituting the genuine knowledge, moderate apriorism (Kant) extends knowledge by the synthetic a posteriori, moderate aposteriorism (the Vienna Circle) excludes the synthetic a priori, and radical aposteriorism (Mill) reduces knowledge to the synthetic a posteriori. This map shows that the synthetic a priori is the key category, because it separates apriorism and aposteriorism, not only in general, but also in their more sophisticated, moderate forms.[46] In fact, defending a priori does not mean defending the synthetic a priori only, because moderate apriorism tolerates analytic sentences as legitimate.

The described combination of AS and AP operates on the level which considers sources of knowledge from a methodological point of view. Moreover, epistemology has always been concerned with genesis of knowledge (the sources of knowledge in the genetic or psychological sense). According to genetic empiricism, experience is the only way of knowledge acquisition. On the other hand, genetic rationalism ascribes a creative cognitive role to reason and typically admits that we have innate knowledge which acts independently of any experience.[47] It is naturally expected that nativism and apriorism are cooccurring on one hand, and empiricism and aposteriorism concur on the other hand. And now we encounter a heavy problem. Since it is assumed that experience cannot provide knowledge which is necessary, the position of aposteriorism creates a challenge: How is analytic (necessary) knowledge possible if every cognition has its beginning in experience?[48] The problem with apriorism is the same as that registered by Kant: How are synthetic a priori sentences possible? As aposteriorists claim: the way *via* nativism is too mysterious to be acceptable.

The problem of aposteriorism plagues its moderate as well as radical version, although not in the same manner. Let me quote a typical voice of a contemporary radical aposteriorist:

> "I would be inclined to believe (following J. S. Mill) that logical and mathematical truths don't differ in their origin from empirical truths -both are results of accumulated experience. A rough example. In a very early stage of their development, people learned to use the words 'not' and 'or'. In certain cases they were sure that something was white, in other cases that it was not white. In many cases they were first unable to decide whether a given thing was white or not (e. g., as a result of a bad light). But they noticed that in many such uncertain cases they finally reached a decision – by means of more thorough and repeated observations, better instruments, etc. Hence they began to believe in 'Everything is white or is not white' and, more generally, in '*p* or not *p*'. Of course, the whole problem belongs to the history of science, of human thought, and is not of a ^philosophical' nature; it may happen that I am completely wrong. In addition, the problem has no fundamental importance." (Tarski 1987, 31).

Contrary to Tarski, I think that the problem of the genesis of logic has a fundamental philosophical, particularly epistemological importance. If we accept (a) Hintikka's argument about the nature of consequence operation, (b) the thesis that experiential knowledge cannot provide absolute analytic sentences, we have a clear tension between (a)-(b) and Tarski's account of formation of logic. It is the main problem for radical aposteriorism: since this view excludes analytic sentences, it seems inconsistent with the very nature of logic, and perhaps some parts of mathematics. On the other hand, moderate aposteriorism must explain how analytic sentences are possible, provided that only experience is the source of knowledge. Thus, we have in fact two possibility questions: (a) How are synthetic a priori sentences possible?, and (b) How are analytic sentences possible? Although both questions concern the a priori, (a) is characteristic of apriorism, but (b) bothers philosophers advocating aposteriorism. In sum, there is a trilemma: either adopt the synthetic a priori (apriorism) or stay with the analytic and the synthetic a posteriori (moderate aposteriorism) or restrict knowledge to the synthetic a posteriori (radical aposteriorism). All options have pluses and minuses. As I have already noted (see Introduction), I will defend moderate aposteriorism, although I do not regard this view as unproblematic. However, I insists once more (see note 44) that separation of genetic and methodological issues connected with empiricism, rationalism, apriorism and aposteriorism is fatal for any reasonable discussion of the related matters.

In order to have a point of reference for a further discussion, it is convenient to have a sample of proposals concerning a priori sentences. I will use a collection made by Wang (1974, 261; I slightly change some formulations): (i) 7 + 5 = 12, (ii) if a point *a* is between *b* and c, then *b* is between *a* and *c,* (iii) existence is not an attribute, (iv) the world is four dimensional, (v) hypocrisy is not red, (vi) there exists something, (vii) $\exists(x=x)$, (viii) every property of positive integers which is expressible in he ordinary theory of numbers defines a class of positive integers, (ix) a cube has twelve edges, (x) I cannot be you, (xi) I could not have been born 15 years later, (xii) 'Frau' is a German word, (xiii) Cantor was the discoverer of the Cantor theorem, (xiv) Wessel was the discoverer of Argand's diagram, (xv) this sentence is true, (xvi) every tone has an intensity and pitch, (xvii) one and the same surface cannot be simultaneously red and green all over, (xviii) spiritual values have a higher place in the scale of values than vital values, (xix) the sum of the internal angles of a triangle is 180°, (xx) the sum of internal angles of a triangle is less than 180°, (xxi) a sentence in a formal system expressing the consistency of this system, (xxii) the principle of mathematical induction, (xxiii) any two things differ in at least finite number of properties, (xxiv) the empty set is different from its unit set, (xxv) every sentence is either true or false, (xxvi) if *A* is analytic and *A* entails *B*, then *B* is analytic, (xxvii) there exists an infinite set, (xxviii) every sentence has a verb, (xxix) $x + 0 = x$, $x + (y + 1) = (x + y) + 1$, (xxx) space is three-dimensional, (xxxi) the best men get what they want, (xxxii) the best men want what they can get, (xxxiii) the law of conservation of energy, (xxxiv) a continuous line joining a point inside a circle to a point outside intersects the circle, (xxxv) the earth did not come into existence five minutes ago, (xxxvi) there is a past, (xxxvii) $e = 1/2\mathrm{mv}^2$, (xxxviii) $f =$ ma, (xxxix) $e = mc^2 + l/2mv^2 +$ .... (xl) two spheres cannot differ only numerically, (xLi) thought is not laryngeal motion, (xii) unity is not a quantity, (xliii) number is not the thing that is counted, (xliv) the difference between two degrees of quality is not itself a quality, (xlv) I ought to promote my own good on the whole (where no one else's good is affected), (xlvi) if I ought to do something, than I can do it, (xlvii) the axioms of prudence, benevolence, and equity, (xlviii) I ought to regard a larger good for society in general as of more intrinsic value than a smaller good, (xlix) one man's good is (other things being equal) of as much intrinsic value an any other man's, (l) the world is a system of necessarily connected parts, (li) an individual is a set of characters, (lii) the characters of an individual are not all equally essential, (liii) things have manifold necessary relations to other things, (liv) every event has a cause, (lv) the world is infinite, (lvi) I see with my eyes, (lvii) the world is finite but unbounded, (lviii) what is done cannot be undone.

It is important to realize that Hao Wang lists examples of sentences claimed to be a priori, analytic or necessary in philosophical debates, not only synthetic a priori ones, although it contains items proposed to serve in this role. I am not able to localize all examples by attributing them to concrete proponents. However, I will briefly comment every case.[49] Thus, (i) and (iii) are Kantian and synthetic a priori for him; (ii) expresses the transitivity of the relation of betweeness and may be interpreted in various ways – for example, Frege and Russell considered it as analytic; (iv) is an application of geometry to physics and can be interpreted in various ways, for example as a convention, (v) is an example of a common sense indubitable, perhaps even necessary truth in a sense; (vi) is a metalogical assumption of classical first-order logic, and it worried Russell as introducing extralogical existential element into logic; (vii) is derivable in first-order logic with identity and it maps problems connected with (vi) into the object language (recall that for some philosophers, no analytic sentence is existential); (viii) is a comprehension axiom for elementary number theory formulated in predicative set theory, and its status depends how set theory is understood – for Russell it was analytic, for Gödel not; (ix) is a definition and may serve as an example of truth by convention; (x) and (xi) are like (v); (xii) is a metalinguistic statement – truly speaking, I do not know why it was included in the list; (xiii) and (xiv) resemble Kripke's examples of necessary sentences known a posteriori (see below); (xv) is a self-referential sentence, but it does not create a paradox, contrary to the Liar, that is, 'this sentence is false'; (xvi) and (xvii) are favourite examples of the synthetic a priori for Husserl and other phenomenologists; (xviii) is an example of an ethical synthetic a priori principle, but only for those who accept such a view; (xix) is synthetic a priori for Kant and Frege, conventional for Poincare, and analytic for Russell; (xx) is true in Non-Euclidean geometry, conventional for Poincare and analytic for Russell; (xxi) is synthetic for Gödel, but see below; (xxii) is analytic for Frege and Russell, but synthetic a priori for Poincare; (xxiii) is perhaps a metaphysical view, but I think about this example as similar to (xii); (xxiv) is a theorem of set theory – see a remark on (viii); (xxv) expresses the principle of bivalence and its interpretation depends on the view about metalogical principle – Russell and Carnap regarded it as analytic, but Tarski as conventional and perhaps forced by experience, at least in some applications; (xxvi) is sometimes considered as a metalogical principle of analyticity, for example, in R. M. Martin's approach (see (35) above) who regards it as analytic, (xxvii) is the axiom of infinity – it worried Russell for the same reasons as (vi) did; (xxviii) is a grammatical principle, and its interpretation depends on the view on grammar – for Chomsky, it is an element of rational, that is, a priori grammatical theory; (xxix) gives the recursive definition of addition in arithmetic of natural numbers and is understood relatively to the view about arithmetic, for example, as analytic by Carnap; (xxxi) and (xxxii) are moral principles (see comments on (xviii)); (xxx) is a priori on Kant's view, but conventional according to Poincare; (xxxiii) is a physical law, and its interpretation depends on how physics is understood – the concept of nomic necessity is sometimes applied to examples of this sort, but a conventionalist account is also possible; (xxxiv) is a geometrical statement (see comments on (xix)); (xxxv) and (xxxvi) give another common sense indubitable truth (see (v), (x) and (xi) above); (xxxvii), (xxxviii) and (xxxix) express physical principles, but true in different physical theories – (xxxvii) and (xxxix) in classical mechanics, (xxxviii) in special relativity (see comments on (xxxiii)); (xl) gives another geometrical example (see comments on (xix)); (xli) formulates a principle of ontology of thought, a priori for Descartes; (xlii), (xliii) and (xliv) are typical ontological principles, very often considered as synthetic a priori, for example, by phenomenologists; (xlv) states the basic principle of moderate egoism (see comments on (xvlii)); (xlvi) captures a famous principle that ought implies can, synthetic a priori for Kant; (xlvii), (xlviii) and (xlil) are ethical principles (see comments on (xviii); (l), (li), (hi) and (liii) express various metaphysical principles, sometimes accepted as a priori, sometimes considered as inductive generalizations, but also sometimes rejected as meaningless; (liv) is a famous principle of causality, synthetic a priori for Kant, eventually analytic for the Vienna Circle; (lvi) and (lvii) possible cosmological models of the world, the latter is closely connected with general relativity (see comments on (xxxiii)); (lvi) is a sentence which would be regarded as analytic by Carnap, as a consequence of the definition of 'seeing'; (lviii) formulates the view that the past cannot be changed which is a popular way of expressing the thesis that truth is eternal, and various interpretation of this proposition are possible.

The above list can be easily extended by adding further examples, including some famous philosophical statements, like *Ex nihilo nihil fit* (coming back to Parmenides), 'Being exists, but Not-Being cannot exist' (Parmenides), *Cogito, ergo* *sum* (Descartes), *Pacta sunt servanda* (Grotius), the principle of sufficient reason (Leibniz), 'Is does not entail ought' (Hume), the categorical imperative (Kant), 'no mental act without a mentalese substance' (Brentano, a generalized Cartesianism), 'content is a metaphysical part of mental acts' (Twardowski) 'Perfect being must exist' (all advocates of ontological proof of God's existence) or 'The realization of Good is good' (Scheler). Each of (i)-(lvii) as well as additional examples require a careful analysis and admit different and mutually conflicting interpretations.

The a priori status of the above examples is defended in various ways. The main justifications of the a priori are as follows:[50] (A) Kantian transcendentalism; (B) psychologism (a priori is generated by human psychic constitution; this position was represented by Fries as a response to Kant); (C) pragmatism (a priori is generated by schemes organizing experience; this position was maintained by C. I. Lewis); (D) linguisticism (the a priori and the analytic are coextensive; this position was one of the main views of logical empiricism); (E) conventionalism (the a priori is generated by conventions; Poincaré and Pap defend this view); (F) anthropologism (the a priori is generated by forms of human life; later Wittgenstein proposed this view); (G) intuitionism (the a priori is generated by self-evidence of some truth; this view was initiated by Brentano and recently developed by Chisholm); (H) rationality-based apriorism (the a priori is generated by human rationality which prohibits to deny a priori sentences in concrete cognitive situations; Putnam defends this account); (I) evolutionism (the a priori is forced by phylogenetic properties of the human kind as opposed to ontogenetic properties of particular individuals; Spencer offered this explanation, following evolutionary theory of Darwin); (J) ontologism (the a priori is rooted in real essential relations holding between objects; Meinong, Husserl, Scheler, and N. Hartmann represented this view usually called 'the theory of the material a priori' superstructured on a special theory of mental acts penetrating essential necessary dependencies).[51]

The variety produced by (A)-(J) consists of views not identical in their principal claims. In particular, not every position consists in defending the synthetic a priori. Linguisticism rejects radically the synthetic a priori, and identifies analyticity and apriority *via* linguistic conventions. Ontologism defends the view that synthetic a priori elements occur in our knowledge, but their source is unclear. Linguistic conventionalism, like Lewis' pragmatism, supports (D), but Poincaré's view accepted the synthetic a priori, exemplified by the principle of mathematical induction. Similarly, psychologism, anthropologism, intuitionism, rationality-based apriorism and evolutionism are coherent with rejecting the synthetic a priori and its admitting as well. Many views were invented in order to show that (D) is incorrect without accepting the synthetic a priori (Pap is perhaps the most clear example of this strategy).

Since I will try to defend moderate aposteriorism which is normally associated with (D), let me characterize linguisticism more closely. The *locus classicus* is this:

> "In saying that the certainty *of a priori* propositions depends upon the fact that they are tautologies, I use the word 'tautology' in such a way that a proposition can be said to be a tautology if it is analytic; and I hold that a proposition is analytic if it is true solely in virtue of the meaning of its constituent symbols, and cannot therefore be either confirmed or refuted by any act of experience [...] I say that validity of *a priori* propositions depends upon certain facts about verbal usage." (Ayer 1946, 21-22).

Two points of linguisticism can be derived from Ayer's explanation. Firstly, disregarding identification of analyticals with tautologies, this view considers all a priori sentences as analytic (the analytic thesis, according to Quinton 1963-1964, 31). Secondly, it takes linguistic conventions as the basis of the a priori. It means that in order to justify a priori sentences one must appeal to conventions governing linguistic usages. These two ingredients of linguisticism operate, however, on different levels. It becomes clear when we analyze typical objections against the linguistic theory of the a priori. Pap (1955) argues that this theory fails because the necessity of logical principles is just prior to linguistic conventions. Moreover, one can also observe that linguistic conventions are recorded on the base of experience (see Broad 1936). Thus, Pap tries to point out that linguistic conventions are not effective in a priori justifications, but Broad suggests that linguisticism traits itself by appealing to conventionalism because justifies apriora by experience. I will argue that (a) the analytic thesis is defensible *via* the complex idea of analyticity which developed above, (b) there is an account of a priori justification which is consistent with (a) and genetic empiricism. It is clear that my account relies heavily on the observation that AP basically connects sentences with their justification. It means that two things are to be checked: the semantic status of apriora and their justification as a priori, because every theory of apriora has to propose at least two theses: the semantic thesis (apriora are analytic or synthetic) and the justification thesis (justification a priori consists in this or that).[52]

My objection to the traditional route of linguisticism is that it either concentrates on the analytic thesis or on justification thesis but not on both together. Incidentally, the term 'linguisticism' obscures the situation because it marginalizes the analytic thesis. Although some other label, for instance, 'the analytic theory of a priori' would be less misleading, will preserve the traditional vocabulary. Defending moderate aposteriorism I am aware that the radical version of this view constantly reappears in philosophical debates and finds its support in various facts, for instance, in applications of computers in mathematical proofs. However, I will not discuss this possibility very much. The main reason is once again Hintikka's argument in favour of stability of consequence operation, which property seems to me incoherent with radical aposteriorism. However, I admit that it can appear to be plausible position. In any case, I believe that any form of aposteriorism should be preferred over any kind of apriorism.

I start with arguments against (a). It is convenient to explore attempts based on metamathematical results. Copi (1949) defends the synthetic a priori by the first Gödel incompleteness theorem. First, he defines analyticity in the following way. Now, every sufficiently rich language, that is, language sufficient for elementary number theory contains non-empirical, non-inductive propositions which are not decidable by syntactical rules of that language. Thus, these undecidable sentences are not analytic. However, by the principle of excluded middle, there is one non-analytic truth which is also non-empirical and non-inductive, therefore synthetic a priori, which destroys, according to Copi, the linguistic theory of a priori.[53] This conclusion is, however, too fast from the point of view of the present paper. We can save the linguistic theory by a simple observation that undecidable true propositions are relative semantic analytic sentences of arithmetic and also pragmatic analyticals with respect to the standard model.[54] Copi's argument is rather against logicism, that is, reduction of mathematics to logic than against linguistic theory of the a priori.

Another metamathematical argument for the synthetic a priori was given by DeLong (1970):

> "[...] *C* [= *CON(AR)* – J. W.] under its intended interpretation could reasonably be classified as *synthetic a priori,* synthetic because it does not follow by definition and general logic, and *a priori* because if arithmetic is consistent, it must be necessarily consistent." (222)

This argument brings necessity into the a priori business. Since I cannot enter into the problem of necessity and its various interpretations, I will adopt the understanding on which *A* is necessary in the world W* (intuitively 'necessary in the real world') if and only if *A* is true in every world W such that *W* is accessible from W*. Let M* be the standard model (the real world) of **AR.** Now M' is an accessible model from M* if M' satisfies arithmetical axioms. Thus, non-standard models of arithmetic are accessible from the standard one. Given our assumptions, the formula CON(**AR**) is necessary in M* if and only if CON(**AR**) is true in every model of **AR.** However, it cannot be true in every model accessible from M* for its unprovability, and it is, by definition, not necessary, at least in the above established meaning. We assume that *CON*(***AR***) is true in M*, but it is done not *simpliciter* a priori, but relatively to pragmatic criteria of standardness. DeLong implicitly introduces this path when he speaks about *C* in its intended interpretation. He could eventually say that the matter concerned *CON*(***AR***), but metaarithmetical statement asserting the consistency of arithmetic serving as a general assumption of both Gödel theorems of incompleteness. However, this change does not help for simple reasons. Either we consider *CON*(***AR***) (in its arithmetical code) as an adequate expression of the sentence 'Arithmetic is consistent' or not. When we argue for the first option, the outlined argument against DeLong is valid, and when we choose the second option, we must discuss the status of 'Arithmetic is consistent' in the metaarithmetical conceptual scheme, but we must begin with criteria of analyticity for it and the whole game starts once again. On the other hand, relative analyticity immediately provides a solution.

A still different attempt of proving by metamathematics that synthetic a priori sentences exist is suggested by Castonguay (1976, 86).[55] He claims that the Church theorem (arithmetic is undecidable) together with the Church thesis (intuitively speaking: intuitive decidability = recursivity) implies that mathematical knowledge is synthetic a priori. Castonguay assumes that analyticity is defined by decidability, that is, he identifies analytic sentences with syntactic analytic sentences defined by (62) or (64). I tried to argue that this account is too narrow, and it should be considerably extended. If I am right, Castonguay's argument proves only that mathematical knowledge is not reducible to purely algorithmic procedures, but not that it is synthetic a priori. Moreover, Castonguay must direct his conclusion to mathematical proofs because it has no clear meaning with respect to mathematical theorems for we are not able to separate decidable and undecidable cases. In fact, undecidability means that we have no general algorithmic criterion to decide every concrete formula as provable (valid) or not. Thus, Castonguay's criterion cannot yield any synthetic a priori sentence.[56] According to the idea of gradualism (see above), proceeding across particular degrees introduces some amount of syntheticity, but is does not need to be interpreted as a commitment to synthetic a priori sentences, at least in the traditional sense. This completes a survey of arguments for the synthetic a priori derived from metamathematics. I conclude that none of the mentioned arguments compels us to accept that synthetic a priori sentences exist because all examples can be interpreted as pragmatic analyticals.

I guess without a detailed analysis that all other cases of sentences that are proposed as a priori and non-analytic, can be treated in the same way.[57] It concerns, for example, sentences about colours and related phenomena (see (xvi), (xvii) and perhaps (xliv) in Wang's list). These examples admit various interpretations (see Delius (1963) and Hardin (1988)), including attempts to show that they are analytic, synthetic a posteriori or synthetic a priori. I am inclined to consider sentences about colours, tones, qualities, etc. as pragmatically analytic in the common-sensical conceptual scheme. As a matter of fact, any person sufficiently mastered in ordinary language would accept (xvii) as true on the base of his or her linguistic competence. The example with tones is more complex, but perhaps even more instructive. Since the concept of tone is more specialized than that of colour, (xvi) seems to be more clearly dependent on the knowledge of meanings than (xvii). However, the assertion of both sentences can be also interpreted as closely related to the development of personal experience of the users of language, and consequently that sentences become synthetic a posteriori; the same remarks apply to (xliv), except that this sentence is only on the tongue of philosophers, not ordinary people. Anyway, the synthetic a priori interpretation of (xvi), (xvii) and (xliv) is not forced by any standards. Metalogical principles ((xxv), (xxvi), and even (Lviii) are pragmatically analytic in an informal metalanguage. The sentence (xlvi) can be formalized in a deontic logic and thereby also becomes pragmatic analyticals, (xlii) is a consequence of appropriate conventions concerning unities and quantities, and (lvi) seems analytic as a result of the definition of seeing.[58]

Pap (1946) defends the a priori in physics. As I already noticed, it does not mean that he admits the synthetic a priori in physics, but his task is to show that physical theories contain aprioristic statements which are not analytic. Thus, Pap aims at the linguistic theory of a priori his criticism, finding this account untenable. Consequently, my goal consists in arguing that Pap's account is consistent with the thesis that all apriora are analytic. Pap says:

> "The theory of *a priori* which will, in this essay, be presented and applied to physical principles, may be called *functional* in so far as the *a priori* is characterized in terms of functions which propositions may perform in existential inquiry, no matter whether they be, on formal grounds, classified as analytic or synthetic. It may also be called *contextual* for statements of the form "x is a priori" or "x is a posteriori" (where the admissible values of x are propositions) will be treated as elliptical or incomplete. A proposition which is a priori in one context of inquiry, may be a posteriori in another context." (viii)

Pap's view is based on a dynamic account of science. He uses a well-known fact that elements of physical theories change their epistemic status. In particular, inductive generalizations are transformed into conventions, and definitions have empirical origin.[59] Newton's laws of motion, the principles of the theory of relativity (special and general), and even the law of causality can be interpreted as functionally a priori. Usually, conventionalization requires idealization. The law of inertia is a good example here. It is an observational generalization that material object are always subjected to external forces. Passing from observation to the theoretical principle 'If no external forces act upon a body, it will continue in its state or motion with uniform velocity along a straight line' consists in introducing an ideal situation marked here by a counterfactual conditional which expresses the law of inertia. Thus, the phenomenon of inertia is once described a posteriori, and once again postulated a priori. Pap is not explicit about the status of the results of conventionalization in the family determined by combining AS and AP. He only says that conventionalized truths are not analytic because their denials are not contradictory. I consider this explanation to be not sufficient, unless one introduces a new category of sentences. If we stay with the analytic, the synthetic a posteriori and the synthetic a priori as categories derived from an exhaustive and mutually exclusive division of sentences, theoretical principles, regardless of whether there are interpreted statically or not, or functionally or otherwise, must belong somewhere. On my account of analyticity, they can be interpreted as pragmatic analyticals in related conceptual schemes.

Although the question how results of conventionalization should be qualified, that is, whether they are analytic or synthetic. Pap's account moves us to the second part of my defence of linguisticism. Clearly, conventionalization is a process or device which changes the status of elements of knowledge. The justification aspect of the a priori brings new problems which go beyond the analytic thesis. Not everything proper for analyticals can be mapped onto the a priori. In particular, it is problematic how the distinction of syntactic, semantic and pragmatic analyticals could be applied to the variety of a priori items, because qualifications 'a priori' and 'a posteriori' are based on quite different standards than attributions of analyticity and syntheticity. However, on other distinction, namely that between absolute and relative analytic sentences provides a heuristic hint for a promising treatment apriora. Saying this, I do not want to suggest that we should automatically transform (63)-(66) into definitions of absolute and relative a priori. I only maintain that the path based on the distinction between absolute and relative meaning of epistemological categories might be fruitful also in the case of the a priori.

Kant's understanding of apriora was definitely absolutistic, because he defined a priori as independent of any experience. Let me repeat that he was conscious of the traditional view on which the a priori was related to a way of justification (from causes to effects). However, there is an ambiguity in Kant's account of the a priori that was pointed out by Reichenbach (1920, Ch. V). Kant at one point says that what is a priori is absolutely valid (for any circumstances), but later he ascribes this category to acts of constitution of objects. These various usages of 'a priori' suggest looking for its relative understanding. A good analogy can be derived from mathematical statistics. Assume that we have a population and we are interested in a statistical distribution of a parameter in that collection of objects. We can select a sample and answer the question a posteriori. However, it is also possible to formulate a statistical hypothesis a priori, that is, before any experience concerning the population in question. It is customary to use the label 'prior probability' (probability a priori) and 'posterior probability' (probability a posteriori) respectively. Using this analogy, we can say that having an experience E, a sentence a priori with respect to E is a sentence A which is (a) involved in a conceptual scheme concerning E, and (b) such that A is accepted before E is employed in its justification. If A is justified by E, we say that it is a posteriori with respect to E. It does not mean that sentences a priori with respect to a particular *E* are completely independent of any experience. Like in the case of probability, they are independent of this concrete experience which is considered.[60]

Similarly to the previous treatment of analyticity, we can relate the concept of the apriori to conceptual schemes. It allows to reformulate (Tl), (T2) and (T3) for AP:

(T1') There is a sentence *A* that for any conceptual scheme **SCH** and any empirical data *E, A* is a priori with respect to **SCH** and *E,*

(T2') For any conceptual scheme **SCH** and empirical data *E* involved in **SCH** there is a sentence *A* such that *A* is a priori with respect to **SCH** and *E,*

(T3') If *A* is an a priori sentence, there is a conceptual scheme **SCH** and empirical data *E* such that *A* is a priori with respect to **SCH** and *E,*

Two things are to be immediately observed here. If we want to keep AP as somehow autonomous in relation to AS, we cannot transform (63)-(67) to fit the case of apriora, but it was of course to be expected. Secondly, (T1')-(T3') are more complex than (T1)-(T3) for more parameters (*E* is added) are involved. This complexity shows that the analytic thesis is supplemented by the justification thesis. According to the first circumstance, we must use (T1') to expose our intuitions. Let me start with (T2'). It can be interpreted as saying that any mature conceptual scheme must contain some a priori elements. It concurs with claims, made by Ajdukiewicz and Pap (and numerous other authors as well) that every theoretical system has principles a priori, not necessary synthetic apriora. At first sight, (T1') defines the concept of the absolute a priori. However, it is correct only when we limit the concept of a priori to the analytic thesis. The justification thesis radically changes the situation. If we restrict the problem of justification of logic to provability alone, any appeal to metalogical decisions is redundant. Theorems of logic are justified internally, that is, within logical systems, either semantically as universally valid or syntactically by provability from selected axioms. However, we also need to explain why this or that consequence operation is selected as proper, and it brings apriority of logic. If one wants to qualify this kind of apriori as absolute, there is nothing wrong with that. The important thing is that nothing compel us to count anything except logic in the domain of the "absolute" a priori. On my part, I am inclined to the view that even apriority of logic is relative with respect to the justification thesis.[61] This position entails that (Tl') becomes a special case of (T3') which determines a general concept of relative a priori sentences.

Now it is immediately obvious that the analytic thesis is sustained.[62] Thus, this part of linguisticism and moderate aposteriorism is justified. Unfortunately, we have troubles with the second part because genetic empiricism is at odds with properties of absolute analytic sentences, that is, theorems of logic. The key problem consists in explaining how empirically acquired information is transformed into logical tautologies. Clearly, not very much is achieved here by pointing out that every a priori justification is relative, although, on the other hand, the situation of the empiricists using the relative a priori is easier than their position associated with the absolutistic concept of the a priori. I confess that I have no straightforward answer. Presumably, three suggestions which do not need to be considered as mutually exclusive, appear here. Firstly, we can say that every a priori justification begins with an experience just because it is relative, that is, a priori with respect to a given conceptual scheme. This means that the relative a priori (as attributed to justification) supports absolute analytic sentences. Secondly, we could look for an explanation how experience in its traditional shape generates logic (compare Tarski's view quoted above). Thirdly, we can try to enlarge the scope of experience by adding its new sources, for example, intuitive or semantic acts constituting a special metalogical experience leading to absolute analytic sentences. The last solution is additionally motivated by the role of metalogical insights in selecting logic. Each solution can be supplemented by evolutionary epistemology.[63] Perhaps it might be helpful to explain why a stable consequence operation is cognitively important. A hint for an argument can be found in probability calculus. One of its axioms says that probability of tautologious events is equal to 1. The rationale of this principle consists in blocking of the maximal diffusion of probability values ascribed to uncertain events to the effect that all probabilities would be equal. Using this analogy, we can say that truth-preserving consequence operation (or equivalently: theories solely consisting of tautologies) was "invented" by human kind through biological evolution in order to protect already accumulated content against excessive diffusion. According to any solution of the problem of the a priori consistent with genetic empiricism, nothing compels us to accept that synthetic a priori sentences exist, unless we say, like in the case of analyticity, that relative apriority which results with relative analyticity always introduces an amount of syntheticity.[64] Like in the case of analyticity, gradualism is also proper for apriority, but nothing dangerous follows for moderate aposteriorism from this account.

Let me show how gradualism with respect to the a priori works by a concrete example. Stenius (1965, 83) defends the following principle:

(S) If *A* is a matter of empirical observation and, therefore, can be known as an a posteriori truth, then *A* cannot be an a priori truth.

The philosophical relevance of (S) is serious because it expresses that no synthetic sentence can be a priori. However, (S) is true only if the absolutist conception of the a priori adopted. Under gradualism and admitting the relative a priori, there is no incoherence in maintaining that we have a priori truths in one conceptual scheme which are a posteriori in another theories.

In order to conclude my defence of moderate aposteriorism I need to consider the relation of the a priori to the necessary. The full analysis of this topic requires taking into account various understandings of necessity, in particular logical, real, deontic, *de dicto, de re,* etc. Since it involves problems going far beyond AS and AP, I must considerably restrict the scope of this (final) part of the present paper. I will mainly focus upon particular examples which suggest that we should abandon the view, which I accept, that the analytic, the a priori and the necessary are coextensional. The problem became popular after Kripke's influential (1971, 1972).[65] He argues that there are contingent a priori truths as well as necessary a posteriori truths.[66] A sentence (Kripke (1972, 56)

(68) Stick *S* is one meter long at $t_{0,}$

is an example of a contingent apriora. On the other hand, the sentence (Kripke (1972, 109-110)

(69) Hesperus is Phosphorus,

is considered as necessary a posteriori. Kripke's arguments for contingency of (68) and aposteriority of (69) are rooted in his views on rigid designators and essentialism. Leaving these controversial matters aside, let me only observe that Kripke seems to understand necessity always in the absolute sense. However, if we divide analyticals into absolute and relative, and analyze apriora as relative with respect to the justification thesis, it suggests that there is also a reason to qualify necessities as absolute and relative. Roughly speaking, *A* is absolutely necessary if and only if it is true in all possible worlds (modulo semantic interpretation of logical constants), but *A* is relatively necessary if and only if *A* is true in all possible worlds selected by theoretical or other postulates, for example, by constraints concerning references. That 'Hesperus is Hesperus' is absolutely necessary by logic (logical truths and their instantiations are always necessary), but (69) is only relatively necessary *via* additional semantic referential clauses.[67]

Kaplan (1979) proposes still another counterexample, namely

(70) I am here,

as analytic and contingent, but it is also fairly plausible to consider as relatively necessary. Thus, we have ways out in order to preserve our conceptual equivalencies and defend moderate aposteriorism as the view which says that every knowledge starts with experience, but some truths are accepted before working with concrete conceptual schemes used for ordering empirical data.

Jan Woleński
The Jagiellonian University

NOTES

[1] See Chapter 1 in the present volume for the history of this issue.

[2] In what follows, I will use the terms 'proposition', 'statement' and 'sentence' interchangeably. The last will be employed more frequently than the rest. In order to avoid some misunderstandings, I note that sentences are understood here as well-formed grammatical units always equipped with meaning.

[3] We have also other related distinctions, for example 'necessary versus possible', 'verbal versus real', 'conventional (or conceptual) versus factual' or 'essential versus accidental'. Perhaps the first distinction, already mentioned, mostly ontological or metaphysical, is of a special importance of us. It is due to a popular definition of analytic truth as true in all possible worlds, that is, by necessity. See also note 52 and the last fragment of this paper.

[4] The literature concerning the problems discussed in the present chapter is enormous. Although I will refer to many views about AS and AP, my survey cannot be complete. I regret that I cannot discuss several interesting suggestions and debates. In some cases, I mention sources in which the reader can find additional information.

[5] See historical remarks in Chapter 1 of this volume, Ueberweg (1857, sec. 83), Gewirth (1953), Pap (1958, Part I) and Proust (1989). In Chapter 1 of this *Handbook*, I tried to interpret some views of pre-Kantian philosophers in terms of analyticity. I will come back to this question in the next sections.

[6] Page-references are to translations or reprints, if they are mentioned in Bibliography.

[7] See Dubislav (1926), Marc-Wogau (1957), Pap (1958, Ch. 2), Proust (43-48).

[8] I do not enter into the question, interesting in itself, whether the relation between *S* and *P* should be taken extensionally or intensionally. Kant's text does not decide this question, although it seems that he had in mind a relation between contents of subject and predicate.

Thus, mathematical symbolism used in interpretation of Kant's view must be taken *cum grano salis.*

[9] However, this conclusion must be somehow qualified. Assume that one will read (Kl) as a semantic definition of analyticity, but (K2) as a syntactic one. It is known that semantic is richer than syntax. Hence, it is not a priori certain that analyticity determined by semantic criterion is extensionally equivalent with analyticity generated by syntax. I come back to this question at the end of this section.

[10] One warning is here in order. Speaking about logical aspects of Kant's views is ambiguous. He had two understandings of logic (see comments by Proust (1989, IV). First, logic was understood by him as formal, that is, independent of content. Secondly, he also appealed to transcendental logic as dealing with production of knowledge. The role of transcendental logic is particularly important in analysis of the synthetic a priori; see historical introduction to this volume, section on Kant.

[11] The numeration (1)-(3) used by Bolzano applies only to this fragment of the present paper.

[12] See Dubislav (1926), Bar-Hillel (1950), Berg (1962, 95-102), Berg (1973, 18-20), Morscher (1973, 222), Proust (1989, 49-108), p. 287, Berg (1992), pp. 79-83, Siebel (1996, ch. 4), Textor (1996, 248-255).

[13] I do not decide whether there is a sharp boundary here. However, it seems to me that it is an important problem, usually neglected by commentaries about Frege. For instance, Proust (1989, 122-133) focuses only on formal correctness of definitions.

[14] Except for the problem of the status of definitions as premises in generating analytic truths, Frege's conception of analyticity is fairly clear. Hence, I do not refer to commentators who usually report what Frege said. Some remarks broadening Frege's perspective are included in Dummett (1991, 23-46).

[15] Note, however, that there is no necessary *iunctim* between logicism and the thesis that logical and mathematical truths are analytic. Logicism requires that logic and mathematics have the same status, i.e., their theorems are either analytic or synthetic, or possess still different semantic characteristic. The last reservation is important for Quine's logicism, but he is also against the AS distinction. I restrict my remarks about logicism to minimum because this topic is extensively discussed in Murawski's chapter in this volume.

[16] Russell seems to ascribe logicism to Wittgenstein. See Russell's Introduction to Wittgenstein (1922, 21)

[17] The differences between Russell and Wittgenstein at this point are not reducible to the problem of logicism in mathematics. In fact, Wittgenstein's text is far from clarity in some respects. At one point, he says that tautologies are propositions (see 6.1), but another time, that they are not (see 4.463, not quoted in my text) for they are not a model of reality. Also, Wittgenstein changed his earlier view about analytic propositions. In his (1961, 21; this text was written in 1914-1916) he says: "There are no such things as analytic *propositions.*" For Russell, tautologies were always fully legitimate propositions. In order to complete the discussions of Russell and Wittgenstein, let me add two historical remarks. First, Wittgenstein's worries about the meaningfulness of tautologies reappeared in the early Vienna Circle, where the first definitions of meaningfulness excluded this sort of expressions. Second, it is interesting that Russell and Wittgenstein used the term 'analytic proposition' only occasionally, and the label 'tautology' more often. See very stimulating remarks about the concept of tautology and its history in Dreben and Floyd (1991). Also see Marion (1998) for recent evaluation of Wittgenstein's philosophy of mathematics. See also the discussion about Wittgenstein's account of apriority in section 3 below.

[18] The list (1)-(13) brings together items mentioned by Mates (1951) and Nordenstam (1972), plus definitions mentioned earlier. (12) was taken from Nordenstam but reformulated in order to make it closer to Lewis' own language. The Mates-Nordenstam list is not

complete, because both authors casually directed their surveys in order to discuss Quine's objections (see below). Hence, I decided to mention several other proposals.

[19] For further comments and remarks see Quine (1963), Bohnert (1963), Butrick (1970), Creath (1987), Friedman (1988), Proust (1989, 167-240), Creath (1990a), Quine (1990), Creath (1992), Tennant (1991), Sarkar (1992), Oberdan (1993), Cirera (1994), and Carnap (1963), (1963a), (1990).

[20] See also Carnap (1966, 265-274). For comments, see Hempel (1963, 703-707) and Carnap's reply in Carnap (1963b, 963-966).

[21] Pollock uses a special sort of implication, but nothing changes if we say that (a)-(e) are provable material implications.

[22] See van Benthem (1984) and Stenius (1984) for a further discussion.

[23] See van Benthem (1974), Castonguay (1976), and Rantala and Tselishchev (1987) for comments and criticism of Hintikka views on analyticity. Also compare Hintikka (1987).

[24] See Sauer (1986), Lemos (1997) and David (1997) for comments. Chisholm proposed several variants of this definition of analyticity.

[25] The list (14)-(62) is arranged chronologically. The last three items break this order, but it is a deliberate choice which will become clear in further considerations.

[26] Note, however, that I minimalized intuitions coming to AS from AP. Both distinctions are taken together in section 3.

[27] The literature about Quine's criticism of the AS distinction is enormous. I cannot review even a small sample of it. My plan is as follows. At first, I will report Quine's arguments from (1951), except his objections against the verifiability conception of meaning (the second dogma of logical empiricism). Then, I will mention some counterarguments against Quine. For a further discussion, see Mates (1951), Perkins and Singer (1951), Martin (1952), Kemeny (1952), (1952a), Gewirth (1953), Wang (1955), Grice and Strawson (1956), Stegmüller (1957, 291-319), Pasch (1958, 11-23), Bennett (1959), Martin (1959, 107-113), Putnam (1962), Quine (1963), Carnap (1963), Kemeny (1964), Aune (1972), Katz (1972, 243-260), Nordenstam (1972), Haack (1977), Orenstein (1977, Ch. 5-6), Priest (1979), Gochet (1986, Ch. I), Bohnert (1986), Hellman (1986), Carnap (1990), Creath (1990a), Katz (1992), Moses (1992), Parsons (1995), Harman (1996), Boghossian (1997), Müller *(1998 passim).* Miller (1998, Ch. 4), Hintikka (1999) and Quine (1986), (1986a), (1991). Quines critic was anticipated to some extent by Tarski (1936). Related criticism, partly inspired by Quine, is to be found in Goodman (1950) and White (1950), (1956). A particularly fascinating material is included in Creath (1990).

[28] The way of defining analyticity *via* the concept of stimulus was anticipated by Perkins and Singer (1951).

[29] Reporting arguments against Quine, I take into account only general ones. For example, I omit those in Katz (1972) because they require entering into details of Katz's semantic theory. I also include my own assessments of certain points raised in the discussion. In general, I do not share Quine's scepticism toward analyticity, but I admit that his criticism created a new situation for the problem.

[30] Even in his (1951). Of course, later he admitted analytic sentences grounded on behavioral criteria.

[31] In fact, Hintikka prefers the adjective 'conceptual' and 'factual' to 'analytic' and 'synthetic', but I will use the traditional terminology, sometimes adding in brackets Hintikka's labels. Another problem with Hintikka's arguments is that he heavily relies on some on views about logic, information, etc. I will try to put matters in a more general framework which is less committed to particularities.

[32] Some cases are naturally disputable. For example, all mixed explications could be counted as pragmatic. Other doubts can stem from vagueness of certain expressions used in particular definitions, for example 'containment', 'idiomatic operators', 'meaning-tautology',

'terminal sentence', 'maximal paraphrase', 'interlinguistic meaning relations', etc. I do not claim that other words used in defining analyticity are crystallically transparent. In fact, there are troubles with 'semantic rules', 'meaning' or even 'logical constant', but we have at least some standards how they should or could be used. However, regardless of whether my all qualifications are correct or not, it is obvious that many quite different ideas were proposed in order to define analyticity.

[33] In any other system (except propositional logic) the completeness theorem either fails or the set of all models is understood in a special way, for example as in the case of second-order logic where not all models are *al pari.*

[34] 'Pragmatic' is here derived from 'pragmatics', not from 'pragmatism'.

[35] Roughly speaking, I make no difference between theories and conceptual schemes leaving, however, these concepts without further explanations.

[36] I am not sure whether sentences about colours should be classified as analytic, but I leave this problem without further comments at the moment.

[37] This idea, at least in the case of mathematics, was strongly motivated by logicism. However, since it reduced, as it is now clear, reduced mathematics in fact not to logic, but to set theory, this motivation is merely historical.

[38] These distinctions help in analysis of some claims about analyticity. Martin-Löf (see (57) above) says that no existential sentence is analytic. However, it is merely justifiable if the concept analyticity is conceived so narrowly that even not the whole of first-order logic is included into it. Also it can be shown that (46b) must be reduced to decidable theories.

[39] Of course, there are various possible objections to the analysis of analyticity *via* (63)-(67). Two seem particularly important. The first is that I did not define the notion of theory (conceptual scheme), and the second that the concept of analyticity is extended beyond a reasonable limit. Since the analysis of matters related to the first objection exceeds the scope of this paper, I can only repeat what I said in note 33. Let the concept of theory (conceptual scheme) function as a primitive. As far as the matter concerns the second challenge, let me note that also other authors propose a very broad concept of analyticity. For example, Kyburg (1983, 296-311) argues that all acceptable generalizations, including statistical ones, are analytic. Anyway, all arguments against Quine suggest that if we want to keep the concept of analyticity, we should broaden its scope beyond Kantian tradition, particularly for making empiricism consistent with degrees of apriority (see the end of the next section).

[40] I will not give a detailed historical account of the concept of the a priori. Since I outlined the history of analyticity and apriority is largely the same, this strategy is reasonable. Consult Schepers, Tonelli, and Eisler (1971) for a general history and Delius (1963) as far as the matter concerns the 20th century, and works quoted in note 5 above.

[41] There is a problem how to make Kant's two claims coherent: (a) every a priori knowledge, including non-pure one, is absolutely independent of all experience, and (b) non-pure knowledge (or its results) use concepts derived from experience. Leaving this question to Kant's scholars, I will understand the Kantian a priori in a uniform way, that is, without focusing on the difference between its pure and non-pure forms, pure intuition and its scope, etc. I also neglect necessity and strict universality as attributes (or criteria) of a priori sentences (for Kant, 'a priori', 'strictly universal' and 'necessary' are co-extensional). My omissions result with simplifications of Kant's views, but I hope that I do not misinterpret his apriorism.

[42] Note however, that important currents in the philosophy of mathematics, namely intuitionism and formalism, share some Kant's views. It should be also remembered that Russell accepted the synthetic a priori in his early (before 1900) philosophy. Einstein seems to share a view similar to that of Russell and Carnap. Einstein (1923, 189) writes: " As far as the laws of mathematics refer to reality, they are not certain; and as far s they are certain, they

do not refer to reality." Hempel (1945) is a good summary of the related views of logical empiricism.

[43] See De Pieris (1988) about Frege's account of a priori and its relation to that of Kant.

[44] I do not like to say that nativism, at least in the Descartes-Leibniz-Kant sense, is a satisfactory solution of the a priori problem. It is not surprising that Frege and Wittgenstein stay away from this position. However, to paraphrase Kant, apriorism without nativism is empty, and nativism without apriorism is blind. Yet it is characteristic that most contemporary comments about apriorism overlook its relation to nativism.

[45] I follow Ajdukiewicz (1947); see also Carnap (1966, 177-183). This fragment of my paper extends remarks about empiricism, rationalism, aposteriorism and apriorism made in Introduction.

[46] Since radical apriorism is now only a historical peculiarity, I will omit this view in my further discussions. Hence, 'apriorism = moderate apriorism' further on.

[47] Of course, there is a great problem of how experience and rational knowledge should be characterized. Typically, sense experience and introspection (inner perception) are considered as sources of experiential (empirical) knowledge, but various activities of reason are proposed as instances of rational knowledge. Unfortunately, instantiations of the rational given by great historical rationalist, like Plato, Descartes, Leibniz or Kant, vary very much which causes troubles with a clear characterization of reason performance. In fact, not very much beyond this vague description of experience and reason will be given below.

[48] I use here the concept of necessity without an explanation. Informally. '*not-A* is necessary' means (see note 3) '*A* true in all possible worlds' or '*A* is inconsistent'. It must suffice here. See also the discussion of DeLong's example below.

[49] The phrase '... is (are) synthetic a priori' means in this comments that a particular example (or particular examples) was (were) proposed as belonging to the category of the synthetic a priori.

[50] I follow Moser (1987, 7) with some cancellations (I omit what he calls 'necessity-based apriorism' and 'reliabilism' because these seem to be too casual and formulated only for a criticism of a particular view about the a priori), changes (I call later Wittgenstein's view simply 'anthropologism', not 'anthropological conventionalism') and additions (I add (A), (E), (I) and (J)). The list covers earlier discussed views of Husserl, Frege and Wittgenstein. Although I stress the role of nativism for apriorism, I decided to give a more specific list. In fact, Kant was the last traditional rationalistic nativist. Later versions of nativism are not necessarily inconsistent with empiricism. For example, even Chomskian nativism can be biologically interpreted (see also note 61).

[51] The main difficulty of this account is that cognitive superactivities recommended by Husserl and other friends of the material synthetic a priori lead to mutually inconsistent results.

[52] I say "at least two" because every full theory of the a priori should also decide which qualifications mentioned in note 3 are applicable to apriora.

[53]Copi's paper raised a discussion whether it is plausible to use metamathematical theorems in philosophical arguments. This discussion is briefly reported in Wolemski (1993, 128). Let me note that Copi's argument was anticipated by Scholz (see (18) below).

[54] Also Gödel and Kemeny defended, after broadening the concept of analyticity, the analytic status of undecidable sentences (see (19), (26), (29) above)). However, I do not suggest that Gödel shared the linguistic theory of a priori, although Kemeny did.

[55] Castonguay argues very carefully. In fact, my presentation makes the argument much stronger. However, since I am interested in a possible way of supporting the claim that synthetic a priori sentences exist, I decided to strengthen Castonguay's reasoning.

[56] Castonguay's argument receives no support from Gödel's remark (see (19) above)) that arithmetic is, due to results of Church and Turing, non-analytic if analytic sentences are

defined as special cases of the law of identity. In the language of the present paper, Gödel's comment suggests that arithmetic is not syntactically analytic.

[57] The same also concerns, for example, Langford's (1951) proof that the sentence 'A cube has twelve edges' is synthetic a priori and Wolniewicz's (1994) argument that the synthetic a priori is related to semantic relations between language and reality. Compare also Essler (1971) where various possible definitions of the synthetic a priori are discussed. Essler's analysis clearly shows that alleged examples of synthetic a priori sentences are related to relatively narrow accounts of analyticity.

[58] Since I am inclined to axiological non-cognitivism, I omit ethical principles. Commonsensical assertions, like (v), (x), (xi), (xxiii) or (xxxv) may be considered as synthetic a posteriori.

[59] The same was also pointed out by Ajdukiewicz (1934a). Note, however, that this kind of conventionalism is different from that of linguisticism. Ajdukiewicz and Pap, following Poincaré and Lewis, do not consider conventionalization in science as playing with words, but as the fundamental change of the epistemological status of some principles.

[60] It is just only an analogy. I do not suggest any particular resemblances between knowledge a priori (a posteriori) and probability a priori (a posteriori). In particular, I neither touch methodological problems of Bayesian statistics nor say that they have any counterpart in the case of knowledge a priori.

[61] There is an important contrast of how the distinction between the absolute and the relative works for AS and for AP. While in the first case it refers to the substantial difference within the domain of analyticals, it only plays a historical role in the context of AP showing that philosophers propose absolute and relative concepts of the a priori.

[62] In particular, I think that my analysis meets difficulties pointed out by Casullo (1992).

[63] Evolutionary epistemology seems to be the only way to explain apriority of grammar in Chomsky's sense which is accessible to genetic empiricists.

[64] Without entering into details, I claim that all recently proposed accounts of the a priori are interpretable *via* relative apriority. It concerns, for example, Chisholm (1977), Kitcher (1980), Putnam (1978), Holland (1992), Anderson (1993), Peacock (1993), Lewin (1995), Harman (1996), Załaniec (1996), Bonjour (1998) or Bealer (1999); some of these papers are reprinted in Casullo (1999), a good anthology of recent approaches to the a priori. However, one reservation is in order. Several mentioned accounts suggest that we should abandon too narrow concept of experience which was probably very disappointing for Kant. Let me add that some suggestions for the concept of the relative a priori I derived from Essler (1971), Nowaczyk (1979), Friedman (1994) and Field (1995-96). In particular. Field distinguishes weak and strong a priori. See also note 66 on a more limited concept of the relative a priori.

[65] These equivalencies were earlier considered as problematic by Sloman (1965). More specifically, he guesses that (a) not every necessary truth is a priori because unprovable mathematical truths are a necessary and unknowable a priori, (b) not all apriora are analytic because the truth-values of such propositions cannot be established *via* purely logical investigations, and (c) not all necessary propositions are analytic because geometrical truths are dependent not only on logical relations but also on the structure of space. It is interesting that Sloman's view about unprovable mathematical truths is different than that of DeLong (see above) who considers CON(**AR**) to be necessary and a priori. This difference show how intuitions concerning a priori vary even in relation to concrete examples. My answer to (b) was given above, and to (a) and (c) will be outlined below.

[66] Kripke's arguments for the necessary a posteriori and the contingent a priori initiated a considerable discussion which cannot be reported here. A very useful general discussion of the Kripke style examples in the context of AS and AP is contained in Bealer (1999, 243-244). See also the discussion in Sidelle (1989) who argues that conventionalism requires that

there are analytic a posteriori truths. However, as the discussion showed, the matter is highly controversial.

[67] It is the reason that some authors, for example, Geirsson (1999) speak about relative apriority as generated by reference conventions. My understanding of relativity of the a priori is wider. I guess that problems raised by van Fraassen (1977) and Zalta (1988) are also solvable by appealing to relative necessity. Another point overlooked by Kripke concerns various uses of the identity concepts (see Kleene (1967, 157-167), van Fraassen (1977, 79). In particular, we have identity which holds in all possible worlds (Leibniz) and, say, equality related to classes of worlds. Thus, relativity enters everywhere outside pure logic.

REFERENCES

Ajdukiewicz, K.: 1934, 'Sprache und Sinn', *Erkenntnis* **4**, 100-138; Eng. tr. 'Language and Meaning', tr. by J. Wilkinson, in Ajdukiewicz, 1978, pp. 35-66.

Ajdukiewicz, K.: 1934a, 'Das Weitbild und die Begriffsapparatus', *Erkenntnis* **4**, 259-287; Eng. tr. 'The World-Picture and Conceptual Apparatus', tr. by J. Wilkinson, in Ajdukiewicz, 1978, pp. 67-89.

Ajdukiewicz, K.: 1947, 'Logika i doświadczenie', *Przeglad Filozoficzny* **43**, 3-22; Eng. tr. 'Logic and Experience', tr. by J. Giedymin, in Ajdukiewicz, 1978, pp. 165-181.

Ajdukiewicz, K.: 1958, 'Le problème du fondement des propositions analytiques', *Studia Logica* **VIII**, 259-272; Eng. tr. 'The Problem of the Foundation of Analytic Sentences', tr. by J. Giedymin, in Ajdukiewicz, 1978, pp. 254-268.

Ajdukiewicz, K.: 1978, *The Scientific World-Perspective and Other Essays 1931-1963,* ed. by J. Giedymin, D. Reidel Publishing Company, Dordrecht.

Anderson, C. A.: 1991, 'Toward a Logic of A Priori Knowledge', *Philosophical Topics* **21**, 1-20.

Aune, B.: 1972, 'On an Analytic-Synthetic Distinction', *American Philosophical Quarterly* **9**, 235-242.

Ayer, A. J.: 1946, *Language, Truth, and Logic,* sec. ed., Victor Gollancz, London.

Bar-Hillel, Y.: 1950, 'Bolzano's Definition of Analytic Proposition', *Theoria,* vol. **16**, 91-117; repr. in Y. Bar-Hillel, *Aspects of Language,* Magnus Press, Jerusalem, 1970, pp. 3-24.

Bealer, G.: 1999, 'The A Priori', in *The Blackwell Guide to Epistemology,* ed. by J. Greco and E. Sosa, Blackwell Publishers, Oxford, 243-270.

Bencivenga, E.: 1986, 'Analyticity and Analytical Truth', *Notre Dame Journal of Formal Logic* **27**, 15-19.

Bennett, J.: 1959, Analytic-Synthetic', *Proceedings of Aristotelian Society* **23**, 164-188.

Berg, J.: 1962, *Bolzano's Logic,* Almqvist & Wicksell, Stockholm.

Berg, J.: 1973, 'Editor's Introduction', in B. Bolzano 1973, pp. 1-30.

Berg, J.: 1992, *Ontology Without Ultrafilters and Possible Worlds An Examination of Bolzano's Ontology,* Academia Verlag, Sankt Augustin.

Bergmann, G.: 1958, 'Analyticity', *Theoria* **XXI,** 71-93; repr. in G. Bergmann, *Meaning and Existence,* The University of Wisconsin Press, Madison, 73-90.

Bogdan, R. (ed.): 1987, *Jaakko Hintikka,* D. Reidel, Dordrecht.

Boghossian, P. A.: 1997, 'Analyticity', in *A Companion to the Philosophy of Language,* ed. by B. Hale and C. Wright, Blackwell, Oxford, pp. 331-368.

Bohnert, H. B.: 1963, 'Carnap's Theory of Definition and Analyticity', in Schilpp, 1963, pp. 407-430.

Bohnert, H. B.: 1986, 'Quine on Analyticity', in Schilpp, 1986, pp. 77-92.

Bolzano, B.: 1839, *Wissenschaftslehre,* Sulzbach; partial Eng. tr.. *Theory of Science,* tr. by B. Terrell, D. Reidel Publishing Company, Dordrecht, 1973.

Bonjour, L.: 1998, *In Defense of Pure Reason A Rationalistic Account of A Priori Justification,* Cambridge University Press, Cambridge.

Borkowski, L.: 1966, 'Deductive Foundation and Analytic Propositions', *Studia Logica* **XIX**, 59-72.

Broad, C. D.: 1936, Are there Synthetic A Priori Truths?', *The Aristotelian Society, Supplementary Volume* **15**, 21-38.

Butrick, R.: 1970, *Carnap on Meaning and Analyticity,* Mouton, The Hague.

Carnap, R.: 1934: *Logische Syntax der Sprache,* Julius Springer, Wien; Eng. tr. *Logical Syntax of Language,* tr. by A. Smeaton, Kegan Paul, Trench, Teubner & Co., London, 1937.

Carnap, R.: 1942, *Introduction to Semantics,* Harvard University Press, Cambridge, Mass.

Carnap, R.: 1947, *Meaning and Necessity,* The University of Chicago Press, Chicago.

Carnap, R.: 1950, *Logical Foundations of Probability,* The University of Chicago Press, Chicago.

Carnap, R.: 1955, 'Notes on Semantics', *Philosophia* **2**, 1972, 3-54.

Carnap, R.: 1958, 'Beobachtungsprache und theoretische Sprache', *Dialectica* **47/ 48**, 236-248.

Carnap, R.: 1963, 'W. V. Quine on Logical Truth', in Schilpp, 1963, pp. 915-922.

Carnap, R.: 1963a, 'Herbert H. Bohnert on Definitions and Analyticity', in Schilpp, 1963, pp. 22-23.

Carnap, R.: 1963b, 'Carl G. Hempel on Scientific Theories', in Schilpp, 1963, pp. 958-966.

Carnap, R.: 1966, *Philosophical Foundations of Physics,* Basic Books, New York.

Carnap, R. and Y. Bar-Hillel: 1953, *An Outline of a Theory of Semantic Information,* Technical Report, No. **247**, Massachusetts Institute of Technology, Cambridge, Mass.; repr. in Y. Bar-Hillel, *Language and Information Selected Essays on Their Theory and Applications,* Addison-Wesley Publishing Company, Reading, Mass, 1964, pp. 221-274.

Carnap, R.: 1989, 'Quine on Analyticity', in Creath, 1990, pp. 427-432.

Cassullo, A.: 1992, 'Analyticity and the Apriori', in Ph. Hanson and B. Hunter (eds.), *Return of the A Priori, Canadian Journal of Philosophy Supplementary Volume* **18**, 113-150.

Cassullo, A. (ed.): 1999, *A Priori Knowledge,* Ashgate, Alershot.

Castonguay, Ch.: 1976, 'Church's Theorem and the Analytic/Synthetic Distinction in Mathematics', *Philosophica* **18(2)**, 77-89.

Chisholm, R.: 1977, *Theory of Knowledge,* sec. ed., Prentice-Hall, Engle-wood Cliffs, New Jersey.

Cirea, R.: 1994, *Carnap and the Vienna Circle Empiricism and Logical Syntax,* Rodopi, Amsterdam.

Coffa, A. J.: 1991, *The Semantic Tradition from Kant to Carnap To the Vienna Station,* Cambridge University Press, Cambridge.

Copi, I.: 1949, 'Modern Logic and the Synthetic A Priori', *The Journal of Philosophy* **46**, 243-245.

Creath, R.: 1987, 'The Initial Reception of Carnap's Doctrine of Analyticity', *Nous* **21**, 471-475.

Creath, R. (ed.): 1990, *Dear Carnap, Dear Van: The Quine-Carnap Correspondence and Related Work W. V. Quine and Rudolf Carnap,* Open Court, La Salle.

Creath, R.: 1990a, 'Introduction', in Creath (1990), 1-43.

Creath, R.: 1992, 'Carnap's Conventionalism', *Synthese* **93**, 141-166.

Dancy, J. and E. Sosa (eds.): 1993, *A Companion to Epistemology,* Blackwell, Oxford.

David, M.: 1997, 'Two Conceptions of the Synthetic A Priori', in Hahn, 1997, pp. 629-651.

De Pieris, G.: 1988, 'Frege and Kant on A Priori Knowledge', *Synthese* **77**, 285-319.

DeLong, H.: 1970, *A Profile of Mathematical Logic,* Addison-Wesley Reading, Mass.

Delius, H.: 1963, *Untersuchungen zur Problematik der sogenannten syntethischen Sätze a priori,* Vandenhoeck & Ruprecht, Gottingen.

Detlefsen, M., D. C. McCarty, and J. B. Bacon: 1999, *Logic from A to* Z, Routledge, London.

Dreben, B. and J. Floyd: 1991, 'Tautology: How not to Use a Word', *Synthese* **87**, 23-49.

Dubislav, W.: 1926, *Über die sog. analytischen und synthetischen Urteilen,* Hermann Weiss-Verlag, Berlin.

Dummett, M.: 1991, *Frege: Philosophy of Mathematics,* Duckworth, London.

Einstein, A.: 1923, *Geometrie und Erfahrung,* Springer, Berlin; Eng. tr. in H. Feigl and M. Brodbeck (eds.), *Reading in the Philosophy of Science,* Appleton-Century-Crofts, Inc., New York, 1953, pp. 189-194.

Essler, R.: 1971, 'Über syntetisch-apriorische Urteile', in H. Lenk (ed.), *Neue Aspekte der Wissenschaftstheorie,* Vieweg, Braunschweig, 195-204.

Field, H.: 1995-96, 'The A Prioricity of Logic", *Proceedings of Aristotelian Society* **96**, 359-379.

Frege, G.: 1884, *Die Grundlagen der Arithmetik,* Breslau; Eng. tr. *The Foundations of Mathematics,* tr. by J. L. Austin, Blackwell, Oxford 1950.

Frege, G.: 1885, 'Über formale theorien der Arithmetic', *Sitzungberichte der Jenaischen Gesellschaft für Medizin und Naturwischenschaften* 12, 94-104; Eng. tr. 'On Formal Theories of Arithmetic', in G. Frege, *Collected Papers on Mathematics, Logic, and Philosophy,* B.Guiness (ed.), Blackwell Oxford 1984, pp. 112-121.

Friedman, M.: 1988, 'Logical Truth and Analyticity in Carnap's 'Logical Syntax of Language'', in W. Aspray and Ph. Kitcher (eds.), *History and Philosophy of Mathematics* (Minnesota Studies in the Philosophy of Science, v. **XI**), Minnesota University Press, Minneapolis, pp. 82-94.

Friedman, M.: 1994, 'Geometry, Convention, and the Relativized A Priori', in W. Salmon and G. Wolters (ed.), *Logic, Language and the Structure of Scientific Theories,* University of Pittsburgh Press and Universitats Verlag Konstanz, Pittsburgh and Konstanz, pp. 20-48.

Geirsson, H.: 1999, Justification and Relative Apriority', *Ratio* **12**, 148-161.

Gewirth, A.: 1953, 'The Distinction between Analytic and Synthetic Truths', *The Journal of Philosophy* L, 397-425.

Gochet, P.: 1986, *Ascent to Truth A Critical Examination of Quines Philosophy,* Philosophia Verlag, München.

Goodman, N.: 1949, 'On Likeness of Meaning', *Analysis* **10**, 1-7; repr. (with revisions) in Linsky, 1952, pp. 65-74.

Gödel, K.: 1944, 'Russell's Mathematical Logic', in P. Schilpp (ed.), *The Philosophy of Bertrand Russell,* Open Court Publishing Company, La Salle, pp. 123-153; repr. in Gödel, 1990, pp. 119-141.

Gödel, K.: 1946, 'Remarks before the Princeton Bicentennial Conference on Problems of Mathematics', in Gödel, 1990, pp. 150-153.

Gödel, K.: 1951, 'Some Basic Theorems on the Foundations of Mathematics and Their Implications', in K. Gödel, *Collected Works, v. III, Unpublished Essays and Lectures,* S. Feferman et al. (eds.), Oxford University Press, Oxford, 1995, pp. 304-323.

Gödel, K.: 1990, *Collected Works, v. II, Publications 1938-1974,* S. Feferman and al. (eds.), Oxford University Press, Oxford, 1990.

Grice, H. and P. Strawson: 1956, 'In Defence of a Dogma', *The Philosophical Review* **65**, 141-148; repr. in H. Grice, *Studies in the Way of Words,* Harvard University Press, Harvard, pp. 196-212.

Gupta, A.: 1982, Analytic and Synthetic Propositions', *Archiv für Geschichte der Philosophie* **64**, 56-63.

Haack, S.: 1977, Analyticity and Logical Truth', in *The Roots of Reference, Theoria* **XLIII**, 129-143.

Hahn, L. E. (ed.): 1997, *The Philosophy of Roderick M. Chisholm,* Open Court Publishing Company, La Salle.

Hardin, C. L.: 1988, *Colours for Philosophers, Unweaving the Rainbow,* Hackett, Indianapolis.

Harman, G.: 1996, Analyticity Regained?', *Nous* **30**, 392-400.

Hellman, G.: 1986, 'Logical Truth by Linguistic Convention', in Schilpp (1963), 189-205.

Hempel, C. G.: 1945, 'Geometry and Empirical Science', *American Mathematical Monthly* **52**, 410-421; repr. in H. Feigl and W. Sellars (eds.) *Reading in Philosophical Analysis,* Appleton-Century-Crofts, New York, 1949, pp. 238-249.

Hempel, C. G.: 1963, 'Implications of Carnap's Works for the Philosophy of Science', in Schilpp, 1963, pp. 685-709.

Hintikka, J.: 1973, *Logic, Language Games and Information Kantian Themes in the Philosophy of Logic,* At Clarendon Press, Oxford.

Hintikka, J.: 1987, 'Rantala and Tselishchev on Surface Information and Analyticity', in Bogdan, 1987, pp. 280-284.

Hintikka, J.: 1999, 'A Distinction Too Few or Too Many? A Vindication of Analytic vs. Synthetic Distinction' (unpublished).

Holland, R.: 1992, 'Apriority and Applied Mathematics', *Synthese* **92**, 349-370.

Kant, I.: 1781, *Kritik der reinen Vernunft,* Riga; Eng. tr.. *Critique of Pure Reason,* tr. by . N. Kemp Smith, Macmillan, 1929.

Kant, I.: 1783, *Prolegomena zu einer jeden künftigen Metaphysik die als Wissenschaft wird auftreten können,* Riga; Eng. tr. *Prolegomena to Any Future Metaphysics,* tr. by L. W. Beck, Bobbs-Merrill Company, Indianapolis, 1950.

Kant, I.: 1800, *Logik,* B. Jäschke (ed.), Königsberg; Eng. tr. *Logic,* tr. by R. Hartman and W. Schwarz, Bobbs-Merrill Company, Indianapolis, 1974.

Kaplan, D., 1979: 'On the Logic of Demonstratives', *Midwest Studies in Philosophy* **4**, 401-414.

Katz, J.: 1972, *Semantic Theory,* Harper & Row Publishers, New York.

Katz, J.: 1992, 'Analyticity', in Dancy and Sosa, 1992, 11-17.

Kemeny, J.: 1952, 'Review of Quine (1952)', *The Journal of Symbolic Logic* **17**, 282-283.

Kemeny, J.: 1952a, 'Review of Mates (1951), *The Journal of Symbolic Logic* **17**, 283.

Kemeny, J.: 1956, 'A New Approach to Semantics', *The Journal of Symbolic Logic* **21**, 1-27.

Kemeny, J.: 1964, 'Analyticity versus Fuziness', in *Form and Strategy in Science,* J. G. Gregg and F. T. C. Harris (eds.), D. Reidel Publishing Company, Dordrecht, pp. 122-145.

Kitcher, Ph.: 1980, A Priori Knowledge', *Philosophical Review* **89**, 3-23.

Kleene, S. C.: 1939, 'On the Term 'Analytic' in Logical Syntax', *The Journal of Unified Science* **IX**, 189-192.

Kleene, S. C.: 1967, *Mathematical Logic,* Wiley, New York.

Kokoszyńska, M.: 1947, 'O różnych ródzajach zdań' (On Various Kinds of Sentences), *Przegląd Filozoficzny* XLIII, 22-51.

Kripke, S.: 1971, 'Identity and Necessity', in M. Munitz (ed.), *Identity and Individuation,* New York University Press, New York, pp. 135-164.

Kripke, S.: 1972, 'Naming and Necessity', in G. Harman and D. Davidson (eds.), *Semantics of Natural Language,* D. Reidel, Dordrecht, pp. 253-355; sec. ed., Blackwell, Oxford, 1980.

Kyburg, H. E.: 1983, *Epistemology and Inference,* University of Minnesota Press, Minneapolis.

Langford, C.: 1949, A Proof that Synthetic *A Priori* Propositions Exist', *The Journal of Philosophy* **XLVI,** 20-24.

Lemos, N. M.: 1997, 'Chisholm, the A Priori, and Epistemic Principles', in Hahn, 1997, pp. 609-628.

Lewin, Y., 1995, Synthetic Apriority', *Erkenntnis* **43**, 137-150.

Lewis, C. I.: 1946, *An Analysis of Knowledge and Valuation,* Open Court Publishing Company, La Salle.

Linsky, L. (ed.): 1952 *Semantics and the Philosophy of Language,* The University of Illinois Press, Urbana.

Marc-Wogau, K.: 1951, 'Kants Lehre von analytischen Urteil', *Theoria,* **XVII**, 99-110; Eng. tr. Kant's Doctrine of the Analytical Judgment', in K. Marc-Wogau, *Philosophical Essays History of Philosophy, Perception, Historical Explanation,* CWK Gleerup, Lund, 1967, pp. 99-110.

Marion, M.: 1998, *Wittgenstein, Finitism, and the Foundations of Mathematics,* Clarendon Press, Oxford.

Martin, J. N.: 1959, *The Notion of Analytic Truth,* University of Pennsylvania Press, Philadelphia.

Martin, J. N.: 1987, *Elements of Formal Semantics An Introduction to Logic for Students of Language,* Academic Press, New York.

Martin, R. M.: 1951, 'On Analytic', *Philosophical Studies* **3**, 42-47.

Martin-Löf, P.: 1994, Analytic and Synthetic Judgments in Type Theory', in P. Parnini, *Kant and Contemporary Epistemology,* Kluwer Academic Publishers, Dordrecht, pp. 87-99.

Mates, B.: 1951, Analytic Sentences', *The Philosophical Review* **60**, 524-534.

Mehlberg, H.: 1958, *The Reach of Science,* University of Toronto Press, Toronto.

Milmed, B.. K.: 1961, *Kant and Current Philosophical Issues,* New York University Press, New York.

Moravscik, J.: 1965, 'The Analytic and the Nonempirical', *The Journal of Philosophy* **42**, 415-429.

Morscher, E.: 1973, *Das logische An-sich bei Bernard Bolzano,* Verlag Anton Pustet, Salzburg.

Moser, P.: 1987, 'Introduction', in P. Moser (ed.) *A Priori Knowledge,* Oxford University Press, Oxford, pp. 1-14.

Moser, P.: 1992, Analyticity and Epistemology', *Dialectica* **46**, 3-19.

Miller, A.: 1998, *Philosophy of Language,* UCL Press, London.

Müller, O.: 1998, *Synonymie und Analytizität Zwei sinvolle Begriffe,* Ferdinand Schoningh, Paderborn.

Nordenstam, T: 1972, *Empiricism and the Analytic-Synthetic Distinction,* Universitetsforlaget, Oslo.

Nowaczyk, A.: 1977, Analytic Sentences in the Semantic System', in Przełęcki and Wöjcicki (1977a), 457-497.

Nowaczyk, A.: 1979, Analyticity and Apriority', in *Semiotic in Poland 1894-1969,* D. Reidel, Dordrecht, 465-483.

Oberdan, T: 1992, 'The Concept of Truth in Carnap's *Logical Syntax of Language*', *Synthese* **93**, 239-260.

Oberdan, T: 1993, *Protocols, Truth and Convention,* Rodopi, Amsterdam .

Orenstein, A.: 1977, *Willard Van Orman Quine,* Twayne Publishers, Boston.

Pap, A.: 1946, *The A Priori in Physical Theory,* King's Crown Press, New York.

Pap, A.: 1955, Necessary Propositions and Linguistic Rules', *Archivio di Filosofia* **3**, 63-105.

Pap, A.: 1958, *Semantics and Necessary Truth,* Yale University Press, New Haven.

Parsons, Ch.: 1995, 'Quine and Gödel on Analyticity', in *On Quine New Essays,* ed. by P. Leonard and M. Santabrogio, Cambridge University Press, Cambridge, 297-313.

Pasch, A.: 1958, *Experience and the Analytic A Reconsideration of Empiricism.* The University of Chicago Press, Chicago.

Peacock, A.: 1993, 'How are A Priori Truths Possible?, *European Journal of Philosophy* **1**, 175-199.

Perkins, R. and Singer, I.: 1951, Analyticity', *The Journal of Philosophy* **XLVIII**, 485-497.

Pollock, J.: 1965, 'Implication and Analyticity', *The Journal of Philosophy* **LXII**, 150-157.

Priest, G.: 1979, 'Two Dogmas of Quineanism', *Philosophical Quarterly* **29**, 289-301.

Proust, J.: 1989, *Question of Form Logic and the Analytic Proposition from Kant to Carnap,* University of Minnesota Press, Minneapolis.

Przełęcki, M. and Wójcicki, R.: 1977, 'The Problem of Analyticity', in Przełęcki and Wójcicki, 1977a, pp. 589-614.

Przełęcki, M. and Wójcicki, R. (eds.): 1977a, *Twenty-Five Years of Logical Methodology in Poland,* D. Reidel Publishing Company, Dordrecht.

Putnam, H.: 1962, 'The Analytic and the Synthetic', in H. Feigl and G. Maxwell (ed.), *Scientific Explanation, Space, and Time* (Minnesota Studies in the Philosophy of Science, v. III), University of Minnesota Press, Minneapolis, pp. 358-397; repr. in H. Putnam, *Mind, Language and Reality Philosophical Papers,* v. 2, Cambridge University Press, Cambridge, 1975, pp. 33-69.

Putnam, H.: 1978, 'There is at least One A Priori Truth', *Erkenntnis* **13**, 153-170.

Quine, W. V.: 1936, 'Truth by Convention', in O. H. Lee (ed.), *Philosophical Essays for A, N. Whitehead,* Longmans, New York, 90-124; repr. in W. V. Quine, *The Ways of Paradox and Other Essays*, Random House 1966, pp. 70-99.

Quine, W. V.: 1951, 'Two Dogmas of Empiricism', *The Philosophical Review* 60, 20-43; repr. in W. V. Quine, *From a Logical Point of View,* Harvard University Press, Cambridge, Mass., pp. 20-43.

Quine, W. V.: 1962, *Word and Object,* The MIT Press, Cambridge, Mass.

Quine, W. V.: 1963, 'Carnap and Logical Truth', in Schilpp, 1963, pp. 385-406.

Quine, W. V.: 1974, *The Roots of Reference,* Open Court Publishing Company, La Salle.

Quine, W. V.: 1986, 'Reply to Herbert G. Bohnert', in Schilpp, 1986, pp. 93-95.

Quine, W. V.: 1986a, 'Reply to Geoffrey Hellman', in Schilpp, 1986, pp. 206-208.

Quine, W. V.: 1990, 'Lectures on Carnap', in Creath, 1990, pp. 47-103.

Quine, W. V.: 1991, 'Two Dogmas Revisited', *Canadian Journal of Philosophy* **21**, 265-274.

Quine, W. V.: 1995, *From Stimulus to Science,* Harvard University Press, Cambridge, Mass.

Rantala, V. and V. Tselishchev: 1987, 'Surface Information and Analyticity', in Bogdan, 1987, pp. 77-90.

Reichenbach, H.: 1920, *Relativitätstheorie und Erkenntnis Apriori,* Springer, Berlin; Eng. tr. *The Theory of relativity and A Priori Knowledge,* tr. by M. Reichenbach, University of California Press, Los Angeles.

Reichenbach, H.: 1951, *The Rise of Scientific Philosophy,* University of California Press, Los Angeles.

Russell, B.: 1919, *Introduction to Mathematical Philosophy,* Alien and Unwin, London.

Russell, B.: 1948, *Human Knowledge Its Scope and Limits,* George Alien and Unwin, London.

Sarkar, S., 1992: '"The Boundless Ocean of Unlimited Possibilities": Logic in Carnap's *Logical Syntax of Language', Synthese* **93**, 191-237.

Sauer, W.: 1986, 'Über das Analytische und das Syntethische Apriori bei Chisholm', *Grazer Philosophische Studien* **28**, 57-77.

Schilpp, P. (ed.): 1963, *The Philosophy of Rudolf Carnap,* Open Court Publishing Company, La Salle.

Schilpp, P. (ed.): 1986, *The Philosophy of W.V. Quine,* Open Court Publishing Company, La Salle.

Schepers, H., G. Tonelli, and R. Eisler: 1971, A priori / a posteriori', in R. Eisler and al. (eds.), *Historisches Wörterbuch der Philosophie,* v. I, Benno Schwabe, Basel, pp. 462-474.

Schlick, M.: 1918, *Allgemeine Erkenntnislehre,* Julius Springer, Berlin; Eng. tr.. *General Theory of Knowledge,* tr. by A. E. Blumberg, Springer-Verlag, Wien, 1974.

Scholz, H.: 1944, 'Einführung in die Kantische Philosophie' (lecture notes), in H. Scholz, *Mathesis Universalis Abhandlungen zur Philosophie als strenge Wissenschaft,* Benno Schwabe, Basel 1961, pp. 152-218.

Sidelle, A.: 1989, *Necessity, Essence and Individuation A Defence of Conventionalism,* Cornell University Press, Ithaca.

Siebel, M.: 1996, *Der Begriff der Ableitbarkeit bei Bolzano,* Academia Verlag, Sankt Augustin.

Sloman, A.: 1965, ''Necessary', 'A Priori' and 'Analytic'', *Analysis* **26**, 1961, 12-16.

Stegmüller, W.: 1957, *Das Wahrheitsproblem und die Idee der Semantik Eine Einfuhrung in die Theorien of A. Tarski und R. Carnap,* Springer-Verlag, Wien.

Stein, E.: 1993, 'Evolutionary Epistemology', in J. Dancy and E. Sosa, 1993, pp. 122-125.

Stenius, E.: 1965, 'Are True Numerical Statements Analytic or Synthetic?', *The Philosophical Review* **74**, 357-372; repr. with additions in E. Stenius, *Critical Essays,* North-Holland Publishing Company, 1972, pp. 68-84.

Stenius, E.: 1972, 'The Concepts "Analytic" and "Synthetic", in R. E. Olson and A. M. Paul (eds.), *Contemporary Philosophy in Scandinavia,* The Johns Hopkins Press, Baltimore, pp. 57-76.

Stenius, E.: 1984, 'Ad van Benthem', *Theoria* **L**, 267-272.

Suszko, R.: 1968, 'Formal Logic and the Development of Knowledge', in I. Lakatos and A. Musgrave (eds.), *Problems in the Philosophy of Science,* North-Holland Publishing Company, Amsterdam, pp. 210-222.

Tappenden, J.: 1993, 'Analytic Truth – It's Worse (or Perhaps Better) than You Thought', *Philosophical Topics* **21.2**, 233-261.

Tarski, A.: 1936, 'Über den Begriff den logischen Folgerung', *Actes du Congrès International de Philosophie Scientifique,* v. 7, Hermann, Paris; Eng. tr. 'On the Concept of Logical Consequence', in A. Tarski, *Logic, Semantics, Metamathematics,* tr. by J. H. Woodger, At Clarendon Press, Oxford 1956, pp. 408-420.

Tarski, A.: 1987, A Letter of Alfred Tarski [to M. White, September 23, 1944], *The Journal of Philosophy* **84**, 29-32.

Tennant, N.: 1991, 'Carnap, Gödel and thew Analyticity of Arithmetic' (unpublished manuscript).

Textor, M.: 1996, *Bolzanos Propositionalismus,* Walter de Gruyter, Berlin.

Ueberweg, F: 1857, *System der Logik und Geschichte der logischen Lehren,* Berlin; Eng. tr. *System of Logic and History of Logic,* tr. by T. M. Lindsay, London 1871.

van Benthem, J.: 1974, 'Hintikka on Analyticity', *Journal of Philosophical Logic* **3**, 419-431.

van Benthem, J.: 1984, 'Analytic/Synthetic: Sharpening a Philosophical Tool', *Theoria* L, 106-137.

van Fraassen, B. C., 1977: 'The Only necessity is Verbal Necessity', *The Journal of Philosophy* **74**, 71-85.

Waissman, R: 1949-1953, 'Analytic-Synthetic', Analysis **10**(1949), 25-40, **11**(1950), 25-38, **11**(1951), 49-61, **11**(1951), 115-124, **13** (1952), 1-14, **13**(1953), 73-89; repr. in F. Waismann, *How I See Philosophy,* Macmillan, London 1968, pp. 122-207.

Wang, H.: 1955, 'Notes on the Analytic-Synthetic Distinction', *Theoria* **XXI,** 158-178.

Wang, H.: 1974, *From Mathematics to Philosophy,* Routledge & Kegan Paul, London.

White, M.: 1950, 'The Analytic and the Synthetic: An Untenable Dualism', in S. Hook (ed.), *John Dewey: Philosopher of Science and Freedom,* The Dial Press, New York, pp. 316-330; repr. in Linsky, 1952, pp. 271-286.

White, M.: 1956, *Toward Reunion in Philosophy,* Harvard University Press, Cambridge, Mass.

Wittgenstein, L.: 1922, *Tractatus Logico-Philosophicus,* with an Introduction by B. Russell, tr. by C. K. Ogden, Kegan Paul, Trench, Trubner & Co., London.

Wittgenstein, L.: 1961, *Notebooks 1914-1916,* tr. by G. E. M. Anscombe, Blackwell, Oxford.

Woleński, J.: 1993, Analyticity and Metamathematics', *Folia Philosophica* **9**, 125-131.

Woleński, J.: 1993a, Analyticity, Decidability and Incompleteness', in *Philosophy of Mathematics.* Part 1, ed. by J. Czermak, Hölder-Pichler-Tempsky, Wien, pp. 379-382.

Wolniewicz, B.: 1994, 'On the Synthetic *A Priori'* in J. Woleński (ed.), *Philosophical Logic in Poland*, Kluwer Academic Publishers, Dordrecht, pp. 327-335.

Zalta, E., 1988: 'Logical and Analytic Truths that are not Necessary', *The Journal of Philosophy* **85**, 57-74.

Żełaniec, W.: 1996, *The Recalcitrant Synthetic A Priori,* Artom, Lublin.

FREDERICK F. SCHMITT

EPISTEMOLOGY AND COGNITIVE SCIENCE

1 INTRODUCTION

I will define epistemology in the traditional way, as the *conceptual* and *normative* study of knowledge. Epistemology inquires into the definition, criteria, normative standards, and sources of knowledge and of kindred statuses like justified belief, evidence, confirmation, rational belief, perceiving, remembering, and intelligence. Cognitive science is, by contrast, the interdisciplinary *empirical* study of cognition in human beings, animals, and machines, and the attempt to *engineer* intelligent cognition.[1] Cognitive science spans work in diverse fields, including empirical cognitive psychology, linguistics, artificial intelligence (AI), neuroscience, and cognitive anthropology.[2] Both epistemology and cognitive science study knowledge, but they have different aims, interests, and methods.

The idea that epistemology and psychology have something helpful to say to one another was common currency in eighteenth and nineteenth century thought (notably, in the work of Hume and Mill). But by 1900, the idea had fallen into disrepute and (with a few distinguished exceptions–Dewey, most notably) remained so until the 1960s. There are still many respected epistemologists who in one way or another think it important to resist the significance of empirical science for epistemology (Siegel 1981, Stroud 1984, 1989, Kim 1988, Feldman 1989, Shatz 1993). The key source of resistance here is that epistemology is conceived as a conceptual and normative study, and to many it has seemed obvious that these studies must be purely a priori and that empirical science cannot be relevant to an a priori study. Psychology, for its part, spent much of this century resolutely avoiding the study of cognitive states and processes, and so discovered little that could have contributed to an inquiry into the nature of knowledge, had epistemology been receptive to its findings. It was not until the advent of information-processing psychology that psychologists began employing concepts of interest to epistemologists (like the concept of information) and targeting states of knowledge that were common ground with epistemology. At this point, there emerged a potential for interaction between the fields that has yet to be fulfilled.

My task in this article is to ask how the findings of cognitive science might be relevant to epistemology traditionally conceived. I will argue that epistemology may profit from the findings of cognitive science in diverse ways. Of course, "epistemology" might be *defined* in a broad enough way to overlap or even encompass cognitive science, but the question of the relevance of cognitive science to epistemology has interest only if epistemology is conceived traditionally as a conceptual and normative enterprise. Again, "epistemology" might be defined in such a way that it is a *pure* a priori study, to which empirical science could therefore

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 841-918.
© 2004 *Kluwer Academic Publishers.*

not be relevant. The interesting question, however, is whether cognitive science could be relevant to epistemology understood as *at least* a conceptual and normative study. Can empirical science aid in such an enterprise? Must we advert to the findings of cognitive science in such an enterprise?

I will assume throughout that cognitive science theorizes at three levels (Pylyshyn 1984, Dennett 1987, Von Eckhardt 1993 for diverse tripartite divisions). First, there is the *intentional* level. Here psychology employs the idiom of folk psychology (talk of intentional states like belief) and charts intentionally describable causal relations (inferential relations, reasoning processes) between intentional states (Dennett's "intentional stance").[3] This is the level at which the findings of cognitive science–such findings as belief perseverance and failure to conform to textbook logic, probability calculus, or statistics–are most immediately relevant to epistemology, couched as it is in the intentional idiom. Second, there is the *cognitive* level, at which intentional generalizations are explained by generalizations about cognitive states, processes, faculties, and architecture (Dennett's "design stance"). I will suppose, for convenience, that cognitive states realize intentional states, that cognitive processes relating cognitive states explain the causal relations between the intentional states they realize, and finally, that cognitive processes are computational.[4] But I doubt whether the bearing of cognitive science on epistemology hangs on any of these controversial assumptions.[5] The cognitive level is more difficult to bring to bear on epistemology than is the intentional level, if only because the relation between cognitive states and intentional states is hotly contested. Nevertheless, it is possible to make fruitful speculative, hypothetical, and heuristic use of work at the cognitive level. Finally, there is the *neural* level, which implements the cognitive level. Though I believe neuroscience will eventually prove highly significant for epistemology by constraining the cognitive and intentional levels, the nature of neural implementation is largely unknown, and it is currently very difficult to make epistemological capital of neuroscientific findings. I will therefore omit discussion of the bearing of neuroscience here.[6]

To forestall disappointment, let me warn that, for reasons of space, I will not attempt to discuss efforts by cognitive scientists that clearly belong to the province of traditional epistemology (e.g., the frame problem). Conversely, I will skip clearly empirical issues that have customarily received treatment from epistemologists (e.g., the issue of innate knowledge of mathematics, linguistics, etc.).[7]

2 SKEPTICISM AND COGNITIVE SCIENCE

Until the late nineteenth century, the history of epistemology was in large measure the history of responses to skepticism (see Hookway 1990 and Schmitt 1992 for surveys of the history of skepticism). It is natural, then, to begin our discussion with the question whether cognitive science bears on skepticism. Do empirical findings from cognitive science aid in answering Cartesian or other traditional forms of skepticism?

2.1 Quine on Skepticism and Naturalized Epistemology

Curiously, the idea that empirical psychology bears on skepticism emerged in recent epistemology in the thinking of a philosopher fond of behaviorism and suspicious of intentional states. In his celebrated article, "Epistemology Naturalized" (1969a), W. V. Quine urges that a species of empirical developmental psychology is the proper heir to the traditional business of answering Cartesian skepticism. This psychology, however, properly *replaces* the project of answering Cartesian skepticism. In brief, the Cartesian skeptical challenge arises in this way (see Descartes 1984 *Meditations* I-III). We have knowledge of the external world only if we are able to rule out prima facie possibilities consistent with our experiences but contrary to our beliefs about bodies–most importantly, the "demon hypothesis," the contrary possibility that our putative experiences of bodies are caused, not by bodies, but by a powerful demon. Yet the demon hypothesis calls all of our beliefs about the external world into question at once. For this reason, we cannot rely on any of our beliefs about the external world in ruling out the demon hypothesis. But we have no other way to rule out the demon hypothesis. Since we cannot rule out all prima facie possibilities contrary to our beliefs about the external world, it follows that we have no knowledge of the external world. (Descartes himself rejects the claim that we have no other way to rule out the demon hypothesis. We can do so, according to Descartes, by arguing that certain propositions must be true whether or not there is a demon, and from these propositions we can deduce the reliability of clear and distinct perception–hence, rule out that we are deceived by a demon.)

In "Epistemology Naturalized," Quine agrees with the skeptic that we have no way to rule out the demon hypothesis, but he regards our inability to rule out the demon hypothesis as a reason to decline the task of answering the skeptical challenge. He does not explicitly reject the Cartesian assumption that knowledge requires being able to rule out contrary possibilities for all beliefs about the external world at once, though he apparently wishes to maintain that we have knowledge despite our inability to rule out the demon hypothesis. But Quine's primary message is that we ought to turn from the task of answering skepticism to an alternative project, albeit an analogous one. The proposed project is to describe and explain how our sensory stimuli give rise, in the course of human cognitive development, to a theory of the external world: a "human subject is accorded a certain experimentally controlled input–certain patterns of irradiation in assorted frequencies, for instance–and in the fullness of time the subject delivers as output a description of the three-dimensional external world and its history" (1969a, 82-3). There is a gap between our "meager sensory input" and our "torrential theoretical output," and it is the new business of epistemology to explain how, given this gap, theory develops from stimuli. The task is not to show how we can justify theory on the basis of the stimuli without relying on any assumptions about the external world, as the Cartesian skeptic requires us to do; it is rather to *describe* and *explain* cognitive development utilizing the full power of empirical science. Reliance on science would be circular if the project were to justify science; but it is a descriptive-explanatory project, a chapter of psychology (1969a, 75-6). Quine proposes here to replace the traditional epistemological enterprise of answering Cartesian skepticism with an analogous enterprise in developmental psychology. It is worth noting that Quine does not intend this *Replacement Proposal* to preempt the normative study of the conditions in which we are justified. It would still be possible to study justified belief in a sense of "justified" that does not presume the ability to rule out the demon hypothesis–an alternative, nonskeptical, though traditional, conception of epistemology (Quine 1986, 664).[8]

Quine makes the Replacement Proposal in "Epistemology Naturalized," but it is not his last word on the traditional epistemological task of answering Cartesian skepticism. In the later *Roots of Reference* (1974) and "The Nature of Natural Knowledge" (1975), he retracts the Replacement Proposal and maintains instead that naturalized epistemology, though a branch of psychology, *can* answer Cartesian skepticism.[9] Naturalized epistemology may proceed as "an enlightened persistence rather in the original epistemological problem" (1974, 3). Here Quine rescinds his earlier concession that relying on science in answering skepticism is circular. For, it turns out, the doubts that give rise to skepticism depend on knowledge: "Doubt prompts the theory of knowledge, yes; but knowledge, also, was what prompted the doubt. Skepticism is an offshoot of science" (1975, 67). In particular, science tells us that there is a gap between sensory stimuli and theory. Thus, naturalized epistemology is both psychology and a traditional epistemological effort to answer Cartesian skepticism. Quine does, however, allow that the skeptic would be entitled to use science to argue against the existence of knowledge by reductio ad absurdum.[10]

In his searching study of Quine's naturalized epistemology, Barry Stroud (1984) lodges two criticisms of Quine's claim that naturalized epistemology can offer an answer to Cartesian skepticism.[11]

(1) Quine's allowance of the legitimacy of a reductio argument is, according to Stroud, a major concession to skepticism, opening up the possibility that we cannot rule out the demon hypothesis and thus that the skeptic's argument succeeds. While Stroud is correct that Quine opens up here the possibility that the skeptic can argue against the existence of knowledge by reductio, Stroud overestimates the skeptic's chances of success. For he misunderstands the nature of the skeptic's reliance on science, as Quine see it. There can be illusions, Quine holds, only if there is veridical experience. Thus, Quine concludes, the skeptic relies on science in a way that rules out global illusion, as in the demon hypothesis. The most the skeptic could hope to establish by relying on science is that there are some illusions and that any particular instance of sensory stimulation could be illusory, though many instances are not. But these points support skepticism only if knowledge requires, not merely that we be able to rule out the demon hypothesis, but that no instances of sensory stimulation on which knowledge is based could be illusory–a much stronger and rather less plausible requirement.

And there is a second reply to Stroud's criticism here: the proposed skeptical reductio is no ordinary reductio. It does not, as an ordinary reductio argument would, first hypothesize for the reductio that we know science, deduce from this the contradictory conclusion that we do not know science, and then discharge the hypothesis to conclude that we do not know science. Rather, it hypothesizes only that science is *true*; thus the conclusion that we do not know science does not contradict the hypothesis. For this reason, the skeptic cannot, as in an ordinary reductio, discharge the hypothesis to conclude that we do not know science. The argument's conclusion is the weaker one that if science is true, then we do not know science. Thus, to argue by this sort of reductio, the skeptic must admit that science is true and indeed that the basis for believing that we do not know science is that science is true. Clearly, such an admission deprives the skeptic's conclusion of its traditional force. Thus, it is by no means obvious that Quine does make a significant concession to the skeptic when he allows that a skeptic may argue against the existence of knowledge by reductio. If Quine is right that the skeptic's doubts depend on science, he may be able to offer an effective answer to skepticism.

(2) Unfortunately for Quine, however, Stroud's second criticism is telling: Quine *is* mistaken in thinking that the skeptic's doubts depend on science, and this vitiates Quine's view that epistemology naturalized is an enlightened persistence in the traditional task of answering skepticism. As Stroud points out, the Cartesian skeptic does not need to establish that there are *actual* illusions, only that we cannot rule out the *possibility* of illusions. The former claim depends for its warrant on science, but the latter claim arguably does not. To establish this claim, the skeptic does not need to establish that illusions are actual, or even that they are *metaphysically* possible. The skeptic need only claim that global illusion is *epistemically* possible–i.e., that a deceiving demon has not yet been ruled out in the skeptical dialectic. But science is not needed to warrant this claim; indeed, the claim might be regarded as standing in no need of warrant, since it merely defines the skeptical dialectic. So Stroud is right: Quine's naturalized epistemology has no prospect of answering Cartesian skepticism.

Although Stroud is right that epistemology naturalized does not answer Cartesian skepticism, one who wishes to reject Cartesian skepticism, as Quine does, need not abandon traditional epistemology and retreat to his earlier Replacement Proposal. One could maintain that there is knowledge, not by answering Cartesian skepticism in the sense of ruling out the demon hypothesis, but by rejecting the Cartesian assumption that knowledge requires ruling out the possibility of a deceiving demon. The Cartesian assumption is, after all, unmotivated and counterintuitive (since it leads to skepticism). Moreover, one could turn to the task of answering various forms of nonCartesian skepticism which clearly admit empirical information, as I will explain below. Let it be noted, too, that nothing we have seen so far shows that we cannot appeal to empirical information in the business of rejecting the Cartesian assumption, as opposed to answering Cartesian skepticism. If, for example, empirical information leads us to deny that there is such a distinction between data and theory as the Cartesian assumption presupposes–i.e., a distinction between sensory experience and hypotheses that explain it–then it would also lead us to reject the assumption (though, let it be noted, Quine himself accepts the distinction between observation beliefs and theory).

Some philosophers deny that it is an option to reject the Cartesian assumption that knowledge requires ruling out the possibility of a deceiving demon, or to refuse the task of answering Cartesian skepticism, or to employ empirical information in our answer to any traditional sort of skepticism. At any rate, these philosophers deny that these moves are an option for anyone wishing to persist in traditional epistemology. Stroud (1989), for example, argues that, whether or not we reject the Cartesian assumption that knowledge requires ruling out the possibility of a deceiving demon, no enterprise that resembles traditional epistemology may rely on empirical information in the business of judging whether we have knowledge in the empirical domain. David Shatz (1993), too, argues that the traditional epistemological project forbids reliance on empirical knowledge. I will briefly consider Shatz's reason for saying so.

Shatz argues the point without employing the Cartesian assumption that knowledge requires ruling out the possibility of a deceiving demon, since that assumption may be disputed on the ground that it rests on an inaccurate analysis of the concept of knowledge. Rather, he adduces a different reason: if we were allowed to rely on empirical information to argue that our physical object theory is preferable to the demon hypothesis, then a subject who subscribes to a mad theory (like the demon hypothesis) would equally be allowed to rely on his or her mad beliefs to establish that this theory is preferable to our physical object theory; yet we would not allow such reliance. But this argument is vulnerable to two responses. (1) The admission that the mad subject is allowed to rely on his or her mad beliefs to argue that the demon hypothesis is preferable is disturbing only if it gives the mad subject some prospect of establishing that the demon hypothesis is preferable to the physical object hypothesis. But it is by no means obvious that there is any prospect of establishing that the demon hypothesis is preferable to the physical object hypothesis. (2) Shatz assumes that an allowable argument for the preferability of a theory is constrained in such a way that *we* can succeed in establishing the preferability of physical object theory, but the mad subject cannot succeed in establishing the preferability of the demon hypothesis. But why assume that an allowable argument is so constrained? Evidently, Shatz assumes that an allowable argument must preclude our being mistaken in our conclusion that physical object theory is preferable to alternative theories. In other words, the means we employ to establish that physical object theory is preferable to the demon hypothesis must establish this conclusion in such a way that the truth of the conclusion follows from the means of establishing it. But this assumption, which I call *independent accessibility internalism*, is questionable (Schmitt 1992, 50-52). The assumption, indeed, entails the Cartesian assumption that knowledge requires ruling out the possibility of a deceiving demon. Thus, Shatz's argument that the traditional epistemological project forbids reliance on empirical knowledge in the end relies on the Cartesian assumption that he hoped to avoid making. Of course, if one rejects that assumption, then there is no longer a Cartesian skeptical challenge that needs to be answered, and the question whether it is permissible to employ empirical information in answering that challenge becomes moot. The assumption of the Cartesian skeptical challenge–that knowledge requires ruling out the possibility of a deceiving demon–does forbid the use of empirical information, as Stroud insists. But that premise is dubious. And once it is rejected, there is no challenge we need to answer.

2.2 Strawson on Skepticism; Academic Skepticism and Pyrrhonism

In *Scepticism and Naturalism* (1985), P. F. Strawson responds to skepticism quite differently from the later Quine. Rather than attempt to *answer* Cartesian skepticism, as Quine does, Strawson rejects the *appropriateness* of the Cartesian skeptical challenge. Unlike Quine's response, Strawson's *is* a response to which empirical information could very likely contribute.

Strawson urges us to abandon "any attempt to justify or validate by rational argument" (1985, 51) the belief in body or other beliefs challenged by traditional skepticism. Toward this end, he develops a line of thought he finds in Hume:

> ...all arguments in *support* of the skeptical position are totally inefficacious; and by the same token, all arguments *against* it are idle. [Hume's] point is really the very simple one that, whatever arguments may be produced on one side or the other of the question, we simply *cannot help* believing in the existence of body, and *cannot help* forming beliefs and expectations in general accordance with the basic canons of induction (1985, 11).

Our beliefs in bodies, in other minds, and our inductively based expectations are "pre-rational, natural, and quite inescapable" (1985, 51). Such beliefs lie outside the proper scope of doubt and justification, instead setting "the natural limits within which, and only within which, the serious operations of reason, whether by way of questioning or of justifying beliefs, can take place" (1985, 51).

Strawson's position rests on two empirical claims: that arguments supporting skepticism are unpersuasive, and that our beliefs about bodies are prerational, natural, and irresistible. If Strawson is right that these empirical claims entail a restriction on the range of beliefs that can be doubted or justified, then there is room for cognitive science to bear on skepticism. For cognitive science could confirm or disconfirm these empirical claims and thus debar or admit skeptical challenges. Indeed, though these empirical claims have a ring of truth, they are tenuous enough to need testing by cognitive science.[12] Cognitive science is, then, relevant to the evaluation of skepticism, as Strawson conceives of skepticism.

David Shatz (1993) has objected to Strawson's rejection of skepticism on the ground that the empirical claims to which it appeals are themselves vulnerable to skeptical doubt. Even if Strawson is correct in claiming that beliefs about bodies are irresistible, this empirical claim is not itself irresistible. So, on Strawson's own showing, it is not exempt from skeptical challenge. Since this empirical claim would be challenged by the skeptic, Strawson begs the skeptic's question when he appeals to the claim in his response to skepticism.

Shatz's objection to Strawson misfires, however. It confuses Strawson's rejection of the *appropriateness* of the skeptical challenge with something very different, an attempt to *answer* that challenge. When one *answers* the skeptical challenge, one accepts the skeptic's assumptions (such as the Cartesian assumption that we know only if we can rule out the possibility of a demon). One then tries to establish, from premises acceptable to the skeptic, that we have knowledge conforming to these assumptions. But when one rejects the *appropriateness* of the skeptical challenge, as Strawson does, one proceeds quite differently: one attacks the skeptic's assumptions, either by denying the need to establish that we have knowledge from premises acceptable to the skeptic or by denying the assumed conditions of knowledge. Clearly, if the skeptic wishes to give a reasoned argument for skepticism, the skeptic must grant an opponent the resources to evaluate the assumptions of the argument. These allowed resources obviously include whatever it takes to evaluate these assumptions. Now, it is plausible enough that evaluating the skeptic's assumed conditions of knowledge is a matter of *a priori analysis* of the concept of knowledge, and not a matter of empirical inquiry. But there is another of the skeptic's assumptions that may be evaluable only on the basis of *empirical* information–namely, the assumption that the category of knowledge or justified belief *applies* to given beliefs queried by the skeptic (i.e., that a given belief is the sort of belief that can be justified or unjustified).[13] If, for example, the category of justified belief is restricted to resistible beliefs, as Strawson claims, then our judgment whether this category applies to a given belief turns on empirical information, since our judgment whether a given belief is resistible turns on empirical information. The skeptic must allow us such empirical information in our evaluation of the assumptions of the skeptical argument. Since Strawson does not attempt to answer the skeptical challenge, but rejects the appropriateness of that challenge, he may rely on whatever empirical information is needed to determine whether the challenge is really appropriate for its target beliefs.

Of course, Strawson may be mistaken in claiming that the category of justification applies only to resistible beliefs–that is an issue that must be decided by a priori analysis. And if Strawson is wrong in claiming this, then cognitive science may not bear on whether the skeptical challenge is appropriate for its target beliefs. My point is only that *if* his claim is correct, then cognitive science *does* bear on skepticism, notwithstanding the fact that its findings are empirical. The skeptic cannot respond by charging that empirical beliefs are in doubt, since at this point in the dialectic, we are still considering how broadly the category of justification applies, and thus we are still engaged in the business of judging the plausibility of the skeptic's assumptions; we have not yet embarked on the leg of the skeptical dialectic that begins with doubt.

I have so far considered, in my discussion of Quine and Strawson, only *Cartesian* skepticism, the kind most resistant to the relevance of empirical information. I have argued that cognitive science bears on Cartesian skepticism if Strawson is right that a skeptical challenge is appropriate only for resistible beliefs. There are, however, many other forms of skepticism besides the Cartesian variety, and in most cases, the arguments for these forms of skepticism clearly depend on empirical claims, and the skeptic must clearly admit empirical claims in evaluating the view. Opponents of the relevance of cognitive science to skepticism often assimilate skepticism to Cartesian skepticism (and equally often assimilate traditional epistemology to the business of answering Cartesian skepticism), a move that makes their position easier to defend.[14] But Cartesian skepticism is by no means the historically dominant representative of skepticism. And other versions of skepticism are clearly more vulnerable to empirical information than Cartesian skepticism is.

Hume subsumed Cartesian skepticism under the category of *antecedent* skepticism, or skepticism supported without any explicit reliance on empirical assumptions. He distinguished the latter form of skepticism from *consequent* skepticism, which explicitly infers skepticism from empirical assumptions (1974, 150). Hume had little interest in Cartesian or antecedent skepticism. His own interest was rather in consequent skepticism, and his project was to draw skeptical consequences from his own empirical (albeit introspectionist) associationist psychology. Arguably, Hume was the first philosopher to employ empirical psychology self-consciously and systematically in evaluating skepticism. He rightly took Pyrrhonism and Academic skepticism, both examples of consequent skepticism, as his forebears (see Schmitt 1992 chs 1-3 for discussion of antecedent and consequent skepticism and their relation to internalism and externalism in epistemology).

These versions of consequent skepticism rely on empirical premises in arguing their case. For example, in Pyrrhonism, one begins with observations about human beliefs–e.g., that people who have different experiences and education form opposing beliefs on various topics (Sextus 1976). The Pyrrhonian argument thus depends on empirical premises. To judge the merits of the argument, we must judge, for example, just how widely beliefs on given topics vary across cultures. To judge this matter we ought to rely not only on the dictates of common experience but on the results of cognitive anthropology. Of course, once we allow empirical information in evaluating the premises of the Pyrrhonian argument, the outcome of the inquiry could be either skeptical or nonskeptical. The same is true of Hume's arguments for skepticism based on his associationist psychology. In short, Pyrrhonism is one form of skepticism that must be evaluated in light of empirical information.

3 WHICH SORTS OF BELIEFS ARE JUSTIFIED?

I have argued that empirical information is relevant to assessing some forms of skepticism. Let us now turn to a different question: is empirical cognitive science relevant to (admissible in and also helpful or necessary for) the task of *systematically assessing which beliefs, or sorts of beliefs, are justified*? This task is part of the theoretical endeavor of gauging our cognitive achievements. (It might also be put to practical use as information in light of which we may revise our beliefs–about which I will, for limitations of space, say very little in what follows.) I will call the task of assessing which sorts of beliefs are justified, for purposes of gauging our cognitive achievements, the task of assessing justification. The task proceeds by assuming a particular account of justification (the correct account, it is hoped) and judging which sorts of beliefs are justified by checking which sorts of beliefs conform to the conditions proposed in the account of justification. Whether cognitive science is relevant to this task depends on whether the given conditions of justification *make* it relevant: are the given conditions of justification the sorts of conditions cognitive science could discover to obtain (or fail to obtain)?[15] (In sections 3 and 4, I will be concerned exclusively with empirical cognitive science, rather than with engineering cognition or with a priori AI, so I will drop the qualification "empirical" in these sections. We will return to AI in section 5.)

It seems that cognitive science *could* be employed to some slightly beneficial effect in assessing which beliefs are justified on most of the currently viable proposed conditions of justification. The important issue, in my view, is whether cognitive science could be very helpful for intelligently assessing justification. On some proposed accounts of justification, cognitive science is even *necessary* for an intelligent assessment of justification. It is these accounts I find most interesting, and I will accordingly review them, one by one. I will end the section by considering two objections to the very idea that cognitive science could be helpful in assessing justification, and I will endorse one of these objections.

3.1 Goldman on Reliabilism and Probability Judgments

In *Epistemology and Cognition* (1986), Alvin Goldman develops a *reliabilist* account of justification, and he systematically considers its consequences for which sorts of beliefs are justified in light of findings from cognitive science. On Goldman's reliabilism, we will need information from cognitive science to assess which sorts of beliefs are justified. Indeed, on his account, we will need cognitive science even to complete the formulation of the conditions of justification. For it is impossible to determine in detail which sorts of beliefs are justified without systematic reliance on cognitive science. There is a sense in which, for Goldman, there is no distinction between the task that concerns us here, assessing justification on given conditions of justification, and the seemingly different theoretical enterprise of selecting the correct conditions of justification (which we will take up in section 4). However, the conditions of justification that must be selected here are not a priori necessary conditions of justification (which, for Goldman, are selected by narrow reflective equilibrium method—see subsection 4.1) but *empirical, contingent* conditions of justification. (Goldman has subsequently reconfigured his reliabilism, without, however, abandoning the claim that cognitive science is relevant–on which more below. But his book *Epistemology and Cognition* remains even after more than a decade the only systematic review of the relevance of cognitive science to epistemology in the literature, and incomparably the most probing and detailed treatment of the subject. I see no alternative to discussing it.)

To see where cognitive science comes into the task of assessing which sorts of beliefs are justified, and equally to appreciate how cognitive science comes into Goldman's theoretical enterprise of selecting the conditions of justification, we need to distinguish three stages in Goldman's account of justification. In the first stage, justification is related to epistemic permissibility:

> A belief is *justified* only if it is permitted by a right system of J-rules.

In the second stage, Goldman proposes a "criterion" of right J-rules:

> A J-rule system is *right* just in case the basic or native belief-forming processes it permits would, when exercised, yield a high ratio of truths to total beliefs.

In the third stage, the aim is to determine the contents of some right system of J-rules by determining which processes would, used in tandem, lead to a high truth ratio. Goldman argues at length that justification turns on the character of our native belief-forming processes and not merely on our beliefs or other psychological states or on what evidence we possess. Roughly, his argument is that, no matter which prior states we assume, and no matter how good the evidence for a belief, a subject can still fail to be justified in the belief if he or she does not exercise the right kind of process in arriving at the belief. Moreover, though a belief may result from an accepted acquired method, this method will not make the belief justified unless it was selected in the right way. Though I will register a qualification shortly, I find Goldman's argument conclusive on this point: acquired methods are justifying only if selected in the right way.[16] The pertinence of native psychological processes, Goldman avers, entails the relevance of cognitive science to epistemology.[17]

The first two stages in Goldman's account of justification are a priori enterprises conducted by a narrow reflective equilibrium method. It is in the third stage that cognitive science comes into the picture. For judging which J-rule systems are right (if any) requires determining which human belief-forming processes are native and which subsets of these native processes would, when exercised in tandem, lead to a high truth-ratio. Both of these third stage tasks require extensive findings from cognitive science. (Of course, the assessment of reliability may require findings from outside of cognitive science.) Goldman accordingly undertakes a review of cognitive science findings that might help the third stage tasks. The aim of his review is primarily to determine the contents of right J-rules. Indeed, we cannot so much as guess what the right J-rules are without a review of cognitive science. In this regard, Goldman's theory differs strikingly from recent foundationalist and coherentist theories. These theories can and do deliver detailed rules of justification or specifications of which beliefs are justified without relying on scientific empirical information. But Goldman must rely on cognitive science to produce a theory of justification comparable in detail to these other theories. Cognitive science is needed not merely to judge which beliefs are justified or which processes are justifying, but, in one sense, to formulate detailed, albeit contingent and empirically-based, conditions of justification–this despite the fact that the first and second stages of his account are a priori.[18]

To appreciate just how Goldman's assessment of the reliability of our belief-forming processes employs cognitive science, it will be desirable to consider at length a specific kind of belief–probability judgments. In recent decades, psychologists who study reasoning have accumulated experimental evidence that people routinely violate a large number of textbook norms of rational probabilistic and statistical judgments.[19] For example, in probabilistic judgments people ignore prior probabilities, while in statistical generalization, people ignore regression to the mean and sample size.[20] I will return to the error of ignoring sample size in subsection 3.2. Here I will focus on Tversky and Kahneman's (1983) finding that people err in their probability judgments by violating the conjunction rule–the theorem of the probability calculus according to which the probability of a conjunction is no greater than the probability of each conjunct.

Tversky and Kahneman gave subjects stories like the following:

> Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations.

Subjects were then asked to rank the probabilities of various propositions, including:

Linda is a bank teller. (T)
Linda is active in the feminist movement. (F)
Linda is a bank teller and is active in the feminist movement. (T and F)

Violations of the conjunction rule were very common. Even in a test in which the relation between the conjunction and its conjuncts was emphasized, 85 percent of the subjects indicated that the conjunction T and F was more probable than T. Tversky and Kahneman also found that even sophisticated subjects (e.g., doctoral students in decision science) fared poorly, and indeed on some tests they fared no better than naive subjects. Tversky and Kahneman explain these findings by the hypothesis that people make probability judgments using a "representativeness heuristic." The probability that Linda is a bank teller is proportioned, according to this explanation, to the degree to which she is representative of bank tellers. Since she is more representative of feminist bank tellers than of bank tellers, the probability assigned the former is higher than that assigned the latter. Tversky and Kahneman found a very high correlation between rankings of representativeness and rankings of probability.

The question Goldman addresses in connection with this finding is, of course, that of the normative status of violations of the conjunction rule and the use of the representativeness heuristic: can probability assignments be rational or justified despite violating the conjunction rule? I will focus here on Goldman's treatment of whether such assessments can be justified. I would like to concede before proceeding, that there is some difficulty in interpreting the empirical findings here: it is unclear whether subjects in the Tversky-Kahneman experiments are reporting *subjective* probabilities or instead assigments of objective probability. I will simply assume for convenience that they report subjective probabilities. Even if subjects are not reporting subjective probabilities, it is plausible enough that the same results would have been obtained if subjects had clearly reported subjective probabilities.

Central to Goldman's treatment of the justification of probability assignments is his distinction between *native belief-forming processes* and *acquired belief-forming methods,* and we must consider carefully the epistemological significance of this distinction before we consider his view of whether violations of the conjunction rule are justified. Native belief-forming processes are those we natively exercise, without training or education, including many perceptual, memorial, and inferential processes, while we acquire the use of belief-forming methods, such as logical, probabilistic, and statistical methods, as a result of training. Goldman argues (persuasively, in my view) that there is a fundamental epistemological difference between native processes and acquired methods. Proper or right acquired methods, unlike proper native processes, yield justified beliefs only when selected by a right native second-order method-selecting process. Goldman argues the point by appeal to an example:

> Suppose Gertrude's mathematical education is seriously deficient: she has never learned the square root algorithm. One day she runs across the algorithm in a pile of papers written by someone she knows to be a quirky, unreliable thinker, and no authority at all on mathematical matters. Despite this background knowledge, she leaps to the conclusion that this rule for deriving square roots (the rule so labeled) is a sound method. She proceeds to follow, and form beliefs in accordance with, this algorithm. She forms beliefs in propositions of the form "x is the square root of y." Are these beliefs justified? Clearly not, for Gertrude has no adequate grounds for trusting the result of this algorithm. (1986, 91)

If Gertrude selects an algorithm in this way, without any reason to believe that it accurately calculates square roots, and uses it to form beliefs about square roots, then, Goldman argues, she is not justified in the beliefs she obtains by using the algorithm, no matter how accurate the algorithm may be. A square-root algorithm must be selected by a right native second-order method-selecting process if it is to yield justified beliefs. Goldman generalizes this point to all acquired methods. Thus, if people fix subjective probabilities by an acquired arithmetical method, the justificatory status of their subjective probabilities depends not only on whether the method is right, but also on whether the method was selected by a right native selection process. Goldman suggests that a right method is a reliable one. (In the case of a method that yields subjective probabilities, a right method is *well-calibrated*–i.e., a method for which the degree of probability assigned a proposition is equal to the frequency of true propositions assigned that degree of probability). He also suggests that a right native second-order method selecting process is one that is *metareliable*–tends to select reliable methods. On Goldman's account, then, the question whether subjective probabilities are justified is a complex one, requiring us to distinguish the case in which subjective probabilities are formed by a native process from the case in which they are formed by an acquired method–the latter case involving in addition an assessment of the metareliability of the native second-order selection process.

Goldman is surely correct about the case of Gertrude: Gertrude's square-root beliefs are clearly not justified merely because she uses an inadvertently reliable square-root algorithm. And this does show that for some beliefs that result from the use of reliable acquired methods, the method must be selected by a right selection process. But is it always true that acquired methods yield justified beliefs only when selected by a right selection process? I am inclined to resist Goldman's generalization. Consider standard arithmetical algorithms for sums and products acquired by normal instruction in primary school. Intuitively, young children use these methods to form justified beliefs about sums and products. Yet they have no proper native second-order process for selecting these algorithms–at least, not if a proper selection process is a psychological process in the individual, and one that must be metareliable. (It is indeed misleading to describe them as *selecting* the algorithms at all.) Students rely on the algorithms the teacher supplies them without possessing any resources for discriminating these algorithms from other, unreliable ones and indeed without even having much basis for imputing reliability to the teacher. It even seems possible for a standard algorithm for sums or products to be acquired accidentally–by a bump on the head–and still be justifying. This suggests that reliable acquired methods may be right in certain circumstances even if not selected by a metareliable selection process. (I would also make the converse point that unreliable algorithms may be right, so long as they are selected by a metareliable selection process. See Schmitt 1992, 163-174 for further discussion of these issues.)

Despite my reservation about Goldman's generalization, I agree with him that methods for assigning subjective probabilities yield justified subjective probabilities only if they are selected by a right native selection process. Intuitively, these methods are closer to square-root algorithms than to standard arithmetical methods for judging sums and products. They are bound to be complex, and more importantly, we are apt to learn them at an advanced point in our development, a point at which we are not entitled simply to take the word of an instructor but must possess evidence for the rightness of the method. For these reasons, using a randomly selected method that luckily turns out to be right (where rightness is, say, conforming probabilities to the probability calculus) does not, intuitively speaking, suffice for justified subjective probabilities. If the subject has no reason to believe that a subjective probability-assigning method is right, or has not selected it in some bona fide way, the mere use of that method does not make his or her subjective probabilities justified.

With these points about the distinction between acquired methods and native processes behind us, let us return to the task of assessing whether subjects can be justified in subjective probabilities that violate the conjunction rule, on Goldman's account of justification. For Goldman, we must distinguish between subjective probabilities assigned by a native process and those assigned by an acquired method. Consider first subjective probabilities assigned by a *native* process. Here what determines whether the subjective probabilities are justified, on Goldman's view, is whether the process contributes to a J-rule system that has a high truth-ratio or, in the case of probabilities, a system that is well-calibrated in the sense that a proposition is assigned a subjective probability equal to the proportion of truths in the class of propositions assigned that probability. It is plausible that if a subjective probability-forming process is well-calibrated, then it will contribute to the good calibration of beliefs and thus be sanctioned by a right J-rule system. Thus, plausibly, if subjective probabilities are natively assigned by a well-calibrated process, they will be justified, on Goldman's view. The remaining question, then, is whether a well-calibrated subjective probability-forming process will conform its output probabilities to the probability calculus. If not, then subjects can be justified in subjective probabilities that violate the conjunction rule. There are two points in favor of a negative answer to this remaining question. First, as Goldman suggests, it is doubtful that native subjective probability-forming processes are capable of conforming subjective probabilities to the probability calculus. Such conformity requires an arithmetical facility that we do not natively possess. Second–a point Goldman does not make–even if we do natively assign subjective probabilities, and even if our processes are well-calibrated, the good calibration of a subjective probability-forming process does *not* entail that it conforms its output to the probability calculus. In fact, it is quite possible for a well-calibrated method to yield subjective probabilities which systematically and grossly violate the probability calculus–e.g., to make all probabilities of conjunctions greater than the probabilities of their conjuncts.) In any event, it is less than obvious that well-calibrated subjective probability-forming processes would generally yield subjective probabilities that obey the conjunction rule. Thus, it is plausible that, on Goldman's view, subjective probabilities can be justified even though they violate the probability calculus.

Goldman does note, however, that there remains on his view one last way in which subjective probabilities natively formed by a well-calibrated process could fail to be justified. Good calibration of the native process could fail to suffice for justification because there are methods available to the subject that are superior in conforming subjective probabilities to the probability calculus. Here Goldman distinguishes two cases: subjects amply trained in a superior method and naive subjects. Suppose someone amply trained in probability calculus nevertheless natively forms subjective probabilities in violation of the calculus. Goldman allows that such a person ought to preempt the native tendency to violate the conjunction rule by using the superior acquired method. In this case, the sophisticated subject's natively assigned subjective probabilities are not justified. Goldman maintains, however, that it is more difficult to decide what to say about a naive subject. Presumably, the case will depend in part on whether that subject has had the *opportunity* to acquire a superior method. If the subject has had such an opportunity and should have acquired that method, then we might put the subject in the same category as the sophisticated subject who already possesses the method but fails to use it. The availability of a method conforming to the probability calculus, then, may entail that subjective probabilities formed by well-calibrated native processes are not justified.

All this is on the assumption that the subjective probabilities that violate the conjunction rule result from a *native* process. What if the subjective probabilities result from *acquired* methods? Goldman does not discuss this case, but presumably he would say that in general a well-calibrated method is justifying if selected by a right selection process, and such a method could yield subjective probabilities that violate the probability calculus. But if the subject has available a superior method, one that conforms subjective probabilities to the probability calculus, then the subject's subjective probabilities are not justified. My primary reservation about Goldman's allowance that subjective probabilities can fail to be justified because there is available a method that conforms to the probability calculus is that there is no basis in Goldman's reliability theory for preferring a method on the ground that it conforms to the probability calculus.

In addition to these points about the justification of subjective probability assignments, Goldman offers an intriguing evaluation of the representativeness heuristic. He does not deny that the representativeness heuristic leads to errors in the cases which violate the conjunction rule. Certainly this is true if the judgments are of objective probabilities. But whether the judgments are objective or subjective probability judgments, the representativeness heuristic could be admissible in a right J-rule system. For it could be *generally* reliable. Indeed, Goldman points out, the representativeness heuristic could be a facet of general matching operations or similarity assessments in cognition.

According to the prototype theory of concepts (Rosch and Lloyd 1978, Smith and Medin 1981), people represent a concept like the concept of bird by a prototype (e.g., by averages of properties across examples–flies, has wings and feathers, etc.), and they judge whether a given object satisfies the concept of bird by measuring its similarity to the prototype: objects sufficiently similar to the prototype are judged to be birds. On this theory, the use of representativeness for making judgments, here understood as similarity to a prototype, is a pervasive feature of human cognition. Moreover, the routine of matching to a prototype is highly reliable, so long as subjects roughly agree in their prototypes and their measure of similarity, since in this case an object judged to satisfy a concept *Y* will really be a member of *Y* under the lexical label employed by the linguistic community. The representativeness heuristic used to make the probability judgment that Linda is a bank teller can be understood as a facet of a matching operation. Linda's known features are matched to the prototype of a bank teller, and the probability is identified with the degree of her similarity to the prototype. If the heuristic is assimilated to the matching operation, then it will inherit the reliability of that operation, even though it gives erroneous results when used to make certain probability judgments. Goldman offers an analogy to the use of generally reliable perceptual processes that generate illusions under certain conditions (e.g., the moon illusion, the Ames-room illusion). Despite yielding erroneous judgments in certain cases, these processes will be sanctioned by a right J-rule system and so will be justifying (unless corrected by learned methods).

Goldman does, however, qualify his endorsement of the representativeness heuristic by observing that its use to make probability judgments differs from the matching operation in one respect: in the categorization cases, the task is to decide whether exemplification of the known features (e.g., flies) suffices for the target property (bird), while in the probability judgment cases, it is not assumed that the known features (e.g., Linda's being outspoken) suffice for the target property (being a bank teller)–a probability of having the target property is assigned by reporting the degree of similarity of the known features to the prototype of the target property. Goldman concedes that this difference could lead us to conclude that the representativeness heuristic is not a facet of the matching operation and does not inherit the general reliability of that operation. However, a more important objection to Goldman's case for the reliability of the representativeness heuristic–an objection that Goldman does not note–is that only *part* of the representativeness heuristic is a facet of the matching operation. For the representativeness heuristic consists not only of measuring the similarity of the given object to the target property, but also of identifying the probability with the degree of similarity. Even if the former part of the heuristic inherits its reliability from the matching operation, the latter is alien to that operation and is the source of the error in violations of the conjunction rule. Despite this, the representativeness heuristic could still be reliable for probability judgments if its unreliability in conjunctive judgments does not overwhelm its reliability elsewhere.

Goldman's treatment of Tversky and Kahneman's results remains a sterling example of a detailed effort to employ cognitive science to assess which sorts of beliefs are justified. It is an effort that reveals how very difficult it is to judge the implications of an empirical finding for which sorts of beliefs are justified, even on an account of the conditions of justification like Goldman's own, which makes empirical findings essential for assessing which sorts of beliefs are justified (and simultaneously for the theoretical task of formulating detailed contingent conditions of justification).

3.2 Kornblith on Induction

Hilary Kornblith has proposed that we conceive of epistemology "as addressed to two questions: (1) What is the world that we may know it?; and (2) What are we that we may know the world?" (1993, 2) Many epistemologists would resist the idea that epistemology is defined by these two questions or even so much as addresses them: the first question properly belongs to metaphysics and the second to the philosophy of mind. Even so, no one can deny that the questions Kornblith wishes to address are traditional philosophical questions: they form the central focus of the Kantian tradition. In this tradition, the attempt to answer skepticism gives way to an explanatory project: we try to say how knowledge is possible by describing what we and the world must be like in order for us to know. Kornblith differs sharply from the Kantian tradition, however, in refusing, with Quine, to distinguish a priori from empirical questions: his project is part of cognitive science. The "must" in "what we and the world must be like in order to know" is a *nomological*, not a logical "must."

Despite the preference here for metaphysics and explanatory cognitive science over the analysis of knowledge, Kornblith does need some working account of knowledge when he addresses his two questions, and it is fair to say that he tacitly assumes reliabilism. (He differs from Goldman in his account of knowledge primarily in embracing wide reflective equilibrium method (1993, 4-5): if we had learned from cognitive science that our processes are *not* reliable, this would, on Kornblith's method, adversely affect the plausibility of reliabilism itself, while for Goldman such a discovery would have no impact on the plausibility of the view. But this difference between them plays no role in Kornblith's inquiry, since cognitive science does not in fact tell us that our processes are unreliable.) Kornblith is accordingly concerned to determine which features we and the world must have if our belief-forming processes are to be reliable. The world, he argues, must contain natural kinds as homeostatic property clusters (self-maintaining clusters of correlated properties). And we, for our part, must have innate dispositions to form natural kind concepts, as well as a propensity to induction on natural kind properties rather than on other properties. Kornblith is concerned not only to make these points but to survey cognitive science for evidence that we in fact satisfy these requirements. Here his enterprise falls squarely under the task of assessing justification (or knowledge), and he pursues the task much as Goldman does, seeking evidence from cognitive science.

Regarding innate dispositions to form natural kind concepts, work in developmental psychology shows that young children judge diverse objects to have the same properties by relying on whether they belong to the same kind, rather than on whether they superficially resemble one another.[21] Regarding a propensity to induction on natural kinds, Kornblith offers no direct argument that we have such a propensity. Rather, he describes a strategy for searching our environment that would enable us to identify natural kind properties, and he notes that induction on natural kind properties enables us to infer reliably from very small samples. I will briefly discuss these points about induction.

In a study of inductive inference, Tversky and Kahneman (1971) found that people routinely draw conclusions about a population on the basis of a very small sample, showing more confidence in their generalizations than is warranted by textbook statistical practice, which proportions confidence in the generalization to the size of the sample. This finding is robust for problems of varying complexity and subject matter, and it extends even to sophisticates in statistics like the members of the mathematical psychology group of the American Psychological Association, who might reasonably be expected to observe textbook statistical practice. Nisbett and Ross (1980) have studied a related phenomenon: in inductive generalization, vividly described examples tend to overwhelm less vivid but more significant statistical information. As Kornblith notes, however, the fact that we project properties on small samples tells us nothing about the reliability of our inductions.

For induction from a small sample, even from a single case, is perfectly reliable when the population is uniform with respect to the property: "If I note that a sample of copper conducts electricity and straightaway conclude that all copper conducts electricity, then I will do just as well as someone who insists on checking a very large number of copper samples for their conductivity" (1993, 92-3). Thus, if we tend to project only on homogeneous populations or natural kind properties, which are highly correlated, our inductions from small samples will tend to be reliable.[22] True, the studies of Tversky and Kahneman show that we are not perfectly sensitive to homogeneous populations and natural kind properties. The question is whether we are fairly sensitive to these properties.

What sensitivity to natural kind properties requires is an ability to detect covariation. Now, the psychological literature on covariation detection brings some bad news. In data-driven assessments of covariation, in which people do not hold antecedent views about the degree of covariation of the properties at issue, people routinely ignore crucial negative evidence regarding covariation (e.g., they judge whether a symptom correlates with a disease by relying almost exclusively on cases in which the symptom and the disease are present and ignoring cases in which one or both are absent) (Nisbett and Ross 1980, 91). People also routinely fail to detect covariation until the degree of covariation is high, and even then they underestimate the degree (Jennings, Amabile, and Ross 1982, 221). In theory-driven assessments of covariation, our performance is even worse than in data-driven assignments: we ignore the data and project covariations dictated by our prior beliefs about covariation.

Despite this bad news about covariation, Kornblith reports some good news of greater significance for the reliability of induction. When the degree of covariation is nearly perfect, we do well in detecting it–a crucial point for our ability to detect natural kinds, where covariation is often nearly perfect. More importantly, we are good at detecting cases of *clustered* covariation in which more than two properties covary–as with natural kinds (Billman 1983, Billman and Heit 1988). Dorrit Billman has developed a computer model for detecting clustered covariation which uses the strategy of *focused sampling*, in which objects are examined for properties, and an object is more likely to be examined for further properties if it has properties which figure in covariation hypotheses that have proven successful. If properties *P* and *Q* have been discovered to covary in the sample searched, then objects having *P* and those having *Q* are more likely to be examined for further properties. In cases in which there are more than two covarying properties, this strategy increases the rate of covariation detection–an effect called *clustered feature facilitation.* Billman offers evidence that human subjects employ such focused sampling. This provides an answer to Tversky and Kahneman's discovery that we routinely generalize from small samples. Generalizing from small samples is reliable for natural kind properties; we are able to detect natural kinds using focused sampling; and we tend to generalize on natural kind properties. We have here an explanation of how we and the world are arranged so that we may know by induction.

I am sympathetic to Kornblith's explanation of the reliability of induction. My primary reservation is that his explanation is overly individualistic. His appeal to focused sampling does seem needed on the assumption that people acquire their knowledge without the benefit of education by their elders. On this assumption, people will be able to use single-case inductions reliably only if they discover natural kinds (or at least clustered correlated properties) by focused sampling. But if individual subjects may rely on reliable teachers to direct them to correlated properties, then they will not need to use focused sampling or restrict their inductions to natural kind properties. In this case, the culturally transmitted experience of the human species takes over the job for which we needed focused sampling. Of course, this cultural-historical explanation of how people are able to identify correlated nonkind properties, and of how people can be reliable in single-case inductions on such properties, would require enough past human experience to permit identification of many such properties. It would also require that subjects do not need much successful induction in order to learn from their elders. It is a question for further exploration whether these requirements are met. Whether a cultural-historical explanation is needed depends on whether we perform very many inductions on correlated nonkind properties. If we do perform such inductions, then focused sampling will be insufficient to explain our inductive success, and we will need a cultural-historical explanation of our success. In any event, we ought not to leave unquestioned the individualism tacitly assumed by Kornblith and the psychologists he cites, as indeed by cognitive scientists generally. It is worth noting in this connection that work in the new field of Artificial Life (Meyer 1996) models the role of culturally transmitted information in reasoning competence, albeit in an evolutionary framework.[23]

3.3 Psychologism and Proper Function Theory

Goldman's reliabilism is not the only currently viable theory of justification that entails that we need cognitive science for a detailed specification of the sorts of beliefs that are justified. There are four other theories that do so as well: psychologism, proper function theory, responsibilism (Kornblith 1982, Code 1987, BonJour 1985), and perspectival internalism (Foley 1986, Schmitt 1993). Since Goldman (1993) has already convincingly argued the point regarding responsibilism, and Foley has argued at length that his version of perspectival internalism leads to a contingent foundationalism in light of plausible empirical (albeit folk) psychological assumptions, I will discuss only the first two theories.

According to *psychologism* (Sober 1978, Cohen 1981),

> a belief is justified (or rational) for a subject just in case it conforms to (or results in the right way from) the subject's psychological competence.

(We say that a belief conforms to competence when competence would produce the belief if freed from interference.) Psychologism assumes a distinction between cognitive competence and cognitive performance roughly analogous to Chomsky's distinction between grammatical competence and grammatical performance (Stein 1996). We explain successful human cognitive performance on reasoning or problem-solving tasks by attributing it to psychological mechanisms that constitute the subject's cognitive competence. We attribute unsuccessful performance on these tasks to interference with the exercise of cognitive competence, either because of structural psychological limitations like the small storage capacity of working memory, or because of some noncognitive factor (e.g., an emotion or desire). On psychologism, the difference between justified belief and unjustified belief is the difference between belief that conforms to (or results from the exercise of) cognitive competence and belief that fails to conform to it through interference.

Psychologism may perhaps provide some specification of the sorts of beliefs that are justified without relying on cognitive science. For arguably we do naively ascribe to ourselves a cognitive competence and explain successful cognitive performance by appeal to our cognitive competence. We naively explain reasoning, for example, by attributing a deductive and inductive competence, and we explain failed reasoning by hypothesizing interference with the exercise of this competence. Nevertheless, our naive ascription of competence would seem to be rather primitive, and it is doubtful that we can, on psychologism, specify in a systematic, detailed way the sorts of beliefs that are justified without relying on cognitive science. Of course, the cognitive science on which we rely here would have to assume a competence/performance distinction to be useful to psychologism, and it would have to explain successful and unsuccessful cognition by appeal to that distinction.

One proponent of psychologism, L. J. Cohen (1981), has, however, denied that findings from cognitive science *could* entail that our naive intuitions about rational belief are mistaken. Indeed, this denial is a consequence of Cohen's very argument for psychologism. His argument for psychologism begins with the claim that the correct conditions of rational belief are those in *narrow* reflective equilibrium for the subject–a reflective equilibrium of our naive intuitions about rational belief (see subsection 4.1 for a fuller explanation of narrow reflective equilibrium). (This is a reflective equilibrium view of which conditions of rational belief are *correct*, not merely a view about selecting, or even justifying the selection of, conditions of rational belief, as we have interpreted reflective equilibrium method up till now.) On this claim, it follows that the subject's intuitions about rational belief in narrow reflective equilibrium are correct. Yet, according to Cohen, intuitions about rational belief that are in narrow reflective equilibrium match competence: we would intuitively judge (in narrow reflective equilibrium) that a belief is rational just in case the belief conforms to our psychological competence. The deliverances of intuition of the form "This belief is rational" in narrow reflective equilibrium ascribe rationality to the beliefs that conform to competence. But our intuitions about rational belief that are in narrow reflective equilibrium are (by the very idea of a correct theory of rational belief as one that is in narrow reflective equilibrium) correct. It follows that psychologism is correct: rational belief coincides with competent belief. Cohen has inferred from his psychologism that we cannot discover empirically that human beliefs are irrational (when they conform to competence): human irrationality cannot be experimentally demonstrated. If we wish to know which beliefs are rational, we need only see which ones are intuitively judged rational in narrow reflective equilibrium.

There are several questionable premises in Cohen's argument for psychologism (see Stein 1996 for a review), but I will content myself with pointing out that even if these premises are accepted, and psychologism along with them, and even if, consequently, human irrationality cannot be experimentally demonstrated, there is still one way in which cognitive science could bear here, albeit at the metalevel. The premise that we would intuitively judge (in narrow reflective equilibrium) that a belief is rational just in case the belief conforms to our psychological competence is in fact an empirical premise–indeed, a questionable one that should be submitted to the findings of cognitive science.[24] Checking whether there is a correlation between intuitively rational beliefs and competent beliefs is business for cognitive science, even though intuitive judgments in narrow reflective equilibrium are naive, and even if attributions of competence are naive attributions. Cognitive science could thus help us judge the plausibility of this premise of Cohen's argument for psychologism and in this way contribute to an assessment of the plausibility of psychologism itself. Of course, if I am right that this premise is empirical, then Cohen's psychologism is itself merely empirically supported, and experimentally demonstrating irrationality has merely been empirically ruled out.

Proper function theory may, like psychologism, need cognitive science to develop a systematic, detailed specification of the sorts of beliefs that constitute knowledge. According to proper function theory, a subject knows a proposition just in case the subject's belief in the proposition conforms to, or results (in the right way) from, the proper functioning of the subject's cognitive faculties (or just in case the belief conforms to or results from the subject's cognitive faculties when they fulfill their proper function). This view has been richly developed, albeit in different ways, by Ruth Garrett Millikan (1984) and Alvin Plantinga (1993). Both authors propose accounts of knowledge, rather than justified belief, and Plantinga goes to considerable lengths to distance his theory from the theory of justified belief. Both authors, too, specify that the proper functioning that leads to knowledge is aimed at true belief.[25]

Millikan proposes a proper function theory of knowledge as an empirical theory of knowledge within evolutionary theory. As I read her, her theory proposes, roughly, the best and most useful sense that evolutionary theory can make of our naive concept of knowledge. For Plantinga, by contrast, proper function theory is a conceptual analysis of knowledge. Millikan endorses an evolutionary etiological account of the proper function of our cognitive faculties, and on this account evolutionary theory, and in particular evolutionary psychology, are essential to judging which proper functions our cognitive faculties have and which sorts of beliefs result from them when they are functioning properly. Plantinga, for his part, rejects an evolutionary etiological account of the proper function of our cognitive faculties. But it would seem that on any account of their proper function (etiological, functionalist, or theological), we could profit from the findings of cognitive science in our study of proper function. For our naive view of the proper functioning of our cognitive faculties would seem to be no better developed than our naive view of the proper functioning of our anatomical organs–the heart, say, before the studies of Harvey.

3.4 Objections to Relying on Cognitive Science in Assessing Which Beliefs Are Justified

I have so far presented the case that various currently viable accounts of justification need cognitive science both to assess justification and to develop systematic, detailed conditions of justification comparable to those provided by traditional foundationalism and coherentism (though of course these conditions will be empirical, contingent conditions, not the a priori necessary conditions offered by the latter account). This is true of Goldman's reliabilism, of psychologism, and of proper function theory. While these theories do entail that we must employ cognitive science to assess justification, I myself regard this fact as a serious objection to the accounts. For, although we do need empirical information to assess justification, we really should not need cognitive science to assess justification. I will discuss two important reasons for saying so, endorsing the second of these reasons. We will see, however, that objecting to the particular way these theories depend on cognitive science by no means precludes us from taking cognitive science to be relevant in other ways, nor does it even preclude us from embracing an alternative version of one of the theories we have canvassed, reliabilism.

3.4.1 The Objection from Accessibility Internalism

The idea that cognitive science could be relevant to epistemology was long in coming because the most influential work in recent epistemology–Roderick Chisholm's (1966) foundationalism–assumed a very strong constraint on the conditions of justification, *accessibility internalism* (Ginet 1975, Alston 1989b, Schmitt 1992 ch. 4), and it was assumed, plausibly enough, that accessibility internalism rules out a role for cognitive science in assessing which beliefs are justified. On accessibility internalism, cognitive science may nevertheless be relevant to other epistemological projects.

Accessibility internalism is best understood as one of two constraints on the conditions of justification according to which justification must be accessible to the subject:

Strong Accessibility Internalism: For any candidate for belief *p* to which the category of justification applies, a subject must be able to tell "internally"–i.e., by reflection or introspection alone–whether he or she is justified in believing *p*.

Weak Accessibility Internalism: The subject can tell by reflection alone whether the belief *p* satisfies whatever conditions happen to be the correct conditions of justification.

I will focus here on weak accessibility internalism, since strong accessibility internalism raises complex issues we cannot address here. For brevity of reference, I will refer to weak accessibility internalism simply as "accessibility internalism." Accessibility internalism is a powerful constraint on justification, strong enough to exclude many proposed conditions of justification. To take one example, reliabilism clearly runs afoul of the constraint, since a subject cannot tell by reflection alone whether the process which forms a belief is reliable. Do *any* conditions actually satisfy it? At first blush, it would appear that some conditions do satisfy it. Here is a representative example (adapted from Chisholm) of the sort of condition that accessibility internalists have assumed conforms to accessibility internalism:

If a subject *S* is appeared to redly (to use Chisholm's language), then *S* is justified in believing that *S* is appeared to redly.

An accessibility internalist who subscribes to this condition assumes that a subject *S* can always tell by reflection alone whether *S* is appeared to redly. (Accessibility internalists typically assume as well that what justifies a subject *S* in believing that *S* is appeared to redly–namely, *S*'s being appeared to redly–also justifies *S* in believing that *S* is justified in believing that *S* is appeared to redly, so that *S* can always tell by reflection alone that *S* is justified in the second-order belief that *S* is justified in believing that *S* is appeared to redly.)

Let us consider first the implication of accessibility internalism for the practical endeavor of revising beliefs in light of our assessment of which beliefs (one's own or those of others) are justified. Accessibility internalism clearly entails that cognitive science is not needed and will not be helpful in this endeavor. For on accessibility internalism, a subject need only reflect to tell whether a belief satisfies the conditions of justification. This clearly entails that cognitive science cannot improve one's assessment of whether one's own beliefs satisfy the conditions of justification for purposes of *self*-regulation. The same arguably holds for one's assessment of whether the beliefs of *others* satisfy the conditions of justification. This is because we can expect an evaluator of the beliefs of others to be clued in to the conditions accessible to those others. One can generally tell by visually inspecting a scene whether others are appeared to redly. Admittedly, one cannot tell this by reflection alone; one needs various kinds of empirical information to do so, but the accuracy of one's judgment does not depend on sophisticated empirical science. Over time, people acquire a knowledge of correlations between types of external conditions and types of their own introspectible states, and they can then employ this knowledge to infer from what they observe conclusions about whether others are in states of these types. The introspectibility of accessible conditions affords people the capacity to judge accurately whether others are in like conditions. So, on accessibility internalism, cognitive science is unnecessary and unhelpful for the practical endeavor of revising beliefs.

Despite this, accessibility internalism does *not* rule out help from cognitive science in the theoretical endeavor of gauging our cognitive achievements. For cognitive science could (and probably already does) bear helpfully on which conditions conform to accessibility internalism itself. And this means that, on accessibility internalism, cognitive science could contribute to the theoretical endeavor in something like the way it does on Goldman's reliabilism. It is an empirical question, for example, whether the condition "I am appeared to redly" is one a subject can always tell to obtain by reflection alone, and cognitive science could be needed to answer this question.

We can see this by pursuing a question raised by Jerry Fodor. Fodor has questioned whether the way things look to one is always introspectively accessible (1990a 249-250). He raises the question by appeal to the example of the Müller-Lyer illusion, in which two lines of the same length look to be of different length when they are terminated with arrowheads pointing in opposite directions, on one line the arrowheads pointing outward and on the other pointing inward (more on this example later). The textbook psychological explanation of this illusion is that the arrowheads cause the perceptual apparatus to read the figure as having depth, the lines being at different distances from the viewer. Since the retinal projections of the lines are the same length, perception compensates by interpreting one line as longer than the other. If we accept this explanation, and we also think, as Fodor does, that the way things look is simply (or is simply described by) a hypothesis produced by the perceptual apparatus, then we must say that certain aspects of looks–in this case, depth of field–are inaccessible to the subject by introspection alone: the subject cannot always tell by reflection alone whether he or she is appeared to "deeply." My point here is not that this psychological explanation of the Müller-Lyer illusion is correct, nor that Fodor's account of looks is correct, though I think a case could be made for both of these claims. My point does not even depend on the assumption that either of these claims is correct. Rather, it is that neither the textbook psychological explanation nor Fodor's account of looks can be ruled out a priori. And since these have implications for the question which appearances are accessible by reflection alone, it would seem that this question is after all an empirical one amenable to scientific investigation. It turns out to be a scientific, empirical question whether depth looks are accessible by reflection alone. And, since we cannot in advance tie the hands of psychologists, it would seem likewise to be an empirical question whether color looks, shape looks, and the rest are accessible by reflection alone. We must leave open the possibility of discovering that these looks are not introspectively accessible to us. And what goes for looks would seem to go double for other conditions that accessibility internalists take seriously as candidate conditions of justification (e.g., conditions of belief content and inference), since looks are the conditions most likely to be accessible by reflection alone. Fodor's example shows that we ought to employ cognitive science to check the accessibility of conditions. Indeed, it seems that cognitive science already gives us reason to doubt our naive judgments of these matters. Thus, cognitive science could (and probably already does) bear on which conditions conform to accessibility internalism. In this way, it could contribute to the business of formulating conditions of justification guided by accessibility internalism. All this is so, despite the fact that, on accessibility internalism, cognitive science does not helpfully bear on which particular beliefs are justified for practical purposes of belief regulation.

I have argued that cognitive science is relevant to the question which conditions of justification satisfy accessibility internalism and thus which conditions of justification are correct. But this conclusion would seem to entail in turn that cognitive science is relevant to epistemology in another way: it is relevant to the plausibility of accessibility internalism itself.[26] For it would seem to show that accessibility internalism cannot itself be selected by narrow reflective equilibrium method; it must be selected by *wide method*–a method that employs cognitive science to select the correct conditions of justification (see subsection 4.1). If, as I have argued, cognitive science is relevant to judging whether candidate conditions of justification conform to accessibility internalism, then it is also relevant to judging the plausibility of accessibility internalism itself. For cognitive science could lead to the conclusion that *no* candidate conditions conform to accessibility internalism. This is more than an abstract possibility. Questions like Fodor's open the possibility that there are simply *no* conditions we can always tell to obtain by reflection alone. But if it turned out that there were no such conditions, then accessibility internalism would itself be mistaken. The situation here is not like the one concerning conditions of justification on a narrow reflective equilibrium theory. If, in light of cognitive science, a condition of justification were shown to entail skepticism, this would *not*, on narrow method, show that the condition is implausible. But if, in light of our best information (including cognitive science), accessibility internalism were shown to preclude there being any conditions of justification at all, this *would* tell against accessibility internalism. It is hard to credit the idea that it could turn out, not merely that there are no justified beliefs, but that there are not even any conditions of justification. After all, we distinguish justified from unjustified beliefs with a high degree of consistency and consensus. Certainly it could turn out that we are mistaken in our judgments, but it is hard to see how it could turn out that there are no conditions of justification at all, not even unsatisfied conditions of justification. For this reason, cognitive science could overturn weak accessibility internalism itself.[27]

To recap, I have argued here that, though on weak accessibility internalism cognitive science can give little aid in assessing for practical purposes whether beliefs are justified, it may help in the theoretical endeavor of selecting the correct conditions of justification. Indeed, empirical findings already on the table may threaten some conditions favored by proponents of accessibility internalism. In addition, cognitive science is relevant to the plausibility of accessibility internalism itself.

3.4.2 The Objection from the General Accuracy of Naive Evaluations of Justification

Accessibility internalism, even of the weak variety, is an implausibly strong constraint on justification. I have discussed it, not because I accept it, but because many epistemologists do. There is, however, a weaker, more plausible *empirical* conjecture that also rules out the need for cognitive science, at least for the practical purpose of belief-revision, but probably for the theoretical endeavor of assessing justification as well. I have in mind the empirical conjecture that our naive judgments of justification are *generally accurate*. It is plausible that people generally succeed in assessing which beliefs are justified. This seems extremely plausible in the case of assessment for purposes of belief revision, but it also seems plausible in the case of theoretical assessment. Indeed, it seems that people have been making accurate judgments of justification for tens of thousands of years, and of course they have done so without ever having made any systematic review of the reliability of our native processes in light of cognitive science, as Goldman's reliabilism demands. No doubt, cognitive science could *confirm* that our judgments of justification are generally accurate. Perhaps it could *improve* the accuracy of our judgments. But we have empirical grounds now for doubting whether it would *disconfirm* the general accuracy of our judgments. This empirical conjecture of course rules out any theory of justification according to which we must rely on cognitive science to assess justification, as Goldman's reliabilism does.

I hasten to observe, however, that this empirical conjecture does not rule out all possible versions of reliabilism. In fact, I endorsed the empirical conjecture and developed a version of reliabilism compatible with it in Schmitt (1992). The empirical conjecture puts two important constraints on any compatible version of reliabilism. First, since naive judgments of justification do not involve a systematic review of the reliability of processes in tandem, justified belief cannot depend on the reliability of many processes working in tandem. Justification turns not on the holistic truth-ratio of a J-rule system, as in Goldman's theory, but on the atomistic reliability of individual processes in isolation.[28] Second, and more pertinent to our concerns in this article, justification turns, not on belief-forming processes that can only be identified by cognitive science, but on processes we can identify on the basis of ordinary experience–"folk processes."

Of course, the case here against Goldman's version of reliabilism rests on the plausibility of the empirical conjecture. But is it true that naive judgments of justification are generally accurate? The argument for saying so is pragmatic. The social institution of epistemic evaluation would appear to be largely successful in its chief cognitive function–fostering justified belief by enabling people to judge which of their own and others' beliefs are justified, and on this basis to retract their unjustified beliefs and acquire justified beliefs. At any rate, people behave as if they tacitly agreed that evaluation is largely successful in this function. For people go on evaluating justification and assign the evaluation of justification a high priority in daily life as if they believed that their evaluations succeed in fostering justified belief. Of course the tacit assumption of success *could* be mistaken. But it seems that there is some reason to accept this assumption (at least, if reliabilism is true). For epistemic evaluation fosters certain beliefs, and these beliefs lead to successful actions. Yet, at the same time, there is some reason to believe that beliefs that lead to successful actions tend to be true. So, assuming that reliabilism is true, there is some reason to believe that the beliefs fostered by the institution of epistemic evaluation lead to justified beliefs. Thus, there is some reason to believe that the institution succeeds in its function. But now, given that the institution succeeds in its function of fostering justified beliefs, it is likely that, generally, people accurately judge which beliefs are justified. It is unlikely that the institution of evaluation would foster justified beliefs if people were not generally correct in their judgments of which beliefs are justified. At least this is so if: a judgment that a belief is unjustified usually leads to retracting that belief; a judgment that a belief is justified usually leads to adopting that belief; and deliberate but erroneous retraction of a belief as unjustified, or deliberate but erroneous adoption of a belief as justified, does not typically increase the stock of justified beliefs. We may infer from all this that our naive judgments of justification are generally accurate. We reach, by this pragmatic argument, the empirical conjecture that naive judgments of justification are generally accurate. And this conjecture in turn favors atomistic folk process reliabilism over Goldman's holistic reliabilism.

It is worth noting that Goldman (1993a) has recently revised his reliabilism to make it compatible with the conjecture that naive judgments of justification are generally accurate. According to this new reliabilism, a belief is justified when it is "obtained through the exercise of intellectual virtues" (e.g., by sight, hearing, memory, or reasoning in approved ways) and unjustified when it is obtained through vices (e.g., guesswork, wishful thinking, and ignoring contrary evidence) (1993a, 97-98). The virtues and vices are, in turn, selected by their reliability and unreliability. The theory will allow that naive judgments of justification are generally accurate if people are accurate in judging whether beliefs result from virtues and vices and also accurate in their judgments of the reliability and unreliability of virtues and vices. If virtues and vices are understood in much the way belief-forming processes are on process reliabilism–and this is the way Goldman seems to understand them–then the above pragmatic argument will support the conjecture that naive judgments of justification are generally accurate, on Goldman's new virtue reliabilism.

However, even if, for the reasons given, we accept the empirical conjecture that people are generally accurate in their judgments of justification, and we also accept my atomistic folk process reliabilism or Goldman's virtue reliabilism, and finally we accept the limitations on the use of cognitive science these views imply, there remain three possible significant roles for cognitive science regarding judgments of justification. We have already mentioned two of them: cognitive science could confirm or improve the accuracy of our judgments. A more important role (or, more exactly, metarole) concerns the pragmatic argument for general accuracy itself. It is, after all, an empirical argument. Thus, both the premises of the argument and its conclusion, the empirical conjecture of general accuracy, however reasonable these may be at the present time, could be overturned in due course by cognitive science.[29] As long as we accept the empirical conjecture, empirical findings like those of Tversky and Kahneman could not lead us to assess our probability judgments as unjustified; but those findings could perhaps cast doubt on the empirical conjecture itself and in this way free us to accept the relevance of cognitive science to the assessment of justification.

In this connection, let me observe that there are three areas of research in cognitive science that specifically address the accuracy of our judgments of reliability and thus bear directly on the empirical conjecture, on a reliabilist account of justification: research on our ability to introspect psychological processes, on metacognition, and on our judgments of the calibration of our processes. The first and last of these raise serious doubts about the accuracy of naive reliability judgments. Since I have reviewed them in some detail in Schmitt (1992 ch. 8), I will report only the upshot here. (For relevant work on metacognition, see Koriat 1994, Miner and Reder 1994, and Schwartz and Metcalfe 1994.)

It is natural to suppose that judgments of reliability rely on introspecting belief-forming processes. We judge the reliability of a process by checking the truth-values of the beliefs that process yields, but we know which beliefs a process yields on various occasions only by introspecting our own processes and relying on others' reports of similar introspections. This is worrisome for the accuracy of our reliability judgments because our ability to introspect our processes has come under attack in recent psychology. First, much processing is now regarded as unconscious and therefore inaccessible to introspection. Second, introspection is itself a fallible process susceptible to interference. Third, some psychologists have claimed that our apparently introspective judgments of processes are neither accurate nor in the end genuinely introspective. In a well known paper, Richard Nisbett and Timothy Wilson (1977) argue that we are often inaccurate in our judgments about cognitive processes and indeed cannot even be said to introspect them–we make causal hypotheses about them instead. While subsequent work on introspection (Ericsson and Simon 1983) has generally been more upbeat about its general accuracy than Nisbett and Wilson were, some psychologists remain pessimistic about introspection (Gopnik 1993). It is not easy, however, to draw epistemological conclusions from the empirical claim that our apparently introspective judgments of processes are inaccurate, even waiving doubts about the correctness of the claim. For one thing, psychologists have been primarily concerned with verbal reports of processes, rather than judgments of the occurrence of processes. Yet our judgments of justification for purposes of the self-regulation of belief would appear to be largely automatic and unconscious, and for this reason they may not be accurately reported even if they themselves are accurate. More importantly, accurate evaluations need not depend on introspecting our processes; they could depend instead on accurate nonintrospective causal hypotheses designed to explain our introspective observations of correlations between types of mental states. Finally, psychologists have studied our judgments of processes primarily in the case of unusual processes in strained experimental conditions; we cannot readily infer from any negative findings about introspection in these circumstances that our beliefs about common processes in everyday circumstances are inaccurate.

Turning now to the literature on calibration, its basic, and unflattering, finding is that people are poorly "calibrated." In particular, people are overconfident in their judgments; at least they are so in judgments that result from problem-solving of moderate or extreme difficulty. That is, people have a higher degree of confidence in their judgments in many instances than is warranted by the frequency of their true judgments in those instances (Fischhoff, Slovic, and Lichtenstein 1977; Lichtenstein and Fischhoff 1977, 1980). Moreover, attempts at improving calibration have had mixed results (Adams and Adams 1958, 1961; Lichtenstein and Fischhoff 1980; Lichtenstein, Fischhoff, and Phillips 1982). Do these findings entail that people overestimate the reliability of their processes and thus, on reliabilism, inaccurately assess justification? There are several obstacles to this inference. First, psychologists have been concerned with our calibration on *topics.* But topics are individuated in a way that may well cut across the epistemically relevant individuation of belief-forming processes. An assessment of justification must judge the calibration or reliability of belief-forming processes; but we cannot readily infer the inaccuracy of judgments of the reliability of belief-forming processes from a subject's poor calibration on topics. Second, psychologists measure the accuracy of people's *reports* of confidence, not of judgments of reliability that might enter automatically and unconsciously into the self-regulation of belief. Third, if overconfidence is systematic and (roughly) a monotonic function of reliability, then *ordinal* comparisons of reliability will still be accurate even where judgments of the degree of reliability are not, and people will be able to make accurate judgments of qualitative reliability equivalent to those they would make if they were not overconfident, merely by setting a higher lower bound on the degrees of reliability they take to be necessary for justification. For these reasons, the findings on calibration, depressing though they may be, do not yet rule out the general accuracy of people's assessments of justification. We do not yet have substantial empirical evidence against the naive empirical conjecture that people are generally accurate in their assessments of justification.

These empirical findings illustrate the ways in which cognitive science could bear on the empirical conjecture and thus on my objection to the need for cognitive science in assessing justification.

4 EMPIRICAL COGNITIVE SCIENCE AND THE CONDITIONS OF JUSTIFICATION

I have so far observed that, on certain accounts of justification (Goldman's and Kornblith's reliabilism, psychologism, proper function theory), cognitive science is needed to assess justification and to provide a detailed specification of the sorts of beliefs that are justified. I have, however, endorsed one ground for doubting whether cognitive science could be needed in the way these accounts require, and I have pointed out that there remains at least one other version of reliabilism that is not committed to this particular reliance on cognitive science. At the same time, I have hastened to add that even if this ground for doubting the relevance of cognitive science in assessing justification is accepted, there remain other significant ways in which cognitive science could bear on epistemology–e.g., by providing a check on this very ground for doubt.

I would like now to turn to the difficult and more profound question whether empirical cognitive science could help us with the theoretical endeavor, not of assessing justification given a particular condition of justification, but with the different theoretical endeavor of *selecting* the correct conditions of justification.

In fact, cognitive science has already influenced epistemology in this theoretical endeavor by calling attention to the role of certain concepts in our everyday thinking about knowledge, concepts which may then form the focal point of an account of knowledge. This influence has spawned some novel accounts of knowledge–notably, Fred Dretske's (1981) information-theoretic account of knowledge, Millikan's (1984) proper function theory, and Alvin Goldman's (1986) reliabilism.[30] But of course this is not *reliance on* or *support from* the findings of cognitive science but merely *inspiration from* cognitive science in the business of formulating conditions of knowledge and justification. I wish to consider whether epistemologists could or should rely on cognitive science to support the choice of conditions of justification.

Clearly, if the enterprise of selecting the correct conditions of justification is a purely a priori business, then it cannot rely on information from empirical cognitive science. Moreover, even an empirical enterprise could bar reliance on cognitive science, if it is restricted to naive empirical information, as in narrow reflective equilibrium method. Information from cognitive science will be admissible only in wide reflective equilibrium method. (Of course, even if a method does admit reliance on cognitive science, it is yet another question, to which I will return in the next subsection, whether cognitive science actually contributes anything to the selection process.) Let us, therefore, turn to the choice between narrow and wide reflective equilibrium methods.

4.1 Narrow versus Wide Reflective Equilibrium Methods

Narrow and wide reflective equilibrium methods are the most popular methods for selecting the correct conditions of justification. (We need not assume that either method necessarily *succeeds* in selecting the correct or even justified conditions of justification; nor need we assume that these methods themselves justify any beliefs.[31]) Both methods employ *reflective equilibrium* to select the conditions of justification (Goodman 1965, Rawls 1971). According to reflective equilibrium method, we select the candidate set of conditions that best organizes and explains our general intuitions about justification and our intuitions about the justification of particular beliefs. The two methods differ, however, in just which intuitions they put in reflective equilibrium. Wide reflective equilibrium method puts in reflective equilibrium not only our naive common sense intuitions about justification but also the findings of empirical science. Equivalently, or nearly enough, wide reflective equilibrium method selects the conditions of justification that best explain the intuitions we would have after being fully informed of empirical science. *Narrow* reflective equilibrium method, by contrast, puts only our naive intuitions in reflective equilibrium. It selects the conditions that best explain our naive intuitions. Naive intuitions are, nearly enough, the intuitions we would have if we were not informed of empirical science. Narrow method is most plausibly directed toward a priori necessary conditions of justification, while wide method would seem best for empirical contingent conditions.

How do the two methods differ in the way they select conditions of justification? Consider first narrow reflective equilibrium method. We have very many naive intuitions that ascribe justification to particular beliefs; we also have the general naive intuition that very many, perhaps most of our beliefs are justified. These intuitions are so numerous and forceful that we may reasonably expect narrow reflective equilibrium method to reject any conditions of justification that make most of our beliefs unjustified (when the justification of beliefs is assessed by naively figuring whether the given conditions are satisfied by the beliefs). Despite this, narrow reflective equilibrium method does not rule out the possibility that, on the selected conditions of justification, our beliefs will turn out to be unjustified in light of empirical science. Such a skeptical outcome would ensue if the intuitions we would have in light of empirical science were sufficiently less sanguine than our naive intuitions. A skeptical outcome of this sort is not, on narrow reflective equilibrium method, a mark against the selected conditions of justification. In short, on narrow reflective equilibrium method, cognitive science is irrelevant to the selection of the conditions of justification, but it could still be relevant to the question whether the selected conditions are satisfied. Whether it is relevant turns entirely on the details of those conditions: are they conditions whose satisfaction we had best assess in light of cognitive science?

Let it be noted that narrow reflective equilibrium method does not forbid reliance on empirical information, as long as it is naive empirical information. Such information is represented in naive empirically-based intuitions, which are given no privilege over other intuitions in the course of putting intuitions in reflective equilibrium. That is, the fact that these are empirically-based intuitions does not enhance their status in narrow reflective equilibrium method, except perhaps to the extent that people give greater intuitive weight to intuitions they believe to be empirically based. Let it also be noted that narrow reflective equilibrium method leaves open an oblique way in which cognitive science could be relevant to the selection process: it could help us figure out just what our naive intuitions are. As narrow reflective equilibrium method is currently conducted, we do not use the results of cognitive science to do this; we simply elicit our intuitions by reflection. But there is nothing in the idea of a narrow reflective equilibrium method that bars help from cognitive science here. What a narrow method rules out is using cognitive science to *correct* our intuitions in the selection process.[32]

On wide reflective equilibrium method, by contrast, all available information is admitted in selecting the conditions of justification. We put all available information, both scientific empirical information and naive opinions, in reflective equilibrium. (Again, it is a further question whether cognitive science actually contributes to the process; equivalently, it is a further question whether the outcome of wide reflective equilibrium is any different from that of narrow reflective equilibrium.) On wide reflective equilibrium method, unlike narrow method, cognitive science enters before we select the conditions of justification. Since naive intuition also goes into reflective equilibrium, and naive intuition tells us that our beliefs are justified, it *is* a mark (though perhaps not a decisive one) against a set of conditions if, in light of empirical science, the conditions entail that we are not justified in our beliefs. In short, on wide reflective equilibrium method, we do not separate the business of selecting the conditions of justification from the business of judging in light of empirical science whether the selected conditions are satisfied. Once we have registered the empirical findings in our choice of the conditions, there is no further process of employing those findings to judge whether the conditions are satisfied; we have already decided whether the conditions are satisfied in the very business of choosing the conditions.

Which method, narrow or wide, is the correct one for selecting the conditions of justification? In my view, the two methods are best viewed not as competitors but as complementary methods aimed at different tasks. Narrow reflective equilibrium method would, with one important qualification to be discussed below (subsection 4.2), seem to be the more appropriate one for selecting conditions of justification designed to define our *common concept* of justification. For it does not seem that our common concept of justification could be shaped by opinions most people do not possess, and no one possessed until recently–by scientifically informed opinions. It must rather be shaped by our common opinions; these are the opinions that must be organized and explained by conditions that define our common concept. Narrow reflective equilibrium method could also be appropriate for selecting *nonconceptual conditions* (whether necessary or contingent) that identify the property of justification, if that property is constructed by us in the sense that it has its features merely in virtue of the way people think about justification (see Craig 1990 for a weaker version of constructivism). Wide reflective equilibrium, by contrast, would seem to be most appropriate for contingent property identification, if the property is a natural one in the sense that it is not constructed by us (see Millikan 1984 for a clear example of naturalism). The method may select among conditions with empirical content (as, for example, in Goldman's third stage selection of conditions of justification).

In light of all this, it is not surprising to find that those who oppose the relevance of empirical science to assessing which beliefs are justified tend to prefer conceptual analysis, narrow method, and a constructivist conception of justification. Proponents of the relevance of empirical science, by contrast, come in various stripes–conceptual analysis and narrow method (Goldman 1986) or property identification and wide method (Kornblith 1993). But most of them do prefer naturalism to constructivism.

If we are not forced to choose once and for all between narrow and wide reflective equilibrium, then we are not deprived once and for all of cognitive science as a resource for selecting the conditions of justification. But the question remains whether and how cognitive science could be helpful in the task of selection.

4.2 The Format of Epistemic Concepts

Epistemologists who use narrow reflective equilibrium method to characterize the concept of justification tend to assume that their task is to find necessary and sufficient conditions for justification. But this assumption–that the concept of justification has the format of necessary and sufficient conditions–is open to question. Cognitive science could enter the selection of conceptual conditions of justification at the very outset by helping us select the format of the concept of justification. (This would of course entail an exception to the narrow reflective equilibrium method stricture against admitting scientific empirical information. It would not, however, have to take us as far as full-fledged wide reflective equilibrium method. The reliance on empirical information could be restricted to the choice of format of the concept of justification.)

Psychologists have argued on empirical grounds for diverse formats for concepts. Concepts have been taken to be sets of features, prototypes (i.e., sets of averages of features across examples), or exemplars (i.e., representations of paradigmatic examples) (Rosch and Lloyd 1978, Smith and Medin 1981). The prototype and exemplar formats have been widely favored in the last two decades. Alvin Goldman has proposed making use of the exemplar format in analyzing the concept of knowledge. His virtue reliabilism, already cited in subsection 3.5.2, treats the concept of knowledge as an exemplar consisting of various epistemic virtues, against which particular examples of beliefs are judged by measuring their similarity to virtuous belief-formation.

Of course, philosophical debate over the relative merits of feature and prototype formats for concepts goes back to Plato. The traditional philosophical dispute has turned on the *necessary* nature of concepts or the analysis of our *concept* of "concept." But narrow reflective equilibrium method has never succeeded in resolving this issue. Moreover, attempts to provide necessary and sufficient conditions for concepts have almost all ended in clear failure, raising the question whether our concepts are really prototypes rather than necessary conditions. If the dispute cannot be settled naively, then we must make an exception to the narrow reflective equilibrium method stricture against reliance on scientific empirical information.

4.3 Foundationalism and Epistemological Pluralism

A number of philosophers have suggested that empirical cognitive science undermines the plausibility of the most venerable account of the inferential structure of justification, *foundationalism*. Of course, if this is so, then foundationalism has empirical content.

Foundationalist theories of knowledge dominated epistemology from its inception in Plato until well into the twentieth century. There are, as one would expect from its long history, many versions of foundationalism (Chisholm 1966, Pollock 1974, Audi 1988, A. H. Goldman 1988; see BonJour 1985 chs 2-4 for an illuminating review). Although these versions differ in important respects, we can make do with the following weak characterization shared by most of them:

> There is a class of justified beliefs (called *foundational* or *basic* beliefs) that derive some of their justification (or have some initial credibility) independently of their relations to other beliefs–i.e., they are justified without being justified wholly on the basis of other beliefs (or wholly in virtue of justifying relations to other beliefs).
>
> All other justified beliefs (called *nonbasic* beliefs) are justified on the basis ultimately of basic beliefs.

In the rationalist tradition, basic beliefs have been taken to be products of a special faculty of intuition, while in the empiricist tradition, they have been taken to be perceptual beliefs, typically beliefs about the appearances of objects ("I am appeared to redly"). In both traditions, nonbasic beliefs have included the products of inductive and deductive inference, as well as beliefs about the physical world inferred from appearances (e.g., the belief that this object is red).

Some philosophers have suggested, however, that foundationalism should be rejected on empirical grounds. "New Look" psychologists (Gregory 1970) have offered empirical reasons for thinking that perceptual beliefs about appearances are *theory-laden*. In a parallel development, post-positivist philosophers of science like N. R. Hanson (1961), Thomas Kuhn (1962), and Paul Churchland (1979) have claimed that scientific observations are also theory-laden. Now, the theory-ladenness of scientific observation clearly does *not* tell against foundationalism–at least, not as we have defined it. For scientific observational beliefs–e.g., beliefs about the features of bodies visually perceived with a telescope–are uncontroversially nonbasic: they clearly rest on theories of the operation of an experimental apparatus. Hence, showing that scientific observational beliefs are theory-laden does not show that any beliefs a foundationalist would take to be basic are really nonbasic. However, the New Look thesis that perceptual beliefs result from top-down perceptual processing and are therefore theory-laden has been thought to contradict foundationalism. Let us ask, then, whether this empirical thesis really does contradict foundationalism.[33]

Let us consider first the case for the theory-ladenness of perceptual beliefs and return later to its implications for foundationalism. New Look psychology proposes that there is no principled distinction between *perception* and *cognition*: "perception involves a kind of problem-solving–a kind of intelligence" (Gregory 1970, 30). In perception, a subject infers distal causes from the character of its sensory states– "betting on the most probable interpretations of sensory data, in terms of the world of objects" (Gregory 1970, 29). Perception is a kind of inference analogous to intellectual theory choice on the basis of underdetermining data. Sensory data or stimuli massively underdetermine the perceptual beliefs about objects themselves (about how objects appear), and background knowledge about objects must be employed to resolve the underdetermination. New Look psychology claims that we rely on beliefs about objects to arrive at perceptual beliefs. Thus, perceptual beliefs themselves are theory-laden.

Jerry Fodor (1990a) has challenged this argument for the claim that perceptual beliefs are theory-laden. He does not wish to deny that *any* perceptual beliefs are theory-laden; perceptual beliefs about the perceivable properties of objects (e.g., their being red or square) may be theory-laden. But Fodor maintains that there is a sort of perceptual belief (or, at any rate, a sort of hypothesis) that is *not* theory-laden. Beliefs of this latter sort are perceptual beliefs (or perceptual hypotheses) about *appearances* (an object's looking red)–"observations," as he calls them (though don't confuse them with the scientific observations we referred to two paragraphs above). Fodor grants the New Look proposal that perception is a sort of inference from sensory data to observations. He also grants that the sensory data underdetermine the observations, so that we must employ assumptions to resolve the underdetermination. However, Fodor denies that the assumptions we rely on here are *beliefs* about the perceivable properties of objects or about appearances. He proposes that the inference that produces observations is *informationally encapsulated*: not just any beliefs about objects are available as premises in the inference (Fodor 1983 pt 3). The premises are in general not beliefs at all but representations of generalizations about size or shape constancy in the way objects appear. These generalizations are inaccessible to consciousness and unavailable for general problem-solving.

Fodor supports his thesis of informational encapsulation by appeal to perceptual illusions like the Müller-Lyer figure, in which equal lines appear to differ in length because bounded by arrowheads pointing in different directions.[34] New Look psychologists cite illusions like the Müller-Lyer figure as examples of how background generalizations about object constancy in appearances enter into perception and influence observation. But Fodor points out that in the Müller-Lyer illusion one background belief clearly does not influence observation: our belief that the lines in the figure are of equal length. We know by measurement that the lines are the same length, yet they still appear to us unequal in length, and, no matter how hard we try, we are unable to bring it about that the lines appear to be of equal length. Fodor concludes that the New Look proposal that perception is a kind of inference does not imply the theory-ladenness of observation–at least, not if theory-ladenness entails that all beliefs relevant to the perceived features of objects are available for observation. The perceptual inference that produces observations does not have access to all relevant beliefs.

Fodor suggests that we think of perception as consisting of two parts: observation and perceptual belief-formation. In observation, we infer from sensory data to a hypothesis about the appearances (e.g., "The lines in the Müller-Lyer figure appear of unequal length"). This is a hypothesis about the distal sources of stimulation, couched in a restricted observational vocabulary–"what you would believe about the appearances if you were going just on the appearances." The inference involved in observation is a modular process, and its premises include generalizations about constancies in the way objects appear. Perceptual belief-formation is also an inference, but here we take as a premise the hypothesis of observation, together with any relevant background beliefs that occur to us, and infer a belief about objects–about the distal sources of stimulation, couched in an unrestricted vocabulary (e.g., if we are innocent of the Müller-Lyer illusion, we infer "The lines in the Müller-Lyer figure are of unequal length").

I find Fodor's argument for the modularity of perception plausible enough, but defending it would take us too far afield (see Churchland 1988 for objections to Fodor, and Fodor's 1990b reply). I propose instead to address the question what bearing the issue of theory-ladenness has on foundationalism. Fodor does not explicitly draw from his modularity thesis any moral for foundationalism, but he is concerned to criticize a related view–what I will call *perceptual pluralism.* I will first discuss the implications of theory-ladenness and modularity for foundationalism and turn subsequently to perceptual pluralism.

It is natural to think that the theory-ladenness of perception, as conceived by New Look psychology, is inconsistent with foundationalism. It is equally natural to think that Fodor's modularity thesis is needed to defend foundationalism from the thesis of theory-ladenness. I wish to cast doubt on these natural thoughts. Suppose that we grant for the moment the New Look thesis that perceptual beliefs are theory-laden because they result from unencapsulated inferences taking as premises background generalizations concerning object constancy. It is natural to think that theory-ladenness rules out foundationalism, by the following line of reasoning. According to theory-ladenness, perceptual beliefs are inferences from sensations, together with premises that include background generalizations. So perceptual beliefs must be justified on the basis of these background generalizations, and they are therefore not basic beliefs. From this it follows that perceptual beliefs are not basic beliefs. And since perceptual beliefs are the best candidates for basic beliefs, it follows that foundationalism is false.

However, this line of reasoning is much too quick to be convincing. Note first a preliminary difficulty for the argument: the perceptual beliefs claimed to be nonbasic are not the perceptual beliefs foundationalists claim to be basic. The perceptual beliefs of New Look psychology are beliefs about the perceivable properties of objects ("This is red"), not beliefs about appearances. It is the latter beliefs that foundationalists claim to be basic. Perhaps, though, the New Look thesis can be revised in line with Fodor's distinction between observational hypotheses, which concern appearances, and perceptual beliefs, which concern the perceivable properties of objects. Then the New Look thesis would be that observational hypotheses (which are beliefs, for New Look psychology) are theory-laden. But even this thesis does not *by itself* contradict foundationalism. It does so only on the additional assumption that the inferences involved in forming observational beliefs are epistemically relevant inferences. That is the assumption that enables the contingent, empirical New Look thesis to bear on foundationalism–the assumption on which foundationalism has empirical content. But foundationalists are apt to reject this assumption.

To see why foundationalists tend to reject the assumption, consider a typical version of foundationalism: that perception delivers not only observational beliefs that represent how things look but before that experiences or sense impressions representing the same. (Note that in Fodor's etiology of perceptual belief, observational hypotheses stand in for sense impressions.) These sense impressions were always assumed *not* to depend on theoretical background beliefs for their formation. But there is nothing to stop the foundationalist from changing the causal story in light of New Look psychology and allowing that the sense impressions do depend causally on background generalizations. The foundationalist may accept this revision with equanimity because, according to typical foundationalism, the chain of *justification* starts only with the sense impressions. The dependence of the sense impressions on background beliefs is therefore irrelevant to the justificatory story. Basic beliefs are justified on the basis of sense impressions. But sense impressions are not in turn justified on the basis of anything else. To paraphrase Chisholm, if I am asked what justifies me in believing that this object is red, I will say, "I am appeared to redly." If I am asked what justifies me in believing that I am appeared to redly, I can only repeat the content of my belief that I am appeared to redly; I can only say "I *am* appeared to redly"–a description of my sense impression. But it makes no sense then to ask what justifies my sense impression. Sense impressions do not themselves have propositional content, and they cannot be doubted in any way analogous to the way in which beliefs can be doubted. Sense impressions are not the sorts of states to which the category of justification even applies. So the fact that background generalizations enter into the etiology of sense impressions does not entail that observational beliefs justified on the basis of these sense impressions are in turn justified on the basis of these background generalizations.[35]

Of course, if observational beliefs themselves resulted from reasoning taking both sense impressions and background generalizations as input, then it would be plausible enough that the justification of observational beliefs does depend on the justification of the background generalizations, and this would contradict foundationalism. But the thesis of theory-ladenness does not by itself entail that background generalizations enter the chain of reasoning that takes a sense impression as input. Clearly, the background generalizations that enter into the Müller-Lyer illusion are efficacious in producing the sense impression and not just the observational belief. Thus, the typical foundationalist can endorse an etiology of observational beliefs that is consistent with the theory-ladenness of perception. The upshot of all this is that theory-ladenness does not by itself contradict a typical foundationalism. Consequently, Fodor's modularity thesis is not needed to defend typical foundationalism from theory-ladenness.

But even though the modularity thesis is not needed for this purpose, it still has implications for a significant epistemological view–*epistemological pluralism.* In fact, one of Fodor's motivations for denying the theory-ladenness of perception is to undermine the support theory-ladenness provides for what we might call *perceptual pluralism.* Proponents of the theory-ladenness of perception took the thesis to lend credence to the psychological claim that people with radically different background generalizations–laypersons and scientists, for example–could perceive the world very differently despite having the same sensations (either, in one version of the view, the same complete history of sensations, or, in another version, the same sensations that would, on foundationalism, warrant their observational beliefs). They perceive the world differently because, in virtue of having different background generalizations, they have different perceptual beliefs–perhaps even different observational beliefs (about appearances). This perceptual pluralism in turn has a significant epistemological consequence. If one thinks of the relation between observational, perceptual, and theoretical beliefs as what supplies warrant for these beliefs, then the psychological claim that people with the same sensations may have very different beliefs makes plausible the epistemological claim that people with the same sensations could be *justified* in radically different beliefs–i.e., it makes plausible *epistemological* pluralism. The thesis of theory-ladenness supports perceptual pluralism, and that idea in turn supports epistemological pluralism.[36]

The modularity thesis, by contrast, is inconsistent with the above argument for perceptual pluralism. For it is inconsistent with the theory-ladenness of perception assumed by the argument. A major attraction of the modularity thesis over the thesis of theory-ladenness, for Fodor, is that the theory-neutrality of observation allows for observational consensus in science, while theory-ladenness does not. If scientists observe the same things, there is a better chance of explaining their agreement on the observations than if they do not observe the same things. We cannot pursue here the matter of explaining consensus in science. The pertinent point for our purposes is that the modularity thesis undermines the argument for the epistemological pluralist view that people can be justified in radically different beliefs, even though they have the same sensations. For that argument rests on perceptual pluralism, and, as we have seen, the modularity thesis is inconsistent with the argument for perceptual pluralism. Of course, empirical work bears here precisely because perceptual pluralism is an empirical thesis and so, for this reason, is epistemological pluralism. To what extent the modularity thesis undermines these views is a question for future empirical work revealing the extent of the informational encapsulation of perception.

We are left, then, with the conclusion that, although the psychological issue of theory-ladenness does not directly affect the plausibility of foundationalism, it does affect the plausibility of epistemological pluralism. Epistemological pluralism does have empirical content.

4.4 Normative Principles of Rational Belief and "Ought" implies "Can"

I have observed that cognitive science could affect the format of the definition of justification. I have resisted the claim that a certain empirical theory, New Look psychology, casts doubt on foundationalism (though I have not denied that some cognitive science finding might yet be relevant to foundationalism). I would like to turn now to an argument that normative principles of rational belief are vulnerable to criticism on the basis of scientific empirical findings, via the familiar constraint that "ought" implies "can."

One business of traditional epistemology is to propose normative principles governing epistemically rational belief. Epistemically rational belief is, by definition, belief we epistemically ought to hold: if it is epistemically rational for a subject *S* to believe a proposition *p*, then *S* epistemically ought to believe *p*. Yet this "ought" is plausibly governed by the constraint that "ought" implies "can": if *S* epistemically ought to believe *p*, then *S* is able to believe *p*.[37] The constraint that "ought" implies "can" underwrites criticism of proposed epistemic principles on empirical grounds, since "can" judgments are empirical judgments.[38] (This gives us reason to prefer wide over narrow reflective equilibrium method in formulating normative principles of rational belief.)

Let me illustrate these points with the requirement of total evidence (RTE): "The credence which it is rational to give to a statement at a given time must be determined by the degree of confirmation...which the statement possesses on the total evidence available at the time" (Hempel 1965, 64). The total evidence available to a subject at a time is the total evidence the subject possesses at the time, where a subject *S* possess evidence *E* at a time just in case *S* believes *E* at the time.

In opposition to RTE, Goldman offers this example:

> Melanie goes to the library one Sunday morning, expecting it to be open. Thus, she apparently has an (activated) belief in the proposition: "The library is open this morning" (= L). The question is whether it is rational of her to believe L. Now if the evidence "available" to Melanie includes all her evidence beliefs in long term memory, then the evidence includes two propositions that jointly entail the negation of L, viz., "Today is Sunday" and "On Sundays the library does not open until 1:00". Thus, if we adopt Carnap's and Hempel's rationality requirement for the application of inductive logic, we would be forced to say that Melanie's belief (high credence) in L is irrational. But this irrationality ascription, it seems to me, is questionable. Melanie's doxastic performance may be open to criticism, but "irrationality" does not seem to be the appropriate charge. (1989, 319)

Goldman provides one explicit ground for rejecting the claim that Melanie's belief in L is irrational: a charge of irrationality attributes a defect of reasoning; yet there is nothing wrong with Melanie's reasoning, only with her evidence retrieval. This strikes me as intuitively correct: even if Melanie *should* have retrieved the evidence, her belief does not count as *irrational,* only unjustified.

But there is another criticism implicit in Goldman's remarks. It is not clear that Melanie *could* under the circumstances have retrieved the evidence. Perhaps she could have done so if she had asked the right questions, but she did not think of these questions and perhaps could not have done so without thinking of many other things she had no reason to think about. At the very least, it would have been uneconomical for her to devote the effort needed to ask the right questions. And thus we do not think she should have retrieved the evidence. And if it is not the case that she should have retrieved the evidence, then there is no basis for the charge of irrational belief.[39] Hence, RTE runs afoul of the constraint that "ought" implies "can" (or at least, "economical") and must be rejected.[40] We have of course speculated here, albeit empirically, about what Melanie could have done in the imagined case. We can expect cognitive science to provide a firmer basis for judgments of feasibility and economy in real cases. But it is reasonable at this point to reject RTE on the grounds we have mentioned. Thus, findings from cognitive science could cast doubt on proposed normative principles of rational belief.

4.5 Idealizing versus Nonidealizing Approaches to Rational Belief

RTE fails because it overidealizes rational belief. It imposes a condition on rational belief in excess of human limitations. Proponents of RTE evidently pay little attention to human limitations when they endorse the principle. This raises an important methodological question: how should epistemologists take account of human limitations when they formulate normative principles of rational belief? Clearly, we cannot ignore human limitations (if "ought" implies "can"). There is, however, a way of saving something from RTE and similar overidealized principles (e.g., the principle: maximize the expected epistemic utility of belief). Instead of saying that rational belief must conform to RTE, say, rather, that rational belief must approximate the ideal of taking total available evidence into account in forming beliefs. More generally, the proposal is that there are certain epistemic ideals, which we may specify without attention to human limitations–such ideals as taking total available evidence into account and maximizing expected epistemic utility. We then take human limitations into account by defining a rational belief as one that approximates these ideals as closely as feasible (or, alternatively, as closely as economical). We may call this an *approximate idealizing approach* to defining rational belief. On this approach, a rational belief is one that falls short of the specified ideals only to an extent excusable by limitations of resources. The approximate idealizing approach is clearly consistent with the constraint that "ought" implies "can." Opposed to the approximate idealizing approach is a nonidealizing approach on which we take human limitations into account from the start, in specifying the ideals themselves. On both approaches, cognitive science is relevant to epistemology. But only on the nonidealizing approach is it relevant to selecting the conditions of justification. On the idealizing approach, by contrast, it is relevant only to the task of assessing justification. Thus, the nonidealizing approach makes cognitive science more deeply relevant to epistemology. We must ask, then, which approach is more plausible.

There is a persuasive objection to the approximate idealizing approach: it does not afford an explanation of the fact that contingent human limitations enter into the conditions of rational belief. Now, on any account of rational belief, the conditions of rational belief will be normative in the sense that they describe conditions that human beings do not *necessarily* or *always* satisfy. But the approximate idealizing approach goes beyond this minimal amount of idealization; it specifies the conditions of rational belief by adverting to ideals that human beings need not even be *able* to satisfy. What point could idealization of this sort have?

Jonathan Adler has tried to motivate the approximate idealizing approach by arguing that it is the only way to secure full *generality* for the conditions of rationality:

> The traditional epistemologist appeals to the facts that we are limited, that we are capable of technologies that greatly expand our cognitive resources, and that we share an interest in knowing with others. Compare the first "fact" to one typical of those cited by naturalists: we can hold no more than about seven distinct units in short-term memory. Knowers who provide an alternative to being finite, while facing anything remotely like our epistemological problematic, are hardly conceivable. But beings who have greater short-term memory capacities are readily envisaged, and if we augment our own meager abilities with technical aids, such beings are actual. The assumption I am making about epistemology is standard for theorizing: Basic principles should be formulated at the most abstract level consistent with not abandoning the problematic. (1989, 240).

Now, Adler unfairly caricatures the opposition to the approximate idealizing approach: no one proposes writing species-specific contingencies into the conditions of rational belief. Everyone recognizes that our concept and standards of rational belief clearly apply to possible subjects without our specific human limitations–genetically engineered humans, smart aliens, and perhaps machines. Opponents of the approximate idealizing approach only bring in species-specific contingencies to show that proposed conditions of rational belief are mistaken because, given our limitations, they cannot apply *to us*, since they violate constraints of feasibility and economy (such as "ought" implies "can").

Nevertheless, Adler offers here an interesting argument in favor of the approximate idealizing approach: it secures *general* coverage. The approach can apply to nonhuman cognizers. But is it true that it is the only, or even the best, way to secure general coverage? Regarding whether it is the *only* way, we may note that a nonidealizing approach can cover more than human cognition simply by *not* specifying that rational beings have *all* human cognitive limitations. A nonidealizing approach will no doubt write *some* human limitations into the conditions of rational belief (e.g., it might write into the conditions that evidence is a chief means to true belief–a condition that would not apply to some possible gods). But a nonidealizing approach can achieve general coverage by, for instance, not specifying human limitations of attention to evidence or of evidence-gathering. Thus, it is not necessary to achieve general coverage the way the approximate idealizing approach does, by abstracting from human limitations to a humanly impossible ideal and then excusing failure by allowing approximation to that ideal.

On the matter of whether the approximate idealizing approach is the *best* way to secure general coverage, the key question is whether a nonidealizing format can offer a principled reason for *not* writing certain human limitations into the conditions of rational belief (e.g., that working memory has a limit of seven chunks), and whether the approximate idealizing approach can offer a principled reason *for* writing certain human limitations into the conditions of rational belief. Nonidealizing and idealizing approaches differ in their inherent tendencies: a nonidealizing approach has an inherent tendency to specificity, while an idealizing approach has an inherent tendency to generality. Nevertheless, a nonidealizing approach can offer a principled reason for generalizing, albeit a pragmatic one–indeed, the very pragmatic reason Adler offers for the approximate idealizing approach. Since the limits of human cognition undergo constant change, and our view of them is in any case uncertain and subject to revision, we should not write these limitations into the conditions of rational belief. On a nonidealizing approach, we write the requirement of evidence into the conditions of rational belief because human beings satisfy epistemic goals by acquiring evidence, but we remain noncommittal about the details: we do not require that rational beliefs be formed in light of all the relevant possessed evidence. (Of course, on a nonidealizing approach, as on the approximate idealizing approach, we can peg how much of the total relevant possessed evidence the subject must rely on to the subject's abilities.) So a nonidealizing approach can balance pragmatic reasons for generalizing against its inherent tendency to specificity and in this way motivate generality.

Can the approximate idealizing approach offer an analogous principled reason for specificity that would countervail its inherent tendency to generality? Here I think the approach is at a disadvantage. An approximate idealist can give no good reason for identifying the ideal with, say, belief that results from total relevant possessed evidence. After all, gods might have means to true, informative, or explanatory beliefs without acquiring evidence. Adler would say that this would abandon "our problematic," but the question is what principled characterization of our problematic the approximate idealizing approach can offer. It appears that the nonidealist is right in claiming that rational belief is essentially tied to acquiring evidence because it is essentially tied to certain human limitations. An approximate idealist has no way to explain why these limitations enter into the conditions of rational belief. The inherent tendency to generality of the idealizing approach drives it all the way to an ideal consisting of the basic epistemic goals of true, informative, or explanatory belief. Of course, this does not force idealists to identify rational belief for human beings with true belief. Rather, it forces them to say that rational belief is what approximates true belief up to the subject's limitations–limitations which force the subject to rely on evidence. But that leads to the very counterintuitive consequence that gods could have rational belief without relying on evidence.

In short, the approximate idealizing approach affords no way to explain why the *general* conditions of rational belief conform in certain respects to *human* limitations. The conditions of rational belief are more deeply parochial than idealists can explain. A nonidealizing approach, by contrast, can explain why this is so, and it is therefore preferable to the approximate idealizing approach. And a nonidealizing approach assigns cognitive science a deeper role in epistemology–a role in selecting the normative principles of rational belief.

This completes our review of the bearing of empirical cognitive science on the tasks of assessing justification and selecting the conditions of justification. We turn now from empirical cognitive science to AI.

5 AI AND THE CONDITIONS OF RATIONAL BELIEF AND ACCEPTABILITY

The most fundamental philosophical question about AI is whether artificial intelligence–intelligent behavior in an electronic digital or analog computer–is so much as possible. The case against the possibility of artificial intelligence is multifarious. Classical AI is said to be impossible because it models inductive reasoning on classical logic. Yet it has proven difficult to develop an inductive logic. For example, the strength of an inductive argument (as measured, say, by the number of confirming instances in the premises) varies crucially with background information, and there is no obvious way to regiment the kinds of background information relevant here. One might try simply to store vast amounts of background information (Lenat, Prakash, and Shepherd 1986), but many doubt whether that will help, or whether it will make an interesting program even if it works. It is possible, too, that reasoning consists of a vast number of highly specific modules organized in an intricate hierarchy, and there is little prospect of discovering such modules without a long term neurophysiological investigation. These reflections may lead to pessimism about the possibility of artificial intelligence. But they should not lead to pessimism about the benefits AI research may bestow on epistemology. For epistemology does not have the ambition of creating artificial knowledge. The discovery that inductive strength depends on background information, while an obstacle to the success of AI, counts for epistemology as an intriguing discovery about knowledge, however far from regimenting the effect of background knowledge we may be, and however far from programming induction we may remain.

The viability of AI is a topic well beyond our scope here, and I will stick with the modest task of reporting a few epistemological efforts inspired by AI. Several epistemologists have recently turned to AI as a source of methods, ideas, and detailed implementation for epistemological theories, just as psychologists have done since the outset of the cognitive revolution in psychology.[41] For those wishing to pursue epistemology in a traditional manner, AI is indeed a more natural source of inspiration and information than empirical cognitive science, for three reasons.

First, much research in AI consists of formal a priori studies similar to and often continuous with logic, probability theory, and statistics, and the findings of these studies may be used in epistemology in much the way the formal findings of logic and probability theory have traditionally been used. The relevance of these findings to epistemology is no more problematic than that of results in logic, and thus the use of these findings in epistemology poses no *special* problem we need to address (with one exception noted below).

One type of AI research develops normatively correct axiomatic systems of reasoning (e.g., probabilistic and nonmonotonic reasoning) and explores their soundness and completeness with respect to plausible semantics. It does so with a keen eye to the computational tractability of these systems and the feasibility of their implementation on existing electronic digital computers. A good example of this research is Pearl (1989), which we will discuss below. A different type of AI research aims at programming computers to simulate or embody good reasoning. Often, the programs are designed by implementing the formal systems developed by the first type of AI research. Interestingly, however, programs are frequently evaluated, not by their approximation to these formal systems, but by the *reliability* of their judgments (Kyburg 1991, Kelly 1995). While AI researchers design such programs by trial and error, this work is no more empirical than, say, the trial and error discovery of proofs in mathematics. To be sure, Gilbert Harman (1973) and Alvin Goldman (1986) have argued, in ways we discussed in section 3.2, that logic, probability theory, and statistics are only *indirectly* or secondarily relevant to conditions of knowledge and justification. But there need be no *special* problem with using an AI finding in epistemology, beyond the familiar difficulties of using formal logic or statistics. I will, however, point out one difficulty with certain uses of sophisticated AI programs in subsection 5.2.

Second, AI researchers are often concerned to describe and implement epistemically rational reasoning or inference. They are already doing epistemology (or applied epistemology), and we should not be surprised if they have made discoveries immediately relevant to epistemology. (But as I said at the outset, limitations of space prevent me from discussing these kinds of discoveries.)

Third, AI research is inspired by intuitive reflection on how we reason and learn, and this is a source of inspiration more apt to yield conditions and strategies that meet *accessibility* constraints than is an empirical psychological study of the strategies human beings actually use. For these three reasons, it is natural for traditionally minded epistemologists to look to AI for epistemological inspiration.

AI can contribute to epistemology in at least six ways. (1) AI researchers have developed axiomatic systems for normative qualitative, probabilistic, and nonmonotonic reasoning that may be imported into conditions of rational belief. We will consider examples of these in section 6. (2) AI researchers have developed reasoning strategies in automated theorem-proving, probabilistic reasoning, and inductive learning that may help characterize correct reasoning. As we will see (subsection 5.1), John Pollock has used AI programming to develop deductive and nonmonotonic reasoning strategies. (3) AI researchers are much concerned with the computational costs and feasibility of reasoning strategies, and their results here could pertain to the costs and feasibility of human beings satisfying certain epistemic conditions and principles. This is an empirical issue, but as I argued in subsection 4.4, it is one that bears on epistemology via the "ought" implies "can" constraint. (4) AI computational models may bridge the gap between intuitive epistemic principles, whose implications for particular cases may be hard to judge, and our epistemic judgments of these cases. We will examine Paul Thagard's use of a connectionist AI model for this purpose (subsection 5.2). (5) AI programs for reasoning can serve epistemology in somewhat the way they serve psychology–by forcing precision and detail in theories and by making clear just what a theory predicts for examples. For this reason, they can reveal complex counterexamples to proposed accounts of reasoning that would escape armchair reflection. Finally, (6) AI research has the potential to correct human chauvinism in epistemology–the tendency, driven perhaps by the legitimate and unavoidable human parochialism of our epistemic concepts, to bind all possible rational reasoning too closely to actual human reasoning. AI could perhaps provide examples of possible computer programs that reason quite differently from human beings but nevertheless reason rationally. I will say more about this in discussing Pollock's work in subsection 5.1.

These are some of the benefits AI research promises epistemology. There are, however, corresponding pitfalls that deserve notice before we sample some work in the area. For one thing, it is easy, in adopting a computational model, inadvertently to interpret the parameters of the model in a way that commits one to implausible epistemic principles. The very advantage a computational model can provide over bare intuitive principles–namely, precise implications for cases–can cost the theory its intuitive plausibility. Second, AI researchers have been tempted to suppose that algorithms that are computationally tractable for computers are tractable for human beings. But the bearing of computational tractability for computers on human feasibility is currently terra incognita. Third, there is a temptation (the converse of advantage (6) above) to elevate a reasoning strategy that is intuitively rational for some possible cognizers (computers) to a necessary condition of rational reasoning for all possible cognizers. This temptation must be stoutly resisted. I will in due course cite examples of all three pitfalls.

5.1 Pollock on Rational Belief: The Case of Deductive Reasoning

In recent years, John Pollock (1987, 1989, 1990a, 1990b, 1991) has undertaken the ambitious project of implementing a general theory of rationality in a computer program. Pollock conceives of his computer implementation, OSCAR, not merely as modeling but as *exhibiting* rationality and thought, and his project has much to contribute to AI studies of automated reasoning. His primary interest, however, is to develop a detailed theory of rationality. According to Pollock, the best way to refine a theory of rationality is to model it and in particular to do so by implementing it in a computer program. The value of computer modeling is that it requires us to specify the conditions of rationality exactly, and it enables us to discover counterexamples to the theory that we would not have discovered by reflection alone.

We will not be able to review Pollock's whole project here. Instead I will focus on the part most accessible to a general reader, the deductive reasoning module. But a brief outline of the project is needed for comprehending any part of it, so I will report Pollock's view of the architecture of rationality as described in "OSCAR: A General Theory of Rationality" (1991).

Pollock holds that human belief-formation encompasses two kinds of belief-forming processes. One is *reasoning*, which is the target of rational evaluation. Reasoning, however, is mostly serial and therefore slow. For this reason, much cognitive processing must be accomplished by nonreasoning processes. The cognition needed for computing the trajectory of a baseball, for example, does not involve reasoning. It is accomplished by a *quick and inflexible (Q & I) module.* Q & I modules have the advantage of speed, but they purchase this advantage at the cost of inflexibility. They are quick because they employ built-in assumptions about the environment; but when these assumptions fail, the modules can be wildly inaccurate. (Up to this point, Pollock is in broad agreement with Fodor on modularity.) Q & I modules are not restricted to the cognition needed for motor skills but are also responsible, as evidence from psychology suggests, for ordinary inductive and probabilistic inference as well (e.g., Kahneman and Tversky's representativeness heuristic). Pollock grants that it is reasonable to rely on these processes. Indeed, we have no practical alternative to relying on them: reasoning in these domains is slow or even computationally unfeasible. Reasoning or intellection does, however, have the advantage of flexibility. Its role is "to deal with cases to which built-in Q & I modules do not apply" and "to monitor and override the output of Q & I modules as necessary" (1991, 192). An automated reasoner with human-like powers will need *both* Q & I modules *and* slow but flexible reasoning.

Turning now to reasoning, Pollock identifies rationality with correct reasoning and distinguishes *theoretical* reasoning, which involves belief updating, from *practical* reasoning, which encompasses all other aspects of rationality (including the rationality of desires, intentions, and emotions). Our interest here of course is in theoretical reasoning. Pollock insists, however, that "theoretical reasoning is often guided by practical reasoning about how best to reason" (1991, 193)–a point ignored by most epistemologists. Such practical reasoning "can affect the strategies we employ in specific circumstances, and we can discover new strategies that are effective" (1991, 193). Practical reasoning in turn relies on prior knowledge (e.g., estimates of the probabilties of an action having various outcomes), and this knowledge is the result of theoretical reasoning. Thus, according to Pollock, some theoretical reasoning must be prior to practical reasoning and not employ practical reasoning.[42]

Since the practical reasoning employed in theoretical reasoning is reasoning about how to reason, it requires introspective knowledge of the agent's own reasoning–of what reasoning has occurred or is now occurring and of the efficacy (reliability) of certain sorts of reasoning. To obtain such knowledge, the agent must be able to form beliefs about the truth of past beliefs and introspect its reasoning processes. It must also be able to reason inductively to generalizations about the reliability of certain kinds of beliefs, sources of beliefs, processes, and methods. So the default theoretical reasoner that provides the premises for the practical reasoning involved in theoretical reasoning must contain a general ("planar") reasoner, an introspective monitoring module, and a truth-evaluation module.

The planar reasoner in this default theoretical reasoner will proceed by generating arguments for conclusions and retracting arguments and beliefs in the face of defeaters. The business of generating arguments falls to an automated default monotonic (or deductive) and nonmonotonic (or defeasible) reasoner. Pollock begins by constructing the portion of this default reasoner dedicated to monotonic or deductive reasoning. Though AI researchers have long attempted to construct automated monotonic theorem-provers for predicate logic, Pollock has found their work of limited use, for three reasons. First, the programs developed in this literature are capable of proving interesting theorems only when tailored to the theorem at hand. Second, the most widely used strategy, resolution refutation (i.e., reductio ad absurdum), has no obvious extension to defeasible reasoning; yet the strategies of the monotonic reasoner should extend to defeasible reasoning. Finally, most AI automated theorem-provers search only a limited range of argument structures, and they lack suppositional reasoning.

A feasible reasoner, Pollock argues, must be guided by its *interests* when it generates arguments. For it is not feasible to generate all possible arguments from a set of premises. AI automated theorem provers, by contrast, generally leave out interests–a cause of massive inefficiency. Pollock therefore proposes, in line with ideas familiar from AI research on problem-solving, that the default reasoner use a combination of *forward chaining* (i.e., reasoning from what is already believed) and *backward chaining* (reasoning from what it is trying to prove, a matter of deriving interests from interests). The aim is to make the conclusions from forward chaining match the interests from backward chaining.

One of Pollock's most significant findings here is that backward and forward chaining employ different natural deduction principles. As he puts it, backward chaining is not just forward chaining in reverse. Some natural deduction principles are useful only for backward chaining–e.g., (1991, 203)

adjunction: $\{p, q\}$ is a reason for $(p \,\&\, q)$
addition: p is a reason for $(p \vee q)$
q is a reason for $(p \vee q)$
– introduction: p is a reason for $-p$.

Pollock argues that these principles govern only "backward reasons": "For instance, consider addition. Suppose we could use addition in random forward reasoning. Then if we adopted *p*, we would be led to adopt every disjunction containing *p* as one disjunct. But there are infinitely many such disjunctions, and most are useless in any given problem" (1991, 203). Thus, we use addition only when we have some reason to be interested in the resulting disjunction. Other natural deduction principles are, similarly, useful only for forward chaining–e.g.,

simplification: $(p \,\&\, q)$ is a reason for p
$(p \,\&\, q)$ is a reason for q

– elimination: $-p$ is a reason for p
modus ponens: $\{(-p \vee q), p\}$ is a reason for q.

Pollock proposes rules of adoption which govern the combination of forward and backward chaining. In simplified form,

> If p is a forwards or backwards reason for q, and you already believe p, then adopt q; and if you are interested in q, and X is a backwards reason for q, then become interested in X.

Pollock suggests that OSCAR's architecture is closer to the architecture of actual human reasoning than typical AI monotonic reasoning programs. In support of this suggestion, he reports that OSCAR is faster than most AI programs using a reductio format, and, unlike the rest, it achieves its results on hard problems without tweaking by a human operator. To explain OSCAR's advantage in speed, Pollock appeals to the fact that OSCAR's interest-driven architecture avoids unnecessary inferences. Only eleven percent of OSCAR's inferences are unnecessary for its 42-line proof of the Schubert steamroller argument in predicate calculus, compared with 479 inferences in the fastest alternative system (Stickel 1986).

We have discussed deductive reasoning. Pollock has also developed a version of OSCAR that reasons defeasibly, one that (unlike proposed nonmonotonic reasoners in AI) can reason suppositionally.

One might question Pollock's account of rational deductive reasoning on various grounds. I will focus here on the question of its psychological plausibility. Pollock aims to develop a reasoner that reasons approximately the way human beings do when they reason rationally. Has he succeeded in this aim? The answer to this question matters because Pollock is after a general theory of rational deductive reasoning. But if OSCAR reasons deductively in a manner alien to human reasoning, a manner in which humans are unable to reason, then the theory's claim to generality will run afoul of the constraint that "ought" implies "can." At best, Pollock will have described one way in which a deductive reasoner *could* reason rationally.

Recent empirical psychology of reasoning brings bad news for Pollock's theory. For it calls into question whether ordinary deductive reasoning proceeds by anything like the strategies of Pollock's default monotonic reasoner or even by anything like natural deduction. There is first the point, all too familiar to teachers of introductory philosophy and logic courses, that people are surprisingly bad at deductive reasoning on a wide range of subject matters and readily commit such fallacies as affirming the consequent and denying the antecedent. Psychologists have long confirmed this observation with studies of Wason's selection task (Wason 1966, 1983) and the THOG problem (Wason 1977, Wason and Brooks 1979, Griggs 1983). Indeed, these studies raise the possibility that invalid reasoning is widespread and fairly systematic. If, as now seems plausible, valid and invalid deductive reasoning employ similar sorts of processes, then people will not have as part of their psychological equipment a valid deducive reasoning module like Pollock's default monotonic reasoner. If there is a deductive reasoning module at all, it will be an automated reasoner that exhibits both valid and invalid reasoning, and thus both rational and nonrational reasoning. Pollock responds to this point by urging that invalid reasoning results from interference with a deductive competence explained by his default monotonic reasoner (1989, 185).[43] But if invalid reasoning is as widespread and systematic as it appears to be, it is implausible to explain valid reasoning by hypothesizing a reasoning competence while explaining invalid reasoning by appeal to interference with that competence. One might just as well hypothesize a competence for invalid reasoning and explain valid reasoning by interference with that competence.

Turning now from the experience of logic teachers and the experimental results to empirical theories of deductive reasoning, psychologists have proposed various theories to explain our curious mixture of success and failure at deduction. There are, broadly speaking, three kinds of theories of deductive reasoning, all aimed in the first instance at explaining judgments of deductive validity, though they explain such judgments by proposing accounts of deductive reasoning, and the judgments they explain in turn explain performance on deductive reasoning tasks like the selection task (Holland et al. 1986 ch. 9). What is notable is that none of these theories posits a competence at deductive reasoning that so much as approximates Pollock's default monotonic reasoner.

The *rule* theory hypothesizes syntactic systems of rules (notably, natural deduction systems) that are incomplete (lack valid rules) and unsound (include invalid rules) (Braine 1978, Rips 1983, 1988). This theory gained currency when it was discovered that under certain conditions people regularly judge validity or invalidity in conformity with syntactic rules. Though the rule theory is inconsistent with Pollock's claim that the default monotonic reasoner reasons approximately the way humans do, it is the empirical theory closest to Pollock's. Indeed, Pollock could amend his default monotonic reasoner to conform to the rule theory. He could do so by allowing that the default monotonic reasoner sometimes reasons invalidly employing strategies of argument construction in common with valid reasoning. He could then propose that our full ability to reason validly is acquired when, as a consequence of practical reasoning, we preempt the invalid reasoning of the default monotonic reasoner and replace it with valid reasoning. This proposal has prima facie plausibility, if indeed there is a default deductive reasoner. But, as we will see, it is far from clear empirically whether there is anything that could accurately be called a default deductive reasoner.

The *model* theory differs from the rule theory in hypothesizing that people make judgements of the validity of arguments by attempting to construct models that falsify the given argument (Johnson-Laird 1983, Johnson-Laird and Byrne 1991). Errors about validity are attributed to the limits of working memory. This theory has so far been developed in great detail only for syllogistic reasoning. The model theory explains invalid reasoning by appeal to the limitations of working memory. It may, however, avoid positing any default monotonic reasoning module at all if the process of constructing models can be assimilated to general processes of imagining or conceiving.

Whatever their plausibility in other respects, neither the rule nor the model theory has been able to give a plausible explanation of the fact that performance on reasoning tests varies greatly with subject matter–a key desideratum for a theory of deductive reasoning. With the exception of performance on practical reasoning or social contract tasks, people do better in reasoning deductively with familiar subject matter than they do with unfamiliar matter. A plausible theory must explain the facilitation provided by familiarity and the fact that syntactic form predicts consistent performance on practical tasks. In fact, there is an approach developed in the last decade that promises such an explanation. The approach has two variants: the *pragmatic reasoning schemas* variant of Cheng and her colleagues (Cheng and Holyoak 1985, Cheng et al. 1986) and what might be called the *social contract reasoning* variant (Cosmides 1989, Cosmides and Tooby 1992). I will discuss the former. The pragmatic reasoning schemas theory hypothesizes schemas for reasoning about pragmatic matters (e.g., rules relating obligation and permission to the performance of actions). It then explains successful judgments of validity as a function of the ease of mapping concrete situations into pragmatic schemas and the degree to which schemas evoke valid inferences. This theory appears to be more potent than both the rule and model theories in explaining various findings concerning performance (e.g., the finding that giving subjects a rationale for their reasoning benefits performance).

Adjudicating between the rule, model, and pragmatic reasoning schemas theories is of course beyond the purview of this article. I bring them up only to observe that on none of the currently viable theories do human beings ordinarily reason deductively by employing anything like Pollock's default monotonic reasoner. The empirical evidence shows not only that people lack anything analogous to Pollock's default reasoner (with its special strategies), but also that they lack anything analogous to a natural deduction reasoner, or indeed any kind of monotonic reasoner approximating the power of natural deduction systems. Even on the rule theory, which is closest to his view, Pollock's default reasoner would need substantial amendment to describe the reasoning processes people actually employ in successful reasoning. But so amended, the theory would no longer allow that outputs of the default reasoner are rational, unless Pollock were willing to embrace a wide spectrum of invalid conclusions as rational. On the amended theory, it would no longer be true that a rational belief is one that results from the operation of the architecture.

This is not to deny that people *sometimes* rationally reason by strategies resembling those of Pollock's default monotonic reasoner. In fact, the reasoner employs strategies much like those taught in introductory logic courses for constructing proofs in natural deduction systems of propositional and predicate logic. More exactly, students in such courses are taught to determine the validity of given arguments by attempting to construct proofs using forward and backward chaining of the sort Pollock's reasoner uses. Early in the development of their logical problem-solving skills, students labor at these strategies, and at this point their reasoning resembles that of Pollock's default monotonic reasoner (save that they use the strategies self-consciously, having selected them on the basis of their teachers' testimony as to its reliability). After repeated trials, successful students gain proficiency in constructing proofs. Arguably, their reasoning at this stage automates the strategies. If so, Pollock's default monotonic reasoner provides a plausible model of the advanced reasoning involved in solving logic problems–a notable achievement, and one that might contribute to the improvement of logic pedagogy.

The default monotonic reasoner does, then, approximate and plausibly model the way people sometimes rationally reason. But I do not believe that this observation provides significant support for Pollock's claim to have developed a psychologically plausible model of rational deductive reasoning. First, the circumstances in which people reason this way are very limited–restricted perhaps to performance in logic courses. Pertinent here is the empirical evidence that courses in propositional logic have little effect on reasoning competence as measured by standard problems from the psychology literature (e.g., Wason's selection task). Cheng et al. (1986) report that after a one-semester introductory course on the propositional calculus and deductive fallacies, students showed a bare three percent improvement in performance on standard problems. The largest improvement was a ten percent decrease in Affirming the Consequent. The best explanation of this pedagogical failure is that the reasoning strategies students learn in logic courses do not reinforce or naturally supplement native reasoning strategies. Native reasoning strategies would appear to differ markedly from learned strategies. This is perhaps why students backslide from learned strategies to their familiar and more comfortable native reasoning strategies as soon as they leave the classroom.

More importantly, even in circumstances in which people use the strategies of the default reasoner, it is doubtful that these strategies are what makes the reasoning rational. The fact that these strategies were selected on the basis of the teacher's testimony would seem at least as important for rationality as the fact that the strategies lead to valid inferences. Arguably, the use of the strategies would not have been rational if they had not been selected on the basis of the teacher's testimony. Nor could Pollock invoke the practical reasoner here and have it do the work of selecting the strategies. For that would require a prior default monotonic reasoner to provide premises for the practical reasoner, and that default reasoner would presumably differ from the one currently under construction–in which case, Pollock would need to describe a different default monotonic reasoner. I am inclined to doubt, then, that Pollock's theory offers a psychologically plausible account of the rationality of deductive reasoning even for actual human rational reasoning that employs the strategies of the default monotonic reasoner.

In short, on the available empirical evidence, people have nothing analogous either to Pollock's default monotonic reasoner or to a monotonic reasoner with the power of a natural deduction system. The strategies of Pollock's default monotonic reasoner are employed only in a very limited class of cases. Moreover, Pollock's theory of rationality misidentifies what makes the reasoning in this limited class of cases rational. Let me remind you, however, that all but the last of these points are psychological points about the realism of Pollock's default monotonic reasoner. They do not by themselves rule out his model as a correct model of rational reasoning. Pollock could always maintain, albeit in heroic violation of the contraint that "ought" implies "can," that our naive assumption that human beings reason rationally is simply predicated on the assumption–false, as it turns out on empirical investigation–that human beings have the default monotonic reasoning module that he posits. But in my view, Pollock ought to give up the idea that his model is the correct model for human rational reasoning and retreat to the more modest proposal that his default monotonic reasoner merely describes a *possible* rational being, even if not a rational human being.

Pollock's theory illustrates some of the virtues, and one of the vices, of employing AI formalisms in epistemology. Attempting to use an AI formalism to develop an epistemological theory has the advantage of forcing us to be explicit and specific about the architecture and processing of justified beliefs. It also suggests particular architectures and processes, and it enhances our ability to generate counterexamples to proposed theories. At the same time, it tempts us to generalize from possible ways of reasoning rationally to a general model of rationality that turns out, on empirical inspection, not to apply to human beings.

5.2 Thagard on Coherence and Acceptability

Coherence theories of justification became increasingly popular in the 1970s as dissatisfaction with foundationalism grew (Lehrer 1974, 1990, BonJour 1985, Harman 1986, Bender 1989). According to coherentism, a subject is justified in a belief just in case the belief belongs to some coherent subsystem of the subject's system of beliefs–alternatively, just in case the belief coheres with the subject's entire system of beliefs. Coherence among the members of a set of beliefs is, in turn, defined either as a cyclical network of basing relations or as a symmetrical relation of mutual support among beliefs. On either of these definitions of coherence, coherentism is inconsistent with foundationalism as we defined it (subsection 4.2), since coherentism denies that any beliefs are justified, even in part, independently of their relation to other beliefs. The network of basing relations and the relations of mutual support that define coherence are usually further characterized in terms of mutual consistency, logical entailment, probabilistic dependence, or relations of explanation.

The coherentist is burdened from the start with two formidable analytical tasks. First, the very format of coherentism requires that there be a single, integrated ordinal scale of coherence on sets of beliefs, since on coherentism, beliefs qualify as justified when they have sufficiently much coherence. The coherentist must therefore amalgamate into one scale the diverse relations–logical, probabilistic, or explanatory–that define coherence (or otherwise choose a single relation from among these with which to identify coherence). Second, relations of logical entailment, probabilistic dependence, and explanation are *local* relations among beliefs. Yet, on coherentism, the justification of a belief is a function of the coherence of the whole (sub)system of beliefs–its *global* coherence. Hence, the global coherence with which the justification of a belief is identified must somehow emerge from local relations of entailment, probabilistic dependence, and explanation. Global coherence must be a function of these local relations.

In fact, coherentism must not only characterize global coherence in terms of local logical, probabilistic, and explanatory relations. It must also characterize the justification of a belief by its particular position in the network of relations of the (sub)system of beliefs. What makes a belief justified varies from one belief to another–as witnessed by the fact that justified beliefs vary greatly in their degrees of justification. On coherentism, the difference in the justification of diverse beliefs derives from the different positions beliefs occupy in the global structure of the coherent (sub)system. Thus, if the justification of a belief is identified with global coherence, then global coherence must be relativized to a belief and assigned, relative to a given belief, in virtue of the position of that belief in the global structure of the (sub)system.

A natural way to try to satisfy this last demand is to begin by defining the justification of a given belief as a function of its local relations with neighboring beliefs–beliefs to which it is immediately related by the relevant logical, probabilistic, and explanatory relations. But clearly, the significance of these relations for the global coherence of the belief must be tempered by the prior epistemic status of the neighboring beliefs. This seems to call for figuring the global coherence of the neighboring beliefs first–but of course that would lead to a regress of calculations. It seems we can pull off this strategy for defining justification in terms of local relations only if we can calculate the local relations of all beliefs simultaneously.

That is where a connectionist computational model may come in. Connectionist models were developed by AI researchers in part to solve the problem of modeling networks of units or nodes, linked by relations, in which the value assigned each node is a function of its global position in the network (its relations to many nodes).[44] At the same time, the value (or activation level) of a node is computed strictly by local computations–i.e., as a function of its relations to neighboring nodes only, not to distant nodes. Thus, the value of a node is globally defined but computed locally. But this is precisely what is wanted to model coherentism computationally, since on coherentism, the coherence with which the justification of belief is identified is global, yet determined by local logical, probabilistic, and explanatory relations. Relations among nodes are represented in the network by weights on the links between the nodes, and the activation level of a node is computed as a function of its prior activation level together with the activation levels of neighboring nodes, weighted by the weights on the links. Obviously a local computation of global functions cannot be accomplished in a single pass of simultaneous parallel computations for each node. In some cases, it can be accomplished in a single pass or cycle of *serial* computations that propagates through the whole network in each cycle. We will see an example of this technique when we discuss Bayesian networks below. A more popular technique, however, is to treat the network as a *dynamical* system in which the computation takes place over many discrete cycles, each involving simultaneous *parallel* computations over all nodes. At each cycle, the value assigned a node is updated by its relations to its immediate neighbors. Over many cycles, the update of the value of a node eventually represents its relations to distant nodes. The network can be said to compute (to an approximation) a unique activation level for each unit if the activation levels stabilize (go asymptotic to a limit) in few enough steps.

Connectionist models have been used to provide local computations for two versions of coherentism. They have been used to model Bayesian conditional updating (by Bayesian networks). They have also been used by Paul Thagard (1989a, 1989b, 1990, 1991, 1992) to model the role of explanatory coherence in scientific theory choice. Here I will focus on Thagard's model of explanatory coherence, as presented in "Explanatory Coherence" (1989a), and postpone discussing Bayesian networks until section 5.

Thagard wishes to characterize the normative acceptability of scientific theories in light of empirical evidence (roughly, justified belief in, adequate support for, or rational belief in those theories on the basis of empirical evidence). He is also interested in modeling in much the same way the psychological state of acceptance for historical cases of theory acceptance, but I will set aside that interest here.

Thagard begins his development of explanatory coherentism by endorsing a number of intuitively plausible principles that relate acceptability to coherence, where coherence is implicitly defined by explanatory, analogical, and logical relations among evidence (or observational propositions) and rival scientific hypotheses.[45] They do so by implicitly defining coherence in terms of explanatory, analogical, and logical relations. Here is a simplified list of principles:

Symmetry: Coherence is a symmetrical relation, and so is incoherence.
Explanation: If P1,..., Pm together explain Q, then each Pi coheres with Q, and the Pi pairwise cohere.
Data Priority: Observation propositions have a degree of intrinsic acceptability.
Contradiction: If P contradicts Q, then P and Q incohere.
Acceptability: The acceptability of a proposition P in a system S depends on its coherence with the propositions in S.
System Coherence: The global explanatory coherence of a system S is a function of the pairwise local coherence of those propositions.

One difficulty with principles like these is that it is hard to apply them to particular cases of theory choice. For it is hard to see just what implications they have for any particular case. Thagard accordingly undertakes to bridge this gap between epistemic principles and particular cases, and he proposes to do so by modeling the principles in a connectionist computer program, ECHO. In ECHO, there is a network of nodes representing propositions, including hypotheses and observation propositions. Links between nodes represent explanatory and logical relations between propositions. Each node is assigned an activation level representing its degree of acceptability (positive up to 1 for acceptability, and negative down to -1 for rejectability). Each link between two nodes is assigned a weight representing a summary of the explanatory and logical relations between the propositions represented by the nodes–their relations of coherence and incoherence. Coherence between propositions is represented by an excitatory link (i.e., a positive weight), while incoherence is represented by an inhibitory link (i.e., a negative weight). The pairwise coherence and incoherence relations between propositions remain constant over cycles. Thus, ECHO does not use a backpropagation algorithm to learn correct weights from an external "teacher" by modifying the weights over cycles, as in many other connectionist systems. The numbers assigned the links therefore represent explanatory, analogical, and logical relations between propositions. The excitatory and inhibitory weights are set experimentally for each case to yield the intuitively correct acceptances in the case.

In running an example of theory choice on ECHO, the activation levels of observation and hypothesis nodes are initially set at 0, and positive activation is propagated through the system from prior nodes directly linked to the observation nodes, each prior node having an activation level permanently fixed at 1. The activation level of each node is modified at the next cycle as a function of its own current activation level as well as the current activation levels of neighboring nodes, modified by the weights on the links. Nodes connected by excitatory links will (other things equal) mutually increase their activation levels over many cycles (if both are positive), while nodes connected by inhibitory links will mutually decrease their activation levels over cycles (if both are positive). In general, as the program goes through its cycles, the weights force changes in the activation levels to raise the levels of coherent propositions. In the typical case, rival coalitions of propositions emerge, each coalition internally coherent but incoherent with its rivals. Often one coalition gains ascendancy in the positive activation levels of its propositions, and its incoherence with rival coalitions suppresses their activation levels to the negative range. Coalitions with positive activation levels are acceptable. Coalitions with negative activation levels are rejected. Thagard has run his program on a variety of episodes in the history of science, including Darwin's evolutionary theory, Lavoisier's oxygen theory, the geological revolution, the plate tectonics controversy, and the dinosaur extinction controversy (Thagard and Nowak 1988, 1990, Thagard 1989b, 1990, 1991). In each case, ECHO endorses the intuitively acceptable theory.

Thagard claims that his model solves the two formidable analytical problems confronting coherentism we mentioned at the outset of this section. First, it integrates into a single measure of coherence the explanatory and logical relations that define coherence. The principles of explanatory coherence by themselves must be integrated if they are to be applied to judge acceptability in given examples. The connectionist model bridges the gap between the principles and judgments of acceptability in given cases by computing a single measure of coherence for a given proposition (the activation level of the node) from initial weights over the whole network that represent explanatory and logical relations. This approach has the advantage that it is not necessary to characterize the notion of coherence by giving necessary and sufficient conditions that make clear judgments in cases–an exceedingly difficult task.

Second, according to Thagard, the connectionist model assigns coherence and acceptability to a proposition as a function of the proposition's position in the global network of explanatory and logical relations, while at the same time computing coherence and acceptability locally. Thus, the acceptability of a proposition is defined by its global coherence but computed by local computations alone.

Turning now to the evaluation of Thagard's theory, it must be admitted that all of Thagard's principles have some initial plausibility. His claim that the model assigns coherence and thus acceptability to a proposition as a function of the proposition's global position, while computing coherence locally, is of course correct and a major contribution to the development of coherentism.

Despite this, there are two worries about Thagard's theory that raise doubts about whether connectionism can contribute to the development of coherentism in the way Thagard proposes, and more broadly, point up the difficulty of applying formal structures from AI to epistemology.[46] One worry concerns the intended status of the AI formalism. This is a worry related to familiar worries about the status of other mathematical (logical, statistical) formalisms in epistemological theories, but there is one difference between most AI formalisms and other familiar formalisms that makes special trouble for the use of AI. Clearly, the connectionist model does not characterize our *ordinary* concept of acceptability. People need not think of or ascribe anything like connectionist computation, still less compute the coherence of a hypothesis using connectionist computation in order to *ascribe* acceptability to a hypothesis in a given case. For one thing, we do not think that a subject's hypothesis is acceptable only if the subject accepts the hypothesis as a result of a connectionist computation of the sort the model describes. To judge whether a subject's hypothesis is acceptable, an evaluator does not need to inquire into whether the subject has computed the coherence of the hypothesis by connectionist computations. Indeed, it would not have occurred to anyone to inquire into such a matter until Thagard proposed his model. So the connectionist computation is not necessary for justification and thus cannot be part of what we ascribe when we ascribe acceptability to the hypothesis. Moreover, we do not think that an evaluator must check to see whether the hypothesis is the one that *would* have been selected by such computation. For it seems unlikely that this could be accomplished by any task other than the connectionist computation itself, and we do not think that when people evaluate acceptability, they as evaluators must implement anything like Thagard's connectionist program. Indeed, I know of no empirical evidence that actual evaluators or subjects implement anything analogous to the program. Thagard himself suggests the connectionist model as a model of psychological acceptance applicable to historical examples of theory choice. But I know of no empirical evidence in favor of this proposal, beyond the very thin evidence supplied by the fact that the model endorses intuitively acceptable hypotheses in diverse examples (once the parameters have been suitably tweaked). Now, it might be urged instead that the model characterizes, not our *concept* of acceptability but the *property* of acceptability. But it is unclear just what property this would be, if not the one fully characterized by our concept of acceptability. We are left, then, with the idea that the connectionist model computes acceptability in a manner that *happens* to be coextensive with acceptability (or happens to be necessarily coextensive with it?). But this idea would not seem to be the solution to a problem. Rather, it presents a new problem: explaining why this sort of computation happens to coincide with acceptability. Admittedly the worry here is no different in kind from familiar worries about the use of formalisms in other epistemological theories, but it does seem to me different in degree. In the case of logic or statistics, there is some prima facie reason to ascribe the formalism to our concept, if only implicitly or inchoately. There is also some prima facie case for saying that we employ an approximation to such a formalism in evaluation (though, as I have argued this turns out on empirical inspection to be mistaken). But in the case of the connectionist model, the degree of arithmetical complexity and formal sophistication of the model renders it unlikely that anything in our concept of acceptability or our practice of evaluation commits us to thinking that way. The very sophistication of AI formalisms raises a doubt about their suitability for the epistemological project of characterizing our concept of acceptability.

There is a second worry about the connectionist model. The model clearly adds assumptions about acceptability *not* expressed or entailed by Thagard's coherentist principles. Now, there need be nothing wrong in general with the model's adding assumptions not expressed by the principles. On the contrary, the model is of theoretical value (as opposed to mere practical value for computing judgments in particular cases) only to the extent that it *does* introduce assumptions that go beyond the principles. (If it did not do so, then the principles would by themselves judge the cases, and the model would be a mere computational device that is in principle dispensable.) But these assumptions must themselves be plausible and, moreover, motivated by the coherentist conception of acceptability, if the theory is to be plausible and the model is to serve the purpose of confirming coherentism by making the intuitively correct judgments of cases. My particular worry here is that Thagard's epistemological interpretation of activation levels is arbitrary and leads to an implausible model of acceptability. The implausibility lies in the comparative nature of the activation levels after cycling. Thagard claims as a virtue of his theory that the model is inherently comparative. But in fact, although acceptability is comparative, it is not comparative in the way the model implies, and this is a deep flaw in the model with no obvious remedy.

I will argue the point by showing how Thagard's theory runs afoul of some intuitively plausible claims about acceptability. It is plausible to suppose that acceptability is essentially tied to degrees of acceptability. It would perhaps be more natural to speak of degrees of justification or support. The acceptability of a proposition would seem to be a function of its degree of justification. Acceptability is (in part) comparative, but degree of justification is not. The degree of justification of a theory (e.g., phlogiston theory) is not a function of the degree of justification of any competing theory (e.g., oxygen theory). We can see this by noting that the arrival of Lavoisier's theory (assuming no change in the evidence) does not *reduce* the degree of justification of phlogiston theory. The degree of justification remains what it was. That is because the relevant degree of justification concerns the theory's relation to the evidence, not its relation to Lavoisier's theory. Phlogiston theory does not become unacceptable because its degree of justification slips but because Lavoisier's theory has a higher degree of justification. All this suggests that a theory is acceptable just in case it has a sufficiently high absolute degree of justification, and its degree of justification is superior to that of any rival theory.

Exactly the same points may be made about degree of coherence. The degree of coherence a theory has is not comparative. Phlogiston theory does not lose its degree of coherence because it is confronted with oxygen theory. It is just as coherent after the arrival of Lavoisier's theory. On a plausible version of coherentism, a hypothesis is acceptable in virtue of its degree of justification. In particular, it is acceptable just in case it has a sufficiently high absolute degree of coherence, and its degree of coherence is superior to that of its rivals.

If all this is right, then the notion of degree of coherence that bears on acceptability is a *noncomparative* notion. This is true even of the notion of degree of coherence that is employed in the comparison between rival hypotheses. In fact, the *same* notion of degree of coherence is employed in both the absolute component and the comparative component of the above coherentist account of acceptability. A hypothesis is acceptable just in case its noncomparative degree of coherence is high on an absolute scale and comparatively high, too–higher than the noncomparative degrees of coherence of rivals. Thus, in a plausible coherentism, we cannot represent acceptability by a degree of coherence that is computed comparatively. This is so even though acceptability is itself in part comparative.

Unfortunately, in Thagard's theory, acceptability is represented by activation levels, and these are computed comparatively. Phlogiston theory gets a low activation level because it is incoherent with oxygen theory. But the incoherence we refer to here is *not* the degree of coherence relevant to the defeat of phlogiston theory, on a plausible version of coherentism. True enough, the incoherence here *is* relevant to the rejection of phlogiston theory because it makes the two theories *rivals*. But this incoherence does not enter into the degree of coherence that defeats phlogiston theory. Phlogiston theory is rejected because its noncomparative degree of coherence is lower than that of its rival, oxygen theory. The activation levels cannot represent the degrees of coherence that figure in what makes the phlogiston theory unacceptable, according to a plausible coherentism. Since the noncomparative degrees of coherence crucial on coherentism are not represented by the activation levels, or by any other parameters of Thagard's model, the model cannot adequately represent acceptability according to coherentism. The mistake here is making the relations of incoherence between rival theories figure in the degrees of coherence of the theories, when they really pertain only to the fact that the theories are rivals. Rivalry is, to be sure, an important factor in acceptability because it determines the reference class of theories by comparison with which the acceptable theory has the highest degree of coherence. But the incoherence that bears on rivalry does not directly enter into the degrees of coherence computed in the competition. In a plausible coherence theory, the determination of rivalry is separate from and prior to the comparison of degrees of coherence. I am inclined to think that Thagard's theory of acceptability exhibits the first pitfall of using AI in epistemology I mentioned earlier: it inadvertently interprets the activation levels in the connectionist model in a way that commits it to an implausible epistemological thesis.

As far as I can see, Thagard's best bet for responding to this criticism is to deny that the notion of acceptability he analyzes is a notion of justification or support. He might say that it is more like rational belief, and though there are rational degrees of belief, there is no degree of rational belief; consequently acceptability differs from justification in not being a matter of highest noncomparative degree of justification or coherence. I am inclined to think that there is such a thing as degree of rational belief, and that it behaves much the way degree of justification does on the story I have outlined. (On a Bayesian account of rational acceptability, for example, degree of rational acceptability would be measured by the expected epistemic utility of acceptance.) But I will content myself with two points here. First, if acceptability is not a matter of degree of justification or coherence, then we are left with the question what the intuitive meaning of the activation levels is supposed to be. Second, if Thagard's notion of acceptability is so different from the notion of justification, his connectionist approach loses its applicability to coherence theories of justification.

In my view, Thagard's theory exhibits some of the virtues and several of the pitfalls of applying an AI formalism to epistemology. It is true that the connectionist model has features that mirror certain formal features of the coherence relations among beliefs. But it is unclear how much resemblance the model bears to actual human belief-formation or to evaluators' ascriptions of acceptability. And there is one key parameter of the model that is given an interpretation corresponding to nothing in epistemological nature.

6 THE TARGETS OF EPISTEMIC EVALUATION

We have considered at length the bearing of cognitive science on the conditions of justification. I have argued that, under wide reflective equilibrium method for selecting the correct conditions of justification, cognitive science could bear on the choice of format for an account of justification, that empirical information from cognitive science could bear on the plausibility of normative principles of rational belief (via the constraint that "ought" implies "can"), and that a nonidealizing approach to the normative principles of rational belief, an approach that begins with empirical information about human cognitive limitations, is superior to the approximate idealizing approach. I have conceded that AI formalisms have the potential to enhance our thinking about justification, though I have raised doubts about particular efforts along these lines.

Let us turn, finally, to a different matter–the targets of epistemic evaluation, i.e., the states to which epistemic epithets like "justified" and "rational" apply. Epistemologists have traditionally assumed that the primary targets of epistemic evaluation are beliefs or degrees of belief (though other items have also been taken as targets: belief-forming, evidence-gathering, and hypothesis-generating processes and methods, cognitive acts and character traits, and speech acts like assertions). It is natural to look to cognitive science to judge the wisdom of these choices.

Suppose, for example, cognitive science turned out to provide good reason to reject our common sense assumption that people have beliefs. Then we would have to admit either that beliefs were not, as we had supposed, the targets of our everyday evaluations, or that our everyday evaluations of beliefs are at best convenient fictions, or otherwise that we were mistaken in supposing that a proposition like "Sam is justified in his belief" entails that Sam has a belief. The last reaction seems the least plausible of the three; but the first and second reactions clearly place us in a challenging predicament. On narrow reflective equilibrium method, empirical evidence against the existence of belief would not cast doubt on the conditions of the theory in narrow reflective equilibrium; rather, it would leave us with skepticism. On wide reflective equilibrium method, by contrast, it would be most natural to rethink the assumption that beliefs are the targets of evaluation. But whether we follow narrow or wide reflective equilibrium method, evidence against the existence of belief would motivate us to ask whether our institution of epistemic evaluation could be modified to evaluate a different, psychologically realistic target, preferably one similar to belief.

In my opinion, cognitive science is unlikely to find against belief. The state most vulnerable to empirical objection is rather *degree* of belief, and I will accordingly make this my focus here.

6.1 Degrees of Belief

Much epistemology assumes a *binary* notion of belief (also called "acceptance").[47] The idea that there are such beliefs–that people simply believe that Napoleon was defeated at Waterloo and do not believe that Wellington was–is well supported by introspection and comports with our practice of ascribing such states without qualification (as does the idea that justification applies to such beliefs and is also binary, as well as coming in degrees).

However, we also distinguish beliefs with respect to degree. I believe that Napoleon lost at Waterloo and that Abe Lincoln was born in 1809, but I am more confident of the former than the latter. Such comparative differences in confidence on a proposition are often treated as arising from different assignments of *degrees of confidence* or *subjective probabilities* to the proposition. The subjective probability of a proposition is sometimes defined as the subject's willingness to bet on the proposition. Such a willingness is clearly not a state of binary belief, not even a cognitive state at all, but a *motivational* state of willingness. It is common ground that we cannot straightforwardly define binary belief in terms of subjective probabilities. For binary belief is neither necessary nor sufficient for high subjective probability. Writers who take subjective probabilities to be a target of epistemic evaluation usually have in mind the idea that subjective probabilities are rational to the extent that they conform to the probability calculus, are updated by Bayesian conditionalization, and perhaps conform to a handful of other principles like van Fraassen's reflection principle.

Alvin Goldman (1986, 324-328) has questioned on empirical grounds whether people natively assign subjective probabilities in a systematic fashion and in conformity with these standard norms. We have already obliquely discussed some of his doubts (subsection 3.1), and I will review them only briefly here. Goldman doubts that people natively have a psychological apparatus that assigns subjective probabilities in conformity with standard norms. First, people are not always (or, for that matter, even very often) maximally confident of tautologies and necessary truths, as required by the probability calculus, standardly interpreted. Second, conforming subjective probabilities to the probability calculus requires an arithmetical facility that people do not natively have. At least, there are no other domains in which people natively show the kind of facility at arithmetic that is needed to conform subjective probabilities to the probability calculus. Third, as we saw earlier, Tversky and Kahneman (1983) and others have shown that people routinely and robustly assign probabilities in ways that violate the probability calculus–e.g., the conjunction rule and Bayesian conditionalization. Tversky and Kahneman explain these violations by appeal to the hypothesis that people employ a "representativeness" heuristic to make probability judgments, assigning probability to an event on the basis of how representative it is of the events in its reference class.

In my view, it is difficult to say whether Tversky and Kahneman's conjunction experiments establish that subjective probabilities routinely violate the probability calculus. In addition to the point mentioned in subsection 3.1, that the experimental subjects may be reporting their assignments of *objective* probabilities, rather than their subjective probabilities, there is another point: it could be that when subjects report a conjunction to be more probable than either of its conjuncts, they do so because the conjunction, combined with the known facts, tells a more coherent story than either conjunct combined with those same facts. In this case, subjects take the request for a probability assignment to require hypothesis testing in which coherence plays a large role. (Note that the test question provides a story so short and hypotheses so arbitrary that it is difficult to assess the plausibility of a hypothesis without bringing in the overall coherence of the hypothesis combined with the facts.) If so, the experimental results show little about subjective probabilities. (Moreover, so understood, the subject's judgments could be consistent with coherentist norms of rationality.)

Despite these reservations about employing the Tversky-Kahneman findings against native subjective probabilities, I believe that Goldman makes a prima facie case against systematic native assignments in conformity with the probability calculus (see also Cohen 1981). Of course, this leaves open the possibility of acquiring methods for conforming subjective probabilities to the probability calculus, and obviously some people do acquire such methods. But clearly these methods are acquired only by those fortunate enough to have mathematical training (e.g., in probability theory), and they would seem to require an arithmetical facility that renders their constant everyday use impractical. More importantly, it is an open question whether people who have training in probability theory are able to override any native tendencies they may have to violate the probability calculus. Tversky and Kahneman (1983) showed that sophisticated subjects test nearly as poorly as naive ones.

At this point, it is worth noting recent work in AI on reasoning under uncertainty that may bear, albeit obliquely, on whether there could be native processes that conform subjective probabilities to the probability calculus. I have in mind here the work of Judea Pearl (1989) and others on Bayesian networks. Early AI work on reasoning under uncertainty avoided probabilities and employed instead degrees of evidential support ("certainty factors," as in the expert system MYCIN–see Shortliffe and Buchanan 1975) or qualitative principles (see Collins and Michalski 1989, Collins 1990).[48] It did so for a reason analogous to Goldman's objection that native subjective probability-fixing processes are arithmetically too complex for human cognition: it was assumed that probabilities are computationally intractable, requiring too much storage space and too many computations for updating. Recently, however, a number of AI researchers have developed representations of probabilities and algorithms for updating them with the aim of rendering probabilistic reasoning computationally tractable.

The problem these researchers wish to solve is that of updating the unconditional probabilities of hypotheses, given the unconditional probabilities of the evidence and the conditional probabilities of the evidence on the hypotheses (typically hypotheses ascribing causes of the observed phenomena reported in the evidence). On the Bayesian approach, we update by identifying the updated unconditional probability of a hypothesis with its conditional probability on the evidence, which is computed in accordance with the Bayesian formula from the unconditional probabilities of the evidence and the hypotheses, and the conditional probabilities of the evidence on the hypotheses.

To render these computations tractable, we first represent probabilities assigned to propositions by a connectionist model–a Bayesian network. We assign each proposition and its probability at a time to a node in a tree, and we assign conditional probabilities of one proposition given another to links between the nodes representing the two propositions. There are nodes representing the evidence (evidence nodes) and nodes representing hypotheses (hypothesis nodes). A Bayesian network *explicitly* represents the unconditional probabilities of all propositions, as well as the conditional probabilities of all pairs of propositions assigned to neighboring nodes in the tree. However, it also *implicitly* represents all the conditional probabilities that follow from the explicitly represented probabilities; these implicitly represented conditional probabilities can be computed from the explicitly represented conditional and unconditional probabilities in polynomial time. (This implicit representation of conditional probabilities holds if, in the network, the most relevant predecessors of each proposition–i.e., the chain of predecessors with highest conditional probability–are identified recursively in some total order, e.g., temporal order.) Pearl has shown that, given a Bayesian network, it is possible to update the unconditional probabilities for all propositions by a message-passing algorithm, in such a way as to conform to Bayesian conditionalization and the probability calculus. Simplifying a bit, the computations of probabilities are local computations in which the update propagates through the network in a single cycle from the evidence nodes to the causal nodes (i.e., the hypothesis nodes). Each unconditional probability at a node is updated in turn by passing through the conditional probability assigned the link to the previously updated parent node. This algorithm avoids the computational complexity of simultaneously computing all unconditional probabilities on the basis of the conditional probabilities. The number of computations involved in an update is linear with the diameter of the network. (Pearl has also shown how to use similar message-passing algorithms to compute probabilistic inferences that take categorical evidence as input and yield the most probable conjunction of hypotheses as output.)

This algorithm clearly provides an answer to those who worry that any Bayesian updating of a large system of subjective probabilities is bound to be computationally unfeasible for currently existing computers. Pearl's algorithm has use in probabilistic reasoning in automated expert systems. This said, it must be admitted that Pearl's algorithm for Bayesian updating provides no direct answer to Goldman's worry that people lack native mechanisms for systematically assigning subjective probabilities in conformity with the probability calculus and standard principles for updating probabilities.[49] Pearl intimates at several points that his algorithm addresses this worry. But all that Pearl shows is that there is an algorithm for Bayesian updating that can run on current electronic digital computers. Clearly, this does not so much as suggest that the algorithm is or can be implemented in native human cognitive mechanisms. To be sure, there is no decisive empirical evidence that it cannot be implemented. But the empirical evidence that people routinely violate norms provides some reason to doubt whether it can be. (Note, too, that the algorithm will be *acquired* as a *method* of updating only by specially trained individuals in possession of computers.)

Though empirical evidence throws in doubt the claim that we systematically represent probabilities or accurately compute their values, there have been a number of attempts to show that reasoning that conforms to plausible qualitative principles proceeds as if it were guided by probabilistic reasoning. I will mention two of these attempts.[50]

First, L. J. Savage (1954) famously showed that subjects whose preference ranking of items conform to certain intuitively plausible axioms will prefer items as if their preferences were governed by subjective probabilities.[51] In particular, if a subject satisfies such axioms as transitivity of preference (and indifference), then there is an assignment of subjective probabilities to states of nature and utilities of the items given the states of nature under which the subject will prefer an item *X* to another *Y* just in case the subjective expected utility of *X* (i.e., the average of the utilities given the states of nature weighted by the probabilities of those states of nature) exceeds that of *Y*. This representation theorem of course does not entail that subjects who conform to the axioms actually have psychological states of subjective probability. It does not entail, for example, that they will have a willingness to bet in certain ways (though it might be claimed that lacking such a willingness is *normatively* inappropriate, given the subject's preferences). The theorem entails at most that it is as if subjects make choices guided by subjective probabilities.[52] However, as Goldman points out, there is substantial empirical evidence that subjects do *not* regularly (and perhaps cannot very often) satisfy the axioms governing preference (MacCrimmon 1968, Tversky 1969, 1975, Kahneman and Tversky 1979, Shafer 1990). In addition, there is empirical evidence that subjects do not conform their choices to subjective expected utility (Tversky and Kahneman 1979). So the representation theorem does not lend any support to the claim that subjects routinely make choices as if guided by probabilities (or that they can do so).

Second, there is a probabilistic semantics for the qualitative logic of defeasible reasoning developed by Ernest Adams (1975) and subsequently elaborated by Geffner and Pearl (1987). This logic was an early nonmonotonic logic designed to accommodate defeasible inferences using a database of rules representing propositions like "Birds fly," "Penguins are birds," and "Penguins don't fly."[53] In classical monotonic logic, these sentences would be translated into universally quantified conditionals and would be inconsistent with each other (assuming there is a penguin). But as used in ordinary in English, the sentences are clearly mutually consistent. So we must either abandon the classical interpretation of the sentences or treat their logic nonmonotonically. Adams's logic does both. Adams proposes an interpretation of the sentences and axioms governing inference that avoids the inconsistency. These intuitively plausible axioms (triangularity, Bayes, and disjunction) are sound and complete with respect to the probabilistic semantics developed around the interpretation of the sentences, a semantics of "extreme" probabilities: "Birds fly" is interpreted as P(Fly x/ Bird x) (i.e., the conditional probability that x flies, given that x is a bird) is infinitesimally close to one.

The extreme probability interpretation has advantages and disadvantages, but what is most important for our purposes here is that the interpretation is clearly incorrect as an account of the semantics of the belief that birds fly, and this raises the question whether the logic correctly describes how we reason on the basis of such beliefs.[54] In fact, "Birds fly" is not a statistical generalization, nor can it be represented by a conditional probability. It means "Normal birds fly" or "Representative birds fly," neither of which even entails the statistical generalization that most birds fly. If these sentences conform to a nonclassical logic at all, it would presumably be a logic that mirrors the representativeness heuristic of Tversky and Kahneman, assuming there is such a logic. Lacking such a logic, it would seem reasonable to proceed classically as far as possible, interpreting the generalizations as classical universally generalized conditionals and treating classical monotonic logic as the inference engine. From "Penguins are birds," "Birds fly" (i.e., "Normal birds fly"), and "x is a penguin," we cannot infer "x flies" in classical monotonic logic, unless we also believe "Penguins are normal birds." This handles the problem of inconsistency. The difficulty for a thoroughgoing classical treatment is that, under the classical interpretation with classical inferences, we can draw a conclusion from "x is a bird" and "Birds fly" only if we believe that a given bird x is normal. The extreme probability interpretation does not require such prior reasons. But perhaps people do often tacitly assume, and have reason to believe, on statistical grounds, that an arbitrary bird is normal. And perhaps if they did not have such reasons, they would not be justified in inferring from their belief "Birds fly" and "x is a bird" to the conclusion "x flies." (Classical monotonic reasoning has its own problems as an account of the psychology of human deductive reasoning, as we saw earlier (subsection 5.1), but nonmonotonic logic fares no better as a description of human psychology.)

This is not to say that classical monotonic reasoning could stand alone without supplementation by any defeasible inferences; for it may be implausible to interpret certain generalizations as conditional on hidden properties like normality–notably, generalizations relating appearances to reality. At the present time, there is little evidence to support either a classical or a nonmonotonic model of defeasible reasoning. Returning at last to the question of what the existence of a probabilistic interpretation of nonmonotonic reasoning shows for human psychology, we should say that it again leaves open the question whether people systematically assign subjective probabilities. We may note in this regard that classical logic also has probabilistic interpretations, just as nonmonotonic logic does. But no one has thought that these interpretations have any psychological significance.

In the end, we must reiterate the point that representation theorems for preference theory and nonmonotonic logic provide no basis for ascribing subjective probabilities as psychological states. For now, it seems warranted to remain skeptical about the psychological reality of systematic subjective probabilities and about the ability of native mechanisms to conform the subjective probabilities we do assign to the probability calculus. Traditional epistemology appears to have chosen wisely in taking binary beliefs to be the primary targets of epistemic evaluation and relegating subjective probabilities to a secondary status at best.

7 CONCLUSION

Can we draw any general conclusions from our discussion? My first, upbeat conclusion is that cognitive science is indeed relevant to epistemology as traditionally conceived, in a number of ways. Cognitive psychology bears on skepticism–at the very least, on the plausibility of the premises of the arguments for Academic and Pyrrhonian skepticism. We must employ findings from cognitive science to assess which sorts of beliefs are justified, on various accounts of justification, including reliabilism, psychologism, and proper function theory. Cognitive science bears on accessibility internalism because it is relevant to the question which conditions of justification satisfy the internalist constraint. Findings from cognitive science challenge historically significant epistemological principles, such as the requirement of total evidence. Work in AI has the potential to fill out both foundationalist and coherentist accounts of justification and rational belief, and it has already been used to address worries about the computational tractability of probability updating. Finally, cognitive psychology tells us which traditional targets of evaluation really exist.

My second, cautionary conclusion is that it is by no means easy to determine just how cognitive science bears on epistemology. To determine this, we must proceed cautiously and attend to quite a few matters at once. Most importantly, we must keep track of our methodology–especially the choice between narrow and wide reflective equilibrium method–and we must distinguish levels of inquiry on which cognitive science might bear–e.g., inquiry into whether the conditions of justification are satisfied, what the conditions of justification are, what the constraints on conditions of justification are, and what the targets of justification are. Regarding the task of assessing which beliefs are justified, I have argued that there are naive empirical reasons to doubt that cognitive science can help us much with the question whether the conditions of justification are satisfied in given instances–contrary to what Goldman's reliabilism, psychologism, and proper function theory entail. But I have conceded that cognitive science may help us with the question whether the naive empirical reason for doubting this is correct. Regarding the endeavor of selecting the conditions of justification, the primary danger in employing cognitive science lies in the very sophistication of cognitive science. Epistemology aims in the first instance to comprehend our everyday concept of knowledge and make sense of our ordinary practice of epistemic evaluation. We are forbidden from importing into our analysis of knowledge sophisticated concepts alien to everyday thought. This is the rationale for narrow reflective equilibrium method. Even here, AI formalisms may help in constructing accounts of the conditions of justification, as long as we resist the temptation to import overly sophisticated concepts or empirically unrealistic computational algorithms. When we turn to the task of selecting the conditions of justification with empirical content (as in normative principles of rational belief), it seems that cognitive science could well be relevant. For example, we may rule out normative principles of rational belief on the basis of empirical information about human abilities, in light of the principle that "ought" implies "can." Thus, cognitive science can be relevant to epistemology as traditionally conceived, and it promises ideas and tools that may advance the field.

Frederick F. Schmitt
University of Illinois at Urbana-Champaign
and Indiana University

NOTES

[1] For a philosophical account of what cognitive science is, see Von Eckhardt (1993). For general surveys of cognitive science, see Johnson-Laird (1988), Stillings et al. (1995), and Green et al. (1996). For collections of philosophical work on cognitive science, see Garfield (1990), Goldman (1993), Hookway and Peterson (1993), and Casati, Smith, and White (1994). For applications of cognitive science to philosophical problems, including epistemology and philosophy of science, see Giere (1988) and Goldman (1993).

[2] I will use the term "psychology" to refer to empirical, as opposed to philosophical, psychology. I assume that there is no obstacle to the relevance of a priori philosophical psychology to epistemology traditionally conceived as a conceptual and normative enterprise.

[3] By "intentional states" I mean mental states that have representational content–notably, propositional attitudes like believing that Napoleon lost at Waterloo.

[4] Not all cognitive science is computationalist–notably, dynamical systems approaches to cognition (Thelen and Smith 1994, Port and van Gelder 1995). So long as these are consistent with ascribing intentional states and processes to knowers in the manner described at the intentional level, these approaches do not rule out any particular general theories of knowledge beyond those excluded by intentional cognitive science.

[5] This is not to say that a requirement of *feasible* computation imposes no interesting constraints on conditions of justification. I will treat that issue in subsections 4.3 and 5.1 (see Cherniak 1986 ch. 4 for discussion). J. R. Lucas (1961) has argued that the mind cannot be (or be adequately represented by) a formal system (e.g., a Turing machine); for though no (sufficiently strong) formal system can prove that the undecidable Godel sentence is true, *we* can see that it is true. I have no sympathy with Lucas's argument (see Bowie 1982 for criticism), but I will not discuss it here, since the argument has no specifically *epistemological* content. The notion of proof to which Lucas refers is the notion employed in proof theory, and this is not an epistemological notion (e.g., one can prove a theorem one does not know to be true because one does not know the axioms of the formal system to be true). Nor does the epistemic "see" in Lucas's argument play an essential role, for we can replace it with a purely nonepistemic doxastic "believe" without loss of plausibility (assuming that the formal system "believes" only what it can prove from its axioms).

[6] There are a few well established assumptions about implementation, and it is possible that we can currently draw epistemological conclusions from them. Consider, for example, the widely accepted one hundred step rule of Feldman and Ballard (1982), according to which quick (transpiring in less than one second) but computationally complex (requiring more than one hundred elementary operations) cognition must process in parallel rather than serially. This rule would exclude, say, a connectionist model for a quick justifying process that reaches equilibrium only after one hundred cycles. But, as far as I know, no one has proposed a connectionist model that runs afoul of this limitation. See Churchland (1989) for other attempts to draw epistemological conclusions from current neuroscience.

[7] There is a good deal of clearly traditional epistemological work by cognitive scientists, especially by workers in AI. However, not everything proclaimed to be "epistemology" by cognitive scientists is epistemology traditionally conceived. The volume *Android Epistemology* (Ford, Glymour, and Hayes 1995) contains no more epistemology in the traditional sense than any other collection of AI research on reasoning, learning, and knowledge representation–which is to say, it contains very little.

[8] For discussion of the Replacement Proposal, as well as other variants of naturalistic epistemology, see Kornblith (1994a) and Maffie (1990, 1995). For a bibliography of naturalistic epistemology, see Schmitt and Spellman (1994a).

[9] See also Quine (1995).

[10] It is interesting that although, in comparing his epistemology naturalized to traditional epistemology, Quine interprets the Cartesian challenge as a challenge from underdetermination–our data underdetermine our physical object theory–he never offers the natural response to skepticism so interpreted: namely, that our physical object theory is simpler or more explanatory than the demon hypothesis. This could be because he doubts whether our physical object theory is superior in these respects, or because he believes that considerations of simplicity cannot be distinguished from empirical considerations. Another possible reason he does not offer this natural response to skepticism, however, is that he senses that the Cartesian challenge is not an underdetermination challenge (contrary to Shatz 1986–see Schmitt 1992: 235-237 for whether Descartes poses an underdetermination challenge).

[11] Stroud also offers a third criticism of Quine: that Quine's naturalistic description itself makes skepticism inevitable, since it posits a gulf between data and theory. This is a mistake on Stroud's part, however. If Quine is right that we can use empirical information to respond to skepticism, then it is possible to answer skepticism. Stroud must instead respond to Quine's claim in the manner of his first two objections. The question of the confirmation of theory on the basis of the data is a matter apart from the business of answering skepticism, and here theoretical desiderata like simplicity can enter, even though they have no power against skepticism.

[12] For philosophical discussion of human control over beliefs, see Alston (1988) and Montmarquet (1993).

[13] It is worth noting in this connection that Strawson might have taken a slightly different and in my view more plausible approach and argued instead, as Hume is sometimes taken to have done, that irresistible beliefs *are* justified. This claim might be defended on the ground that justified beliefs are epistemically permissible beliefs, and what we cannot resist doing must be permissible, since "ought" implies "can" (but see Schmitt 1992: 246-247 for doubts about this defense). This would call for a theory of justification contrary to the theory assumed by the skeptic, on which justified belief requires a degree of success in aiming at the truth. While a skeptic must reject the dictum that "ought" implies "can," the skeptic must allow debate about the nature of justification before entering the skeptical doubts, and this debate could lead to embracing the view that "ought" implies "can."

[14] Guilty of this to one extent or another are Stroud (1984, 1989), BonJour (1985), Fumerton (1987), Winblad (1989), and Shatz (1993).

[15] There are various arguments in the literature purporting to show a priori or empirically that *most* of our beliefs are justified–a conclusion that would rule out a broadly defeating role for the findings of cognitive science. *A priori* arguments appeal to principles of charity in interpretation or reflective equilibrium accounts of justification. I will mention the argument from a reflective equilibrium account in subsection 3.4 (and observe that it is really an empirical argument, despite advertising to the contrary). The arguments from charity seem to me to have no force. Charity with regard to *truth* entails at most that a majority of our beliefs are true. It doesn't follow from this that any beliefs are *justified*, even on a reliabilist theory of justification. For example, on Goldman's reliabilism, exercising processes that together would yield beliefs with a high truth-ratio (in relevant counterfactual circumstances) suffices for justification. But it doesn't follow from the fact that the processes we exercise yield *actual* beliefs with a high truth-ratio that these processes would if exercised in relevant counterfactual circumstances yield beliefs with such a truth-ratio (though perhaps charity does entail that the totality of beliefs formed in counterfactual circumstances has a high truth-ratio). The argument from truth charity is even less potent on the atomistic reliabilism I favor (see subsection 3.5.2). Charity with regard to *rationality* says that beliefs are to be interpreted so as to make most of them rational. Of course this does entail that most of our beliefs are rational

if we have beliefs. But, so far as I can tell, the best reason for that view (as opposed to the principle of humanity, which merely maximizes rational beliefs and is consistent with ascribing little rationality) is that we can literally ascribe beliefs only on the assumption of *perfect* rationality. But people who hold that view don't think we have any *beliefs* and so don't infer from charity that our beliefs are rational (Dennett 1987). So much for a priori arguments for rationality. What of *empirical* arguments? I will discuss and endorse (in subsection 3.5.2) a pragmatic argument for the accuracy of evaluation that also leads to the conclusion that many beliefs are justified. A different empirical argument, which I also think has some force, appeals to evolution. See Stich (1993) and Stein (1996) for discussion of both a priori and empirical arguments. Of course, even if some a priori or empirical argument does establish that most of our beliefs are justified (rational), this does not rule out the possibility that cognitive science might help show that many beliefs are unjustified or help discover which sorts of beliefs are justified. So it does not rule out an important role for empirical science in assessing which sorts of beliefs are justified.

[16] See also Schmitt (1992: 163-174, 225-227).

[17] There are two worries about Goldman's project that I will not discuss here, though both bear immediately on the use of cognitive science to assess which sorts of beliefs are justified. One is that reliabilism embroils us in a *circularity* when it comes to assessing whether our beliefs are justified, since we must use our belief-forming processes to judge the reliability of our processes. (See Stroud 1989, Winblad 1989, and Shatz 1993 for versions of the objection. See Goldman 1986 and Alston 1993 for replies.) Another objection is that the use of our own processes to assess reliability is bound to *rubberstamp* the reliability of our processes (see Schmitt 1992 ch. 2 for a reply). It is worth mentioning a related criticism. Winblad has charged, and Goldman (1989) concedes, that, on Goldman's theory, it is impossible for induction to lead to the result that induction is unreliable. For Goldman imposes a nonundermining condition on justification, according to which one is justified in a belief as a result of a process only if one does not believe that the process is unreliable. Hence, if induction were to lead to the belief that induction is unreliable, this belief would undermine its own justification. In my view, however, this argument does not go through on a cautious formulation of the nonundermining clause. Note that the belief that undermines is subsequent to the process undermined. So if the nonundermining clause requires the undermining belief to be prior to the process undermined, the argument fails. And there is a rationale for requiring that the undermining belief be *prior* to the process: the point of a nonundermining condition is to motivate people not to exercise a process in the presence of a belief in the unreliability of that process. Clearly, if the nonundermining condition is to serve this purpose, the relevant belief in unreliability must be prior to the possible exercise of the process (except where the subject can foresee the output of the process before exercising it.)

[18] It is surprisingly easy to overlook this point. Feldman (1989) and Dretske (1988) both do. Dretske asks: "If you do not have to know how people behave to determine how they *should* behave, why (other than determining *whether* they are behaving as they should) must one know how they behave?" (1988: 268). The answer is that, for Goldman, the a priori investigation yields only a criterion for J-rules; it does not tell us which J-rules there are. Yet a specification of the J-rules is essential for describing how people should behave. On Goldman's theory, describing how people should behave is, at one level of generality, *equivalent* to determining whether they are behaving as they should (or more accurately, able to do so). This is a simple consequence of the fact that, for Goldman, justification requires only conforming to some right J-rule system, and any J-rule system is right if it has a high truth-ratio. At the highest level of generality, the question what one should believe is a priori–you may believe anything some right J-rule system sanctions. If more specificity is desired, then you must check which J-rule systems people can conform to in light of their psychology–a question that can be answered only by cognitive science.

[19] See also the papers by Tversky and Kahneman in Kahneman, Slovic, and Tversky (1982).

[20] See Cohen (1981) and Stein (1996 ch. 3) for a critical review of these experimental results. For experimental results that raise doubts about Tversky and Kahneman's results, see Fiedler (1988) and Gigerenzer (1991).

[21] Notably, Gelman and Markman (1986, 1987), Wellman and Gelman (1988), and Gelman and Wellman (1991).

[22] See Mitchell, Keller, and Kedar-Cabelli (1986) for AI work on single-case induction and explanation-based generalization.

[23] We can also expect evolutionary psychology to address the matter in due time. Though the work in evolutionary psychology represented in Barkow, Cosmides, and Tooby (1992) is individualistic in ignoring the role of reliance on testimonial information in the evolution of cognition, there is already work on the coevolution of noncognitive social relations (Caporael and Baron 1996, Wilson 1996). See Schmitt (1994b) for discussion of social issues in epistemology.

[24] In fairness to Cohen, he appears to *define* competence as the competence to make judgments of rationality in reflective equilibrium. On this definition, rational belief will, by stipulation, match conformity to competence, if the latter is defined as being sanctioned by competent judgments. But then the question becomes whether the psychological competence that produces our beliefs in general (which is, after all, the competence found wanting by Tversky, Kahneman and company) subsumes the competence to make judgments of rationality, or indeed whether we even *have* a competence to make judgments of rationality in reflective equilibrium. These are, again, empirical questions.

[25] This makes trouble for both versions. Millikan's is in trouble because evolutionary psychology so far raises nothing but doubts about whether proper functioning is knowledge-producing and even whether our faculties are aimed at truth (Cosmides and Tooby 1992). Plantinga's has a different problem. For we currently have, at best, a tenuous view of the proper functioning of our cognitive faculties. What if it turned out, as may well be the case, that they were not aimed at truth? Plantinga would be faced with the following options: deny that there is any knowledge; relinquish the requirement that proper functioning be aimed at truth; relinquish the proper function theory. The third option would seem the best, even on narrow reflective equilibrium method (since the consideration adduced here is a hypothetical, "narrow" consideration). Plantinga has made his account of knowledge hostage to a tenuous view of proper functioning. One might take that to be an objection to the account.

[26] One might wonder whether my claim that cognitive science *could* lead to the conclusion that I cannot tell by reflection alone whether I am appeared to deeply entails that *in fact* I cannot tell by reflection alone whether I am appeared to deeply. In this case, my claim entails, not only that cognitive science could rule candidate conditions inadmissible under accessibility internalism, but also that *no* condition satisfies accessibility internalism–and hence, as I argue in the next paragraph in the text, accessibility internalism is mistaken. But this line of reasoning fails. The mere *potential* of cognitive science to show that I cannot tell by reflection whether I am appeared to deeply does not by itself entail that I cannot tell this by reflection. Presumably, it would be possible for me to tell by vision whether there is a desk in my office, even though cognitive science has the potential to show that I cannot do so. What goes for telling by vision also goes for telling by reflection.

[27] This point, if sound, generalizes to many constraints on conditions of justification: although narrow reflective equilibrium method is plausible for conditions of justification, constraints on conditions of justification that could, if strong enough, rule out there being any conditions of justification, as accessibility internalism could, must be selected by wide reflective equilibrium method.

[28] See Schmitt (1992 ch. 7) for another objection to holistic reliabilism.

[29] At this point we would, however, face the question whether the epistemic status of cognitive science so depends on the accuracy of past evaluations, or on the justification of the beliefs deemed justified, that their inaccuracy undermines this epistemic status and thus deprives our discovery of inaccuracy of the force it would otherwise have.

[30] Note, however, that in his review of Goldman (1986), Dretske (1988) appears to be skeptical of the value of cognitive of science for determining which sorts of beliefs are justified. It may be, however, that this appearance derives rather from a lack of interest in the project of specifying the sorts of beliefs which are justified, or perhaps from confusion about the structure of Goldman's theory of justification (see note 18 above).

[31] For criticism of reflective equilibrium method as a general account of justification, see Stich (1993). For an extensive review of reflective equilibrium method and its relation to cognitive psychology, see Stein (1996 ch. 5).

[32] One might suggest that empirical findings could enter into the selection process in a different way. Consider an approach on which we test conditions by first drawing out the sorts of beliefs that are justified on the condition, in light of empirical findings–much as Goldman proposes that we do for reliabilism. We then criticize the account on the ground that these sorts of beliefs are not justified according to naive intuition. (Stich and Nisbett 1980 make a parallel criticism of a narrow reflective equilibrium account of *justification*: the gambler's fallacy is in narrow reflective equilibrium for naive subjects, but it is not intuitively justified. However, they appeal to *wide* intuition here, not narrow.) Now, is this method narrow or wide? It does employ empirical information, but not as a test of the plausibility of the proposed conditions, only as a way of deducing their consequences for cases. That said, I do not know of any examples of this (hybrid?) method.

[33] I will leave aside here the semantical arguments for theory-ladenness (Sellars 1963, Churchland 1979).

[34] For Fodor's full story on modularity see his (1983). For discussion of modularity, see Garfield (1987), Barkow, Cosmides, and Tooby (1992), and Karmiloff-Smith (1992).

[35] It is true that once the foundationalist admits that sense impressions depend on contingent background generalizations about object constancy, it must also be admitted that the information represented in the sense impression is fallible, dubitable, and corrigible. But it does not follow that foundationalists must admit that observational beliefs themselves are fallible, dubitable, and corrigible. For the information represented in the sense impression that is rendered dubitable by reliance on contingent background generalizations is information about the distal properties of objects, while the observational belief is merely a belief about the appearances. In the Müller-Lyer figure, for example, one background generalization is to the effect that when objects appear a certain way, they are separated by depth. This is a defeasible generalization that happens to be mistaken for the Müller-Lyer figure. Applying it to the figure leads to the false conclusion that the lines are of unequal length, and the sense impression then falsely represents the lines as being of unequal length. But once the lines are so represented, the representation defines the appearances, and the belief that the lines appear of unequal length is true. (Needless to say, I am not claiming that observational beliefs *are* infallible, indubitable, or incorrigible. I referred in subsection 3.4.1 to Fodor's point that in the Müller-Lyer figure, according to the best available psychological explanation of the illusion, there are looks–namely, depth of field–of which the subject is unaware. That undermines not only the omniscience of observational beliefs, but also the claim that subjects can tell by introspection, for a variety of perceivable properties, whether or not the appearances represent those properties; and that in turn casts doubt on whether the subjects can tell by introspection that the appearances represent those properties even when they do represent them. But my point here is not to cast doubt on the foundationalist claim that observational beliefs are infallible, etc., but only on the claim that the theory-ladenness of sense impressions implies that they are not infallible.)

[36] It should be noted that foundationalism, as we have defined it and in its typical versions, is neutral in the dispute between epistemological pluralism and its opposite, epistemological monism, for the same reason that it is consistent with theory-ladenness. Typical foundationalism is, however, inconsistent with the *argument* for epistemological pluralism, as we have stated it. For the argument assumes that the relation between observational, perceptual, and theoretical beliefs supplies warrant to observational beliefs, but on typical foundationalism, what supplies warrant to observational beliefs is not this relation but rather sense impressions. Nevertheless, the argument can be reformulated to dispense with this assumption in favor of another consistent with foundationalism. It need only be assumed that the relation among sense impressions and beliefs is what supplies warrant for beliefs, and people with the same sensations can have different sense impressions, as well as different beliefs.

[37] Although I regard the constraint that "ought" implies "can" with suspicion for most modes of evaluation, and in particular for epistemic justification (Feldman 1988, Schmitt 1992: 94-98), I do find it plausible for "rationally ought."

[38] See Cherniak (1986 ch. 4) for the use of the "ought" implies "can" (or "economical") constraint to argue against requiring rational cognizers to believe or reason in conformity with substantial fragments of classical logic. Cherniak employs complexity theory to make the argument.

[39] Goldman makes a similar criticism of Gilbert Harman's (1973, 1986) inferential holism, according to which justifying inference takes our "total view" as input and yields a total view as output. Goldman points out that belief-processing is restricted to activated beliefs, and we can process only a few of these at a time.

[40] Jonathan Adler (1989) lodges four objections to Goldman's criticism of RTE. (1) RTE is not supposed to be "executable"–i.e., it is not a regulative principle. Thus, Adler says, the case of Melanie is not an objection to RTE but to guidance drawn from it. But, as Goldman responds, his case of Melanie tells against a nonregulative version of RTE, so long as it imposes a condition of rational belief. (2) RTE is an epistemological principle. So, according to Adler, cases that fall short of it may be *practically* rational but are not *epistemically* rational. But surely Melanie's reasoning is not merely practically rational. Moreover, her cognitive behavior is excusable on epistemic grounds. If she cannot retrieve the evidence, or even if it is epistemically uneconomical to do so, she has an *epistemic* excuse. (3) RTE may be proposed as a principle of *commitment* rather than rational belief: people are committed to what is indicated by the total evidence they possess. If someone points out that relevant evidence a subject possesses overturns the subject's actual belief, then, Adler urges, the subject ought not to hold the belief because committed to not doing so. But, in response, if commitment to not believing entails that the subject ought not to hold the belief, then RTE is mistaken even on a commitment interpretation: what if the subject is unable to avoid the belief, or unable to retract it even after it has been pointed out that it is indicated by the total possessed evidence? If, by contrast, commitment to not believing instead entails that one ought to admit that the total evidence does not favor the belief when this is pointed out, then again a commitment version of RTE is mistaken: what if one is unable to admit this? (4) RTE, Adler suggests, may be viewed as a general policy of evidence gathering or retrieval. But, to reply, if the policy is merely "retrieve as much of the possessed evidence as is feasible (economical)," then we have already written limitations into the condition to protect it from counterexamples, and if the policy does not have the qualification of feasibility, then it is vulnerable to counterexample, assuming the underlying principle is that one ought to follow the policy.

[41] For a general survey of AI research, see Boden (1996). For a survey of the use of computer models in psychology, see Boden (1988). For general collections of foundational or philosophical articles on AI, see Grimson and Patil (1987), Graubard (1988), Boden (1990),

and Cummins and Pollock (1991). For skepticism about the prospects of AI, see Dreyfus (1992) and Putnam (1992 ch. 1).

[42] There is a question, however, why this prior theoretical reasoning must be reasoning at all, rather than merely inflexible inference of the sort that goes on in Q & I modules, to which the category of rationality may not apply. Clearly, the prior theoretical reasoning would have to be inflexible if practical reasoning is necessary for flexibility. Perhaps the answer is that Q & I modules are not capable of providing the kinds of generalizations that practical reasoning must work from–e.g., generalizations about the reliability of beliefs. Only reasoning can provide such generalizations.

[43] Pollock distinguishes between warrant, which is an idealized conformity to competence, and justification, which is not idealized (1989: 126ff). Accordingly, one might wonder whether warrant might be coextensive with conformity to deductively valid reasoning, even if deductively justified belief is not. Roughly, justified belief results from strategies that would on repeated application approximate warrant. But the considerations canvassed in the text tell against even the view that people have a competence that would approximate deductively valid reasoning on repeated application. Warrant is no more coextensive with conformity to deductively valid reasoning than justified belief is.

[44] For a presentation of connectionist ideas, see Rumelhart and McClelland (1986). For philosophical work on connectionism, see Clark (1989), Bechtel and Abrahamson (1991), and Ramsey, Stich, and Rumelhart (1991).

[45] Thagard describes his coherentism as characterizing inference to the best explanation. His principles, however, do not entail that a hypothesis is acceptable only if it results from an inference to the best explanation, as an inference to the best explanation account of acceptability would do. Nor is it plausible that the connectionist computation he uses as a model counts as an inference from beliefs about the explanatory power of rival hypotheses to the acceptance of one of those hypotheses, as inference to the best explanation has traditionally been conceived (Harman 1973, 1986, Thagard 1988, Lycan 1988, Lipton 1991).

[46] There are other objections to Thagard's theory that also deserve mention. First, Thagard's claim that the model integrates explanatory and logical relations is misleading. For these relations are represented by weights combined by update. To be sure, the assignment of weights to all links implicitly compares the relative significance of explanatory and logical relations in the pairwise coherence relations between propositions. But the theory provides no rationale for the comparison. Nor is there any argument for saying that these relations are genuinely comparable in the way the theory requires. Another worry, noted by Thagard, is that there are many degrees of freedom in the model. Parameters are assigned values without any rationale besides the desire to match our intuitive judgments of acceptability in cases (as, e.g., the use of the decay parameter needed to induce a degree of skepticism in the system by arbitrarily reducing the activation levels of the nodes over cycles). And some of these parameters are tailored to specific cases simply so that the model makes the intuitively correct judgments of acceptability in the cases (e.g., the assignments of weights to the links). These unmotivated assignments are needed if the model is to judge cases and make the intuitively correct judgments. But to the extent that these assignments are unmotivated by coherentism, the fact that the model makes the intuitively correct judgments fails to confirm coherentism.

[47] But some philosophers (Perry 1980) distinguish belief from acceptance.

[48] For classic articles, see Shafer and Pearl (1990 chs 4, 5, and 8).

[49] There is another concern: that the algorithm may have limited application to real cases of updating. For the algorithm works only for a network without loops (cycles of at least three nodes connected by links in which each node is dependent on the next). If a network has a loop, then cyclical computation will typically ensue, so that the same evidence will be mistakenly counted repeatedly in updating the unconditional probabilities of hypotheses and the computation will not terminate. See Pearl (1988: 44, 195-223).

[50] Probabilistic interpretations have also been given of certainty factors (Heckerman 1986) and the logic of qualitative conditional independence (Pearl 1988 ch. 1).

[51] See Maher (1993) for a sophisticated defense of this approach.

[52] Note, too, that the theorem does not entail that any subjective probabilities the subject does assign, however appropriately, must enter into the calculation of expected utility for purposes of ordering preferences. The theorem induces no connection between the virtual subjective probabilities it entails (when preferences are properly ordered) and any actual subjective probabilities the subject may have.

[53] For exposition of nonmonotonic logics, see Konolige, Brewka, and Dix (1996).

[54] For further difficulties with the extreme probability interpretation, see Neufeld and Poole (1988) and for response see Pearl (1990: 178-180).

REFERENCES

Adams, E.: 1975, *The Logic of Conditionals,* D. Reidel, Dordrecht.

Adams, J. K. and P. A. Adams: 1961, 'Realism of Confidence Judgments,' *Psychological Review* **16**, 465-492.

Adams, P. A. and J. K. Adams: 1958, 'Training in Confidence Judgments,' *American Journal of Psychology* **71**, 747-751.

Adler, J. E.: 1989, 'Epistemics and the Total Evidence Requirement,' *Philosophia* **19**, 227-243.

Alston, W. P.: 1988, 'The Deontological Conception of Epistemic Justification', in Tomberlin, 1988.

Alston, W. P.: 1989a, *Epistemic Justification: Essays in the Theory of Knowledge,* Cornell University Press, Ithaca.

Alston, W. P.: 1989b, 'Internalism and Externalism in Epistemology', in Alston, 1989a.

Alston, W. P.: 1994, *The Reliability of Sense Perception,* Cornell University Press, Ithaca.

Audi, R.: 1988, *Belief, Justification, and Knowledge: An Introduction to Epistemology,* Wadsworth, Belmont, Ca.

Barkow, J., L. Cosmides, and J. Tooby (eds.): 1992, *The Adapted Mind: Evolutionary Psychology and the Generation of Culture,* Oxford University Press, Oxford.

Bechtel, W. and A. A. Abrahamsen: 1991, *Connectionism and the Mind,* Blackwell, Oxford.

Bender, J. W. (ed.): 1989, *The Current State of the Coherence Theory: Critical Essays on the Epistemic Theories of Keith Lehrer and Laurence BonJour, with Replies,* Kluwer, Dordrecht.

Billman, D. O.: 1983, *Procedures for Learning Syntactic Categories: A Model and Test with Artificial Grammars,* Doctoral dissertation, University of Michigan, Ann Arbor.

Billman, D. O., and E. Heit: 1988, 'Observational Learning from Internal Feedback: A Simulation of an Adaptive Learning Method,' *Cognitive Science* **12**, 597-625.

Boden, M.: 1988, *Computer Models of Mind: Computational Approaches in Theoretical Psychology,* Cambridge University Press, Cambridge.

Boden, M. (ed.): 1990, *The Philosophy of Artificial Intelligence,* Oxford University Press, Oxford.

Boden, M. (ed.): 1996, *Artificial Intelligence,* Academic Press, New York.

BonJour, L.: 1985, *The Structure of Empirical Knowledge,* Harvard University Press, Cambridge.

Borch, K. and J. Mossin (eds.): 1968, *Risk and Uncertainty,* St. Martin's, New York.

Bowie, G. L.: 1982, 'Lucas' Number is Finally Up', *The Journal of Philosophical Logic* **11**, 279-285.

Braine, M. D. S.: 1978, 'On the Relation between the Natural Logic of Reasoning and Standard Logic', *Psychological Review* **85**, 1-21.

Caporael, L. R. and R. M. Baron: 1996, 'Groups as the Mind's Natural Environment', in Simpson and Kenrick, 1996.

Casati, R., B. Smith, and G. White (eds.): 1994, *Philosophy and the Cognitive Sciences*, Holder-Pichler-Tempsky, Vienna.

Cheng, P. W. and K. J. Holyoak: 1985, 'Pragmatic Reasoning Schemas', *Cognitive Psychology* **17**, 391-416.

Cheng, P. W., K. J. Holyoak, R. E. Nisbett, and L. M. Oliver: 1986, 'Pragmatic versus Syntactic Approaches to Training Deductive Reasoning', *Cognitive Psychology* **18**, 293-328.

Cherniak, C.: 1986, *Minimal Rationality,* MIT Press, Cambridge.

Chisholm, R.: 1966, *Theory of Knowledge,* 1st edn (2nd edn 1977, 3rd edn 1989), Prentice-Hall, Englewood Cliffs, N. J.

Churchland, P.: 1979, *Scientific Realism and the Plasticity of Mind,* Cambridge University Press, Cambridge.

Churchland, P.: 1988, 'Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor', *Philosophy of Science* **55**, 167-187.

Churchland, P.: 1989, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science,* MIT Press, Cambridge.

Clark, A.: 1989, *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing,* MIT Press, Cambridge.

Clay, M. and K. Lehrer (eds.): 1989, *Knowledge and Skepticism*, Westview Press, Boulder, Co.

Code, L.: 1987, *Epistemic Responsibility,*University Press of New England, Hanover, N.H.

Cohen, L. J.: 1981, 'Can Human Irrationality Be Experimentally Demonstrated?' *Behavioral and Brain Sciences* **4**, 317-331.

Collins, A.: 1990, 'Fragments of a Theory of Human Plausible Reasoning', in Shafer and Pearl, 1990.

Collins, A. and R. Michalski: 1989, 'The Logic of Plausible Reasoning: A Core Theory', *Cognitive Science* **13**, 1-49.

Cosmides, L.: 1989, 'The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task', *Cognition* **31**, 187-276.

Cosmides, L. and J. Tooby: 1992, 'Cognitive Adaptations for Social Exchange', in Barkow, Cosmides, and Tooby, 1992.

Craig, E.: 1990, *Knowledge and the State of Nature,* Clarendon Press, Oxford.

Cummins, R. and J. Pollock: 1991, *Philosophy and AI: Essays at the Interface*, MIT Press, Cambridge.

Dennett, D.: 1987, *The Intentional Stance,* MIT Press, Cambridge.

Descartes, R.: 1984, *The Philosophical Writings of Descartes* 2 Meditations, trans. J. Cottingham, R. Stoothoff, and D. Murdoch, Cambridge University Press, Cambridge.

Dretske, F.: 1981, *Knowledge and the Flow of Information,* MIT Press, Cambridge.

Dretske, F.: 1988, 'Review of Alvin Goldman's *Epistemology and Cognition*', *The Journal of Philosophy* **85**, 265-270.

Dretske, F.: 1995, *Naturalizing the Mind,* MIT Press, Cambridge.

Dreyfus, H. L.: 1992, *What Computers Still Can't Do: A Critique of Artificial Reason,* MIT Press, Cambridge.

Ericsson, K. A. and H. Simon: 1983, *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge.

Evans, J. St B. T. (ed.): 1983, *Thinking and Reasoning: Psychological Approaches,* Routledge and Kegan Paul, London.

Feldman, R.: 1988, 'Epistemic Obligation', in Tomberlin, 1988.

Dretske, F.: 1989, 'Goldman on Epistemology and Cognitive Science', *Philosophia* **19**, 197-207.

Fiedler, K.: 1988, 'The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors', *Psychological Research* **50**, 123-129.

Fine, A. and J. Leplin (eds.): 1988, *PSA 1988*, vol. 1, Philosophy of Science Association, East Lansing.

Fischhoff, B., P. Slovic and S. Lichtenstein: 1977, 'Knowing with Certainty: The Appropriateness of Extreme Confidence', *Journal of Experimental Psychology: Human Perception and Performance* **3**, 552-564.

Fodor, J.: 1983, *The Modularity of Mind*, MIT Press, Cambridge.

Fodor, J.:1990a, 'Observation Reconsidered', in Fodor, 1990c.

Fodor, J.: 1990b, 'A Reply to Churchland's 'Perceptual Plasticity and Theoretical Neutrality'', in Fodor, 1990c.

Fodor, J.: 1990c, *A Theory of Content and Other Essays,* MIT Press: Cambridge.

Foley, R.: 1987, *The Theory of Epistemic Rationality,* Harvard University Press, Cambridge.

Ford, K., C. Glymour, and P. J. Hayes: 1995, *Android Epistemology,* MIT Press, Cambridge.

Foss, B. (ed.): 1966, *New Horizons in Psychology,* Penguin, Harmondsworth.

French, P. A., T. E. Uehling, and H. K. Wettstein (eds.): 1980, *Midwest Studies in Philosophy 5: Studies in Epistemology*, University of Minnesota Press, Minneapolis.

Fumerton, R.: 1987, 'Nozick's Epistemology', in Luper-Foy, 1987.

Garfield, J. (ed.): 1987, *Modularity in Knowledge-Representation and Natural-Language Understanding,* MIT Press, Cambridge.

Garfield, J. (ed.):: 1990, *Foundations of Cognitive Science: The Essential Readings,* Paragon, New York.

Geffner, H. and J. Pearl: 1987, 'A Framework for Reasoning with Defaults', *Technical Report R-66*, Cognitive Systems Laboratory, University of California, Los Angeles.

Gelman, S. A. and E. M. Markman: 1986, 'Categories and Induction in Young Children', *Cognition* **23**, 83-208.

Gelman, S. A and H. M. Wellman: 1991, 'Insides and Essences: Early Understanding of the Nonobvious,' *Cognition* **38**, 213-244.

Gelman, S. A and H. M. Wellman: 1987, 'Young Children's Inductions from Natural Kinds: The Role of Categories and Appearances', *Child Development* **58**, 132-141.

Giere, R.: 1988, *Explaining Science,* University of Chicago Press, Chicago.

Gigerenzer, G.: 1991, 'How to Make Cognitive Illusions Disappear: Beyond 'Heuristics and Biases'', *European Review of Social Psychology* **2**, 83-115.

Ginet, C.: 1975, *Knowledge, Perception, and Memory,* D. Reidel, Dordrecht.

Goldman, A. H.: 1988, *Empirical Knowledge,* University of California Press, Berkeley.

Goldman, A. I.: 1986, *Epistemology and Cognition,* Harvard University Press, Cambridge.

Goldman, A. I.: 1989, 'Replies to the Commentators', *Philosophia* **19**, 301-324.

Goldman, A. I.: 1993a, 'Epistemic Folkways and Scientific Psychology', in Goldman, 1993c.

Goldman, A. I.: 1993b, *Philosophical Applications of Cognitive Science,* Westview Press, Boulder, Co.

Goldman, A. I. (ed.): 1993c, *Readings in Philosophy and Cognitive Science,* MIT Press, Cambridge.

Gopnik, A.: 1993, 'How We Know Our Minds: The Illusion of First-Person Knowledge of Intentionality', in Goldman, 1993b.

Goodman, N.: 1965, *Fact, Fiction, and Forecast*, 2nd edn, Bobbs-Merrill, Indianapolis.

Graubard, S. R.: 1988, *The Artificial Intelligence Debate: False Starts, Real Foundations,* MIT Press, Cambridge.

Green, D. W. et al.: 1996, *Cognitive Science: An Introduction,* Blackwell, Oxford.

Gregory, R.: 1970, *The Intelligent Eye,* McGraw-Hill, New York.

Griggs, R. A.: 1983, 'The Role of Problem Content in the Selection Task and in the THOG Problem', in Evans, 1983.

Grimson, W. Eric and R. S. Patil (eds.): 1987, *AI in the 1980s and Beyond,* MIT Press, Cambridge.

Guttenplan, S. L. ed.: 1975, *Mind and Language,* Clarendon Press, Oxford.

Hahn, L. E. and P. A. Schilpp (eds.): 1986, *The Philosophy of W. V. Quine,* Open Court, LaSalle, Il.

Hanson, N. R.: 1961, *Patterns of Discovery,* Cambridge University Press, Cambridge.

Harman, G.: 1973, *Thought,* Princeton University Press, Princeton.

Harman, G.: 1986, *Change in View: Principles of Reasoning*, MIT Press, Cambridge..

Heckerman, D.: 1986, 'Probabilistic Interpretations for MYCIN's Certainty Factors', in Kanal and Lemmer, 1986.

Heil, J. (ed.): 1993, *Rationality, Morality, and Self-Interest,* Rowman and Littlefield, Lanham, Md.

Hempel, C. G.: 1965a, *Aspects of Scientific Explanation,* Free Press, New York.

Hempel, C. G.: 1965b, 'Inductive Inconsistencies', in Hempel, 1965a.

Holland, J. H., K. J. Holyoak, R. E. Nisbett, and Paul R. Thagard: 1986, *Induction: Processes of Learning, Inference, and Discovery,* MIT Press, Cambridge.

Hookway, C.: 1990, *Scepticism,* Routledge, London.

Hookway, C. and D. Peterson (eds.): 1993, *Philosophy and Cognitive Science,* Cambridge University Press, Cambridge.

Hume, D.: 1974, *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, L. A. Selby-Bigge (ed.), 3rd edn, P. H. Nidditch (ed.), Oxford University Press, Oxford.

Jennings, D. L., T. M. Amabile, and L. Ross: 1982, 'Informal Covariation Assessment: Data-based versus Theory-based Judgments', in Kahneman, Slovic, and Tversky, 1982.

Johnson-Laird, P. N.: 1983, *Mental Models,* Harvard University Press, Cambridge.

Johnson-Laird, P. N.: 1988, *Computers and the Mind: An Introduction to Cognitive Science,* Harvard University Press, Cambridge.

Johnson-Laird, P. N. and R. M. J. Byrne: 1991, *Deduction,* Erlbaum, Hillsdale, N. J..

Johnson-Laird, P. N. and P. C. Wason (eds.): 1977, *Thinking: Readings in Cognitive Science,* Cambridge University Press, Cambridge.

Kahneman, D. and A. Tversky: 1979, 'Prospect Theory: An Analysis of Decision under Risk., *Econometrica* **47**, 263-291.

Kahneman, D., P. Slovic, and A.Tversky eds.: 1982, *Judgment Under Uncertainty: Biases and Heuristics,* Cambridge University Press, Cambridge.

Kanal, L. N. and J. F. Lemmer (eds.): 1986, *Uncertainty in Artificial Intelligence,* North-Holland, Amsterdam.

Karmiloff-Smith, A.: 1992, *Beyond Modularity,* MIT Press, Cambridge.

Kelly, K. T.: 1995, *The Logic of Reliable Inquiry,* Oxford University Press, Oxford.

Kim, J.: 1988, 'What is 'Naturalized Epistemology'?', in Tomberlin, 1988.

Konolige, K., G. Brewka, and J. Dix: 1996, *A Tuturial on Nonmonotonic Reasoning,* Cambridge University Press, Cambridge.

Koriat, A.: 1994, 'Memory's Knowledge of Its Own Knowledge: The Accessibility Account of the Feeling of Knowing', in Metcalfe and Shimamura, 1994.

Kornblith, H.: 1983, 'Justified Belief and Epistemically Responsible Action', *Philosophical Review* **92**, 33-48.

Kornblith, H.: 1993, *Induction and Its Natural Ground: An Essay in Naturalistic Epistemology*, MIT Press, Cambridge.

Kornblith, H.: 1994a, 'Introduction: What is Naturalistic Epistemology?', in Kornblith, 1985b.

Kornblith, H. (ed.): 1994b, *Naturalizing Epistemology* 2nd edn, MIT Press, Cambridge.

Kuhn, T.: 1962, *The Structure of Scientific Revolutions,* University of Chicago Press, Chicago.

Kyburg, H.: 1991, 'Normative and Descriptive Ideals', in Cummins and Pollock, 1991.

Lehrer, K.: 1974, *Knowledge,* Oxford University Press, Oxford.

Lehrer, K.: 1990, *Theory of Knowledge,* Westview Press, Boulder, Co..

Lenat, D. B., M. Prakash, and M. Shepherd: 1986, 'CYC: Using Common Sense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks', *AI Magazine* **6**, 65-85.

Lichtenstein, S. and B. Fischhoff: 1977, 'Do Those Who Know More Also Know More About How Much They Know? The Calibration of Probability Judgments', *Organizational Behavior and Human Performance* **20**, 159-183.

Lichtenstein, S. and B. Fischhoff: 1980, 'Training for calibration', *Organizational Behavior and Human Performance* **26**, 149-171.

Lichtenstein, S., B. Fischhoff, and D. Phillips: 1982, 'Calibration of Probabilities: The State of the Art to 1980', in Kahneman, Slovic, and Tversky, 1982.

Lipton, P.: 1991, *Inference to the Best Explanation,* Routledge, London.

Lucas, J. R.: 1961, 'Minds, Machines and Godel', *Philosophy* **36**, 120-124.

Luper-Foy, S.: 1987, *The Possibility of Knowledge: Nozick and His Critics,* Rowman and Littlefield, Totowa, N. J.

Lycan, W. G.: 1988, *Judgement and Justification,* Cambridge University Press, Cambridge.

MacCrimmon, K. R.: 1968, 'Descriptive and Normative Implications of the Decision-theory Postulates', in Borch and Mossin, 1968.

Maffie, J.: 1990, 'Recent Work on Naturalized Epistemology', *American Philosophical Quarterly* **27**, 281-294.

Maffie, J.: 1995, 'Towards an Anthropology of Epistemology', *The Philosophical Forum* **3**, 218-241.

Maher, P.: 1993, *Betting on Theories,* Cambridge University Press, Cambridge.

Metcalfe, J. and A. P. Shimamura: 1994, *Metacognition: Knowing about Knowing,* MIT Press, Cambridge.

Meyer, J.-A.: 1996, 'Artificial Life and the Animat Approach to Artificial Intelligence', in Boden, 1996.

Millikan, R. G.: 1984, 'Naturalist Reflections on Knowledge', *Pacific Philosophical Quarterly* **65**, 315-334.

Miner, A. C. and L. M. Reder: 1994, 'A New Look at Feeling of Knowing: Its Metacognitive Role in Regulating Question Answering', in Metcalfe and Shimamura, 1994.

Mitchell, T., R. Keller, and S. Kedar-Cabelli: 1986, 'Explanation-based Generalization: A Unifying View', *Machine Learning* **1**, 47-80.

Montmarquet, J. A.: 1993, *Epistemic Virtue and Doxastic Responsibility,* Rowman and Littlefield, Lanham, Md.

Neufeld, E. and D. Poole: 1988, 'Probabilistic Semantics and Defaults', *Proceedings of the 4th AAAI Workshop in Uncertainty in AI,* Minneapolis.

Nisbett, R., and L. Ross: 1980, *Human Inference: Strategies and Shortcomings of Social Judgment,* Prentice-Hall, Englewood Cliffs, N. J.

Nisbett, R. and T. D. Wilson: 1977, 'Telling More Than We Can Know: Verbal Reports on Mental Processes', *Psychological Review* 84, 231-259.

Pearl, J.: 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference,* Morgan Kaufmann, San Mateo, Ca.

Pearl, J.: 1990, 'Probabilistic Semantics for Nonmonotonic Reasoning', in Cummins and Pollock, 1990.

Perry, J.: 1980, 'Belief and Acceptance', in French, Uehling, and Wettstein, 1980.

Plantinga, A.: 1993, *Warrant and Proper Function*, Oxford University Press, Oxford.

Pollock, J.: 1974, *Knowledge and Justification,* Princeton University Press, Princeton.

Pollock, J.: 1987, 'Defeasible Reasoning', *Cognitive Science* **11**, 481-518.

Pollock, J.: 1989, *How to Build a Person: A Prolegomenon*, MIT Press, Cambridge.

Pollock, J.: 1990a, 'Interest-driven Suppositional Reasoning', *Journal of Automated Reasoning* **6**, 419-462.

Pollock, J.: 1990b, *Nomic Probability and the Foundations of Induction,* Oxford University Press, Oxford.

Pollock, J.: 1991, 'OSCAR: A General Theory of Rationality', in Cummins and Pollock, 1990.

Port, R. F. and T. van Gelder: 1995, *Mind as Motion: Explorations in the Dynamics of Cognition*, MIT Press, Cambridge.

Putnam, H.: 1992, *Renewing Philosophy,* Harvard University Press, Cambridge.

Pylynshyn, Z. W.: 1984, *Computation and Cognition: Toward a Foundation for Cognitive Science*, MIT Press, Cambridge.

Quine, W. V.: 1969a, 'Epistemology Naturalized', in W. V. Quine, 1969b.

Quine, W. V.: 1975, 'The Nature of Natural Knowledge', in Guttenplan, 1975.

Quine, W. V.: 1969b, *Ontological Relativity and Other Essays*, Columbia University Press, New York.

Quine, W. V.: 1974, *The Roots of Reference,* Open Court, LaSalle, Il.

Quine, W. V.: 1986, 'Reply to Morton White', in Hahn and Schilpp, 1986.

Quine, W. V.: 1995, *From Stimulus to Science,* Harvard University Press, Cambridge.

Ramsey, W., S. Stich, and D. E. Rumelhart: 1991, *Philosophy and Connectionist Theory*, Erlbaum, Hillsdale, N. J..

Rawls, J.: 1971, *A Theory of Justice,* Harvard University Press, Cambridge.

Rips, L. J.: 1983, 'Cognitive Processes in Propositional Reasoning', *Psychological Review* **90**, 38-71.

Rips, L. J.: 1988, 'Deduction', in Sternberg and Smith, 1988.

Rosch, E. and B. B. Lloyd (eds.): 1978, *Cognition and Categorization,* Erlbaum, Hillsdale, N. J.

Rumelhart, D. E. and J. L. McClelland (eds.): 1986, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* 2 vols, MIT Press, Cambridge.

Savage, L. J.:1954, *The Foundations of Statistics,* Wiley, New York.

Schmitt, F. F.: 1992, *Knowledge and Belief,* Routledge, London.

Schmitt, F. F.: 1993, 'Epistemic Perspectivism', in Heil, 1993.

Schmitt, F. F.: 1994a, 'Naturalizing Epistemology: A Bibliography', in Kornblith, 1994b.

Schmitt, F. F.: 1994b, *Socializing Epistemology: The Social Dimensions of Knowledge,* Rowman and Littlefield, Lanham, Md..

Schwartz, B. L. and J. Metcalfe: 1994, 'Methodological Problems and Pitfalls in the Study of Human Metacognition', in Metcalfe and Shimamura, 1994.

Sellars, W.: 1963a, 'Empiricism and the Philosophy of Mind., in Sellars, 1963b.

Sellars, W.: 1963b, *Science, Perception, and Reality,* Routledge and Kegan Paul, London.

Sextus Empiricus: 1976, *Sextus Empiricus* 1 Outlines of Pyrrhonism, trans. R. G. Bury, Harvard University Press, Cambridge.

Shafer, G.: 1990, 'Savage Revisited', in Shafer and Pearl, 1990.

Shafer, G. and J. Pearl: 1990, *Readings in Uncertain Reasoning,* Morgan Kaufmann, San Mateo, Ca.

Shatz, D.: 1987, 'Nozick's Conception of Skepticism', in Luper-Foy, 1987.

Shatz, D.: 1993, 'Skepticism and Naturalized Epistemology', in Wagner and Warner, 1993.

Shortliffe, E. H. and B. G. Buchanan: 1975, 'A Model of Inexact Reasoning in Medicine', *Mathematical Biosciences* **23**, 351-379.

Shrager, J. and P. Langley (eds.): 1990, *Computational Models of Discovery and Theory Formation,* Erlbaum, Hillsdale, N. J.

Siegel, H.: 1980, 'Justification, Discovery and the Naturalizing of Epistemology', *Philosophy of Science* **47**, 667-676.

Simpson, J. A. and D. Kenrick (eds.): 1996, *Evolutionary Social Psychology,* Erlbaum, Hillsdale, N. J.

Smith, E. E. and D. Medin: 1981, *Categories and Concepts,* Harvard University Press, Cambridge.

Sober, E.: 1978, 'Psychologism," *Journal for the Theory of Social Behavior* 8, 165-191.

Stein, E.: 1996, *Without Good Reason: The Rationality Debate in Philosophy and Cognitive Science,* Clarendon Press, Oxford.

Sternberg, R. J. (ed.): 1988, *Advances in the Psychology of Human Intelligence* vol. 4, Erlbaum, Hillsdale, N. J..

Stich, S. P.: 1993, *The Fragmentation of Reason: Preface to a Pragmatic Theory of Cognitive Evaluation,* MIT Press, Cambridge.

Stich, S. P. and Richard E. Nisbett: 1980, 'Justification and the Psychology of Human Reasoning,' *Philosophy of Science* **47**, 188-202.

Stickel, M. E.: 1985, 'Schubert's Steamroller Problem: Formulations and Solutions', *Journal of Automated Reasoning* **2**, 89-101.

Stillings, N. A., Steven E. Weisler, C. H. Chase, M. H. Feinstein, J. L. Garfield, and E. L. Rissland: 1995, *Cognitive Science: An Introduction,* MIT Press, Cambridge.

Strawson, P. F.: 1985, *Scepticism and Naturalism,* Columbia University Press, New York.

Stroud, B.: 1984, *The Philosophical Significance of Scepticism,* Oxford University Press, Oxford.

Stroud, B.: 1989, 'Understanding Human Knowledge in General', in Clay and Lehrer, 1989.

Thagard, P.: 1988, *Computational Philosophy of Science,* MIT Press, Cambridge.

Thagard, P. 1989a, 'Explanatory Coherence', *Behavioral and Brain Sciences* **12**, 435-467.

Thagard, P. 1989b, 'Connectionism and Epistemology: Goldman on Winner-take-all Networks', *Philosophia* **19**, 189-196.

Thagard, P. 1990, 'The Conceptual Structure of the Chemical Revolution', *Philosophy of Science* **57**, 183-209.

Thagard, P. 1991, 'The Dinosaur Debate: Explanatory Coherence and the Problem of Competing Hypotheses', in Cummins and Pollock, 1991.

Thagard, P. 1992, *Conceptual Revolutions,* Princeton University Press, Princeton.

Thagard, P. and G. Nowak: 1988, 'The Explanatory Coherence of Continental Drift', in Fine and Leplin, 1988.

Thagard, P. and G. Nowak: 1990, 'The Conceptual Structure of the Geological Revolution', in Shragger and Langley, 1990.

Thelen, E. and L. B. Smith: 1994, *A Dynamic Systems Approach to the Development of Cognition and Action,* MIT Press, Cambridge.

Tomberlin, J. E. (ed.): 1988, *Philosophical Perspectives 2: Epistemology,* Ridgeview, Atascadero, Ca..

Tversky, A.: 1969, 'Intransitivity of Preferences', *Psychological Review* **76**, 31-48.

Tversky, A.: 1975, 'A Critique of Expected Utility Theory', *Erkenntnis* **9**, 163-173.

Tversky, A. and D. Kahneman: 1971, 'Belief in the Law of Small Numbers', in Kahneman, Slovic, and Tversky, 1982.

Tversky, A. and D. Kahneman: 1983, 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment', *Psychological Review* **90**, 293-315.

von Eckhardt, B.: 1993, *What Is Cognitive Science?,* MIT Press, Cambridge.

Wagner, S. J. and R. Warner (eds.): 1993, *Naturalism: A Critical Appraisal*, University of Notre Dame Press, Notre Dame.

Wason, P. C.: 1966, 'Reasoning', in Foss, 1966.

Wason, P. C.: 1977, 'Self-contradictions', in Johnson-Laird and Wason, 1977.

Wason, P. C.: 1983, 'Realism and Rationality in the Selection Task', in Evans, 1983.

Wason, P. C. and P. G. Brooks: 1979, 'THOG: The Anatomy of a Problem', *Psychological Research* **41**, 79-90.

Wellman, H. M. and S. A. Gelman: 1988, 'Children's Understanding of the Nonobvious', in Sternberg, 1988.

Wilson, D. S.: 1996, 'Incorporating Group Selection into the Adaptationist Program: A Case Study Involving Human Decision-making', in Simpson and Kenrick, 1996.

Winblad, Douglas: 1989, 'Skepticism and Naturalized Epistemology', *Philosophia* **19**, 99-113.

DAVID BLOOR

SOCIOLOGY OF SCIENTIFIC KNOWLEDGE

The work of sociologists of knowledge and socially oriented historians of science should be of interest to epistemologists for one clear and overriding reason. It furnishes a theory of knowledge which exhibits knowing as a social process, and knowledge as a collective accomplishment. Such a claim should not be underestimated. The sociology of knowledge challenges much that has been put forward in the name of epistemology. There are a number of dimensions along which that challenge proceeds. First, the work, which has generated a social conception of knowledge is concrete rather than abstract. All too often philosophers have distanced themselves from the contingencies of real, historical cases in favour of logical formalism and displays of technical virtuosity. Second, the sociological approach is naturalistic rather than normative. The word 'normative' is not the opposite of 'naturalistic', but one way to evade the discipline of naturalistic enquiry is to retreat from the world of fact, the 'is', into a world of unsituated 'oughts', ideals and free-floating values. Concern with how a 'true' or 'rational' scientist ought to behave can be an excuse for avoiding the question of how actual passages of scientific work proceed. Third, and most important of all, the sociology of knowledge challenges the widely held individualism that permeates epistemology.

Over the past twenty years the sociology of scientific knowledge has grown apace. Historical, empirical and theoretical monographs have appeared – and have frequently been met with deep opposition. All too often that opposition has been grounded in a misunderstanding of the claims and conclusions that have been advanced. The causes of the misunderstanding are not far to seek. They are to be found in the three major differences just identified. In attempting to absorb work based on such different premises, critics have mistaken the new approach for an incompetent contribution to their own enterprise, rather than an attempt to reconstruct epistemology along new lines. The detailed discussion of examples has been seen as a failure to clear away contingencies and rise to the level where the real issues are to be found. Again, the naturalistic orientation has seemed like a failure to deliver the required evaluations – and hence as toleration of work that should be condemned. Finally, the rejection of individualism has been experienced as a rejection of the essential attributes of cognition itself. Given this breakdown of communication, which I shall illustrate, it seems appropriate to go over the basic ground slowly and carefully. I shall not, however, offer a general survey of the empirical or historical literature in the sociology of scientific knowledge.[1] I shall mention case-studies, and look at one in some detail, but my discussion will focus on the epistemological lessons that can be drawn from such work. Nevertheless, as a preliminary, it may be useful to differentiate the field which is of special concern to

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 919-962.
© 2004 *Kluwer Academic Publishers.*

us from the more general category of work that falls under the heading of the sociology of science.

Sociologists of science deal with themes such as the funding of science, the distribution of rewards and resources, refereeing practices and patterns of citation, recruitment and career trajectories and the general 'ethos' and 'value system' of the scientific profession. For a useful survey of the literature on these topics see Zuckerman 1988. One of the pioneers of the sociology of science was R.K. Merton, and his influential work may be taken as representative of many of the assumptions and preoccupations of the field. Merton began his involvement, in the 1930s, with a study of 17th C. science. He sought to explain the foci of scientific enquiry and to document the stimulus to scientific development provided by Protestant theology. He was led to an emphasis on the 'norms' and 'values' of science which, he argued, continued to inform science and help further the institutional goal of accumulating reliable knowledge. (For an account of the place of Merton's work in the ideological context of American academic life see Hollinger 1995.)

Merton largely took for granted that, in the proper functioning of the institution, the rational appraisal of evidence and the testing of theories were autonomous processes. The inner, rational core of scientific thinking was not itself social. Thus, he would routinely contrast the rational and social properties of science. In a paper on the interaction of science and military technique, first published in 1935, he wrote characteristically that the, "foci of scientific interest are determined by social forces as well as the immanent development of science". (Merton 1973, 204.) Notice that there are two things here, "social forces" and "immanent" developments. The immanent development of science, on this perspective, is helped or hindered, but not constituted by, the way society (and science itself) is organised.

Such a position is widespread in both sociology and philosophy. Indeed it might be called the standard position with regard to the sociology of knowledge. A classic statement is to be found in Karl Mannheim's *Ideology and Utopia* (1936). According to Mannheim, it is only when, "the process of knowing does not actually develop historically in accordance with immanent laws" that we can conclude that "extra-theoretical factors" and "existential determination" have been at work (p.239). Clearly Merton's social forces are Mannheim's extra-theoretical, existential determinants. They are things which can inhibit or facilitate the working of science's inner logic, that is, its immanent development. The tradition was continued by Lakatos's 1971 distinction between the (rational) "internal" and (non-rational) "external" history of science. Only the latter, he asserted, is amenable to socio-psychological, causal explanation. More recently still the structure has been re-iterated by Laudan (1977). All of the different developments in the sociology of scientific knowledge, whether contemporary laboratory studies, or sociologically informed historical work, have been devoted to overcoming this dualistic, indeed Manichean, conception of knowledge.[2]

Merton's concern in the 1930s and 40s with the norms and values of science was presented, at least in part, as a response to the anti-Semitism and racial pseudo-science of the Hitler regime. The writings of the German physicists Stark and Lennard are cited in this connection. These Nobel prize- winning experimentalists called for an Arian, as distinct from a Jewish, science. In opposition to such tendencies Merton draw attention to those aspects of the scientific tradition that embodied the norm of 'universalism', i.e. the requirement that knowledge claims be assessed by impersonal, general criteria. To entertain doubts about relativity theory, simply because its inventor was Jewish, would be a clear violation of such a principle.[3] These political circumstances surrounding the emergence of Merton's 'functionalist' picture of science has cast a long shadow over subsequent discussions. It is still difficult to disentangle theoretical issues in the sociology of knowledge from past anxieties about the autonomy of scholarship and the treatment of scientists under totalitarian regimes.

It may help to offset these fears if I show how the sociology of knowledge grows naturally out of a treatment of science whose credentials are, I hope, beyond reproach. I shall take as my starting point a highly sophisticated work in the philosophy of science, written by a physicist, and use it to set the stage for the sociological approach. The work in question is Pierre Duhem's 1906 classic *The Aim and Structure of Physical Theory.* I am not assuming that everyone thinks Duhem's work is right. His ideas are not to be accepted uncritically and, indeed, it will rapidly prove necessary to move beyond his formulations. Nevertheless, reference to this work should ensure that all parties to the discussion are speaking the same language.[4]

I THE SOCIAL STRUCTURE OF PHYSICAL THEORY

Duhem argued, famously, that no hypothesis in physics could be tested in isolation. (He distinguished physics from more descriptive enterprises such as physiology in this respect, but I shall follow the many precedents which rightly treat the argument as quite general.) His point was that every test of an hypothesis H involves further, auxiliary hypotheses A, concerning the initial conditions of the test, the working of the test apparatus and any instruments it involves. Chemical experiments involve assumptions about the purity of the chemicals, while physical experiments require assumptions about the closure of the system, its thermal, electrical or magnetic isolation, or the working of instruments such as interferometers or cathode ray tubes. If a prediction about the experimental outcome drawn from H appears to be wrong, then the locus of blame is unclear. The experimenter is dealing with the logical conjunction A and H. The negation of a conjunction is a disjunction, so which should be negated, H or A? Logically, either conclusion could be drawn. It may be possible to rescue H by adjusting the auxiliary assumptions, replacing A by A*, so as to reconcile theory and observation. H has not been absolutely *proven* false until it has been absolutely *proven* that no such rescue is possible.

The Duhem argument is a proposal about the burden of proof and therefore, in practical terms, it involves us in the question of 'proof for whom?' It takes us into the realm of credibility and all the considerations which bear on the issue of which persons find which propositions convincing, and why they react in the way they do. It exposes the scope for acts of choice and discretion, and hence raises the question of why choice is exercised as it is. Why was Ptolemy rejected in favour of Copernicus, at that time and place, given that the Copernican system failed to fit in with current theories of dynamics and also needed propping up by appeal to lots of epicycles? Why did the fluid theory of heat take over from the kinetic theory in the

18th century, only to be displaced again by that approach in the 19th? Why did central European mathematicians experience a sense of crisis in the foundations of mathematics after World War I? Brouwer had been presenting the case for intuitionism for a number of years. Why suddenly take notice now?

If such questions are addressed in a properly historical fashion, as they should be, we can see how even skeletal and abstract issues can lead us towards the sociology of knowledge. Just because Duhem was raising a 'logical' point about hypothesis testing does not mean that he was not at the same time, and with those same words, raising a sociological issue. One of the most important lessons to be learned from the sociology of knowledge is that we must be beware of false dichotomies. The idea that, in general, logical questions must be distinguished from sociological questions is just one such false dichotomy. We shall soon encounter others.

Does Duhem's argument mean that scientists could always rescue an hypothesis, or that they could render any hypothesis consistent with any data, by adjusting the auxiliary assumptions? No it does not. Duhem was drawing attention to an important, holistic feature of our system of knowledge, but he was not presenting the overall structure as undifferentiated or its adjustment as effortless. On the contrary, the physicist's room for manoeuvre in reconciling a given hypothesis with accepted data is a limited one. It is limited by the current horizon of understanding, by human ingenuity and by the credibility of any new auxiliary hypothesis that may be proposed.

Duhem identified these limitations in different terms to those I have just used. He referred to what he called "good sense", and (provocatively) to the "faith" the scientific community has in certain theoretical assumptions. Consider the point about faith first. Any test must take something for granted. If, when a prediction is wrong, suspicion should fall on something previously taken for granted then it, too, can be tested, but then that test must also take something else for granted. Not everything can be independently tested, and in practice tests stop at ideas, assumptions, theories and procedures which have come to be routinely accepted without such tests. It is this taken for granted element that Duhem called "faith". Good sense is less sharply defined. Duhem spoke of it as a variable thing, but expressed this variability in terms of a differential access to the truth of the matter, as if there may be a fixed ideal of good sense, but some people can intuit it better than others. It is difficult to know how seriously to take this idiom, and it is at this point that we can begin to improve on his account by offering a naturalistic and causal reading of the constraints he expressed in intuitive terms.

A plausible reading of Duhem is that good sense involves a responsiveness to the costs and benefits of trying to rescue a hypothesis. If a physicist were so wedded to an hypothesis that he was led to make numerous *ad hoc* adjustments in the accepted laws of physics, perhaps treating the laws of optics as different in the vicinity of his apparatus, or giving the gravitational constant a different local value, this may be deemed a lack of good sense. The pay-off would have to be very great to justify the modification. Why? Because what is being compromised is the coherence of knowledge, and in practice that means the possibility of different physicists making use of one another's work. It refers to the ease with which they can communicate, the manner in which their efforts are co-ordinated, and the extent to which they are participants in a collective enterprise whose outcome can exist as a body of shared culture. A carefully crafted structure of classifications and permissible inferences will be torn apart. What is at stake for the group of persons involved is whether they can sustain the collective good we call science. Without this co-ordination all we should have, as Kuhn later pointed out, would be persons who might be recognised as scientists, but something less than science (Kuhn 1962, 13).

Duhem's good sense and faith can be identified as social phenomena. The credibility attaching to different responses to the anomalies thrown up by experiment and observation is an index of a social process. Just as costs only make sense in the context of a market, so credibility only makes sense when the scientist's behaviour is set in its social context. What Duhem identified in terms of common-sense psychology and religious metaphor I have re-identified as a social phenomenon, and as a property of the social context. Notice that context has been defined here in a purely internal manner, that is, as one concerning the scientist, *qua* scientist, relating to other scientists, but it is a social context none the less. It is social in that it concerns interaction, co-ordination, co-operation and collective action. This is not a weak or trivial sense of 'social'; it is the most basic and root sense of the word. Clearly a great deal that is also social has been left out of the story, for example, how different groups of scientists may attach different values to the costs and benefits of any given proposal, or how an interest in its preservation will crystallise around any stable element of culture. I have also ignored the contingencies connecting what I have called "internal" social relationships with a broader spectrum of interactions. Despite these omissions the foundation for a richer and more detailed analysis has been laid down.

II CONVENTIONALISM

It is interesting to reflect on how Duhem's insights have been absorbed into the mainstream of epistemological thinking. All competent epistemologists know their Duhem, but what is it they know? Many must see in Duhem exactly what Popper saw, namely an astute delineation of the processes by which a piece of knowledge may be protected from criticism. For Popper, Duhem showed us how we can do something we should not do, namely, turn a piece of putative empirical knowledge into a mere convention by resorting to a systematic strategy of protection. Duhem's argument about the logic of experimental testing is seen as an argument about the evasion of testing. On this view, Duhem has revealed something to us about the pathology of knowledge. Turning something into a convention is set in contrast to the proper conduct of our cognitive business. (Popper 1959, ch.iv.)

Feyerabend's early work had a similar general form to that of Popper's. Feyerabend (1963) was disturbed by the widely accepted tendency he detected, both in science and philosophy, to operate with a highly conservative consistency principle. This took the form of a requirement that new knowledge claims be consistent with existing and accepted knowledge. New developments should not disturb existing achievements. The consistency requirement, argued Feyerabend, was simply a human policy whose effect gave current beliefs a spurious appearance of fitting the facts. The fit was really an artefact of the politics of knowledge and violated the sound, old empiricist requirement that knowledge claims be exposed to criticism by comparison with experience. It resulted in a self-serving gloss being put on all the deliverance's of experiment and observation. Once again, conventionalisation was seen as the villain of the piece.

To cure the problem, Feyerabend recommended that scientists be encouraged to cultivate a plurality of alternative theories. This would offset the conventionalisation, or canonisation, of any single understanding of nature. The dialectic of his early argument points clearly towards Feyerabend's later position. Passing through the phase of pluralism he was led to the advocacy of epistemological anarchy (Feyerabend 1988. 'Anarchy' means dissolving conventional structures and hierarchies and operating on the level of individual spontaneity. It has always been a seductive social form, but it can only be achieved at a cost. All the benefits of organisation and co-ordination must be sacrificed but, for some, the price seems worth paying.

A variant on these themes has been proposed by Helen Longino (1992). Her demonstration of the advantages of encouraging feminist approaches to science depends on the premise that pluralism is, in itself, not merely a cognitive virtue, but the overriding cognitive virtue. This premise is never stated explicitly, but her general argument has no plausibility without it. Certainly the benefits of pluralism are easily appreciated. New eyes looking at things from a new point of view can help overcome old blind-spots. New voices speaking on behalf of new groups will offer perspectives that were unwittingly, or deliberately, ignored in the past. All of these virtues are real ones. There is always a case to be made for pluralism and democracy in the realm of the mind, but that argument has to be a local one. At any given time it may be wise to go for more pluralism; but at other times the need may be to close ranks and increase coherence. Treating either strategy as a universal principle is folly, but this is exactly what Longino has done, tacitly, for pluralism.

In presupposing the virtues of pluralism Longino gives no consideration to the possibility that an increased sensitivity to new voices, and a corresponding insensitivity to old ones, might result in the loss of knowledge. What if the novelties are merely novel errors or old errors re-cycled? And what if the tradition embodies important but unfashionable truths? The particular case has to be argued with particular reasons, not general presumptions. Longino mentions specific feminist criticisms of work in primatology and biology, but these function merely as illustrations of her general claim. There is, however, no principle which guarantees that we will gain the advantages, and avoid the penalties, of pluralism and participation.[5]

Whatever insights one might want to attribute to the work of Popper, Feyerabend, and, in a somewhat different key, Longino, they cannot provide us with the appropriate sensibility for responding to Duhem. At best, they will lead to a one-sided appreciation of what he can tell us. The rationalistic emphasis on pluralism, diversity and criticism, obscures the generality of the conventional machinery that Duhem identified.[6] We do not just conventionalise things that we want to protect from criticism. Conventionalisation is a ubiquitous process which also, and necessarily, plays a central role in exposing parts of our knowledge to criticism. It is the fulcrum on which criticism turns. A more sympathetic reading of Duhem, such as that given by Hesse (1974), makes the generality and utility of conventionalisation abundantly clear.

To appreciate this generality, and free ourselves from a one-sided picture of the role of the social, we may appeal to the idea that all theories are born refuted. All theories face trouble and anomaly all the time. Newtonian mechanics couldn't be reconciled with the motion of the moon for some sixty years. The predicted motion of the perigee was half that observed (Kuhn 1962, 81). Lavoisier's oxygen theory of combustion and acidity could not explain the behaviour of what was called marine acid gas, our hydrochloric acid (Conant 1966). And Einstein's work on relativity immediately ran into Kaufmann's experiments (Hon 1995) – as we shall see in more detail. Expressing things in this way reminds us that it may be very good to protect a theory from refutation, because if we did not we should have nothing left at all. The world is vastly complex and, by comparison, even our best theories are schematic and simplified. So conventionalisation is needed to sustain truth as well as, on occasion, protect error. Conventionalisation is an epistemic virtue as well as epistemic vice. In itself it is a neutral, ubiquitous feature of all systems of belief.

Conventionalising a body of ideas or concepts is simply the expression of a shared willingness to use them. This readiness to deploy ideas as a shared resource to cope with (inevitable) difficulties is the key phenomenon to which any epistemology must do justice. On a micro-level it may be pictured as follows: Person A is prepared to use theory T partly because it seems to be the best available, but also partly because of the knowledge that persons B, C and D (and so on) themselves routinely use it. Indeed, the knowledge that B, C and D use it, is itself part of what its appearing to be the best available theory consists in. If B, C and D abandoned it that would count as a reason for A abandoning it. Perhaps B, C and D know something that A doesn't know. In any case, A wants to proceed in a fashion which allows for the exchange of information and the possibility of co-operation and co-ordination of effort. This description of the behaviour of A, B, C and D simplifies the processes involved, making it look calculating when, in fact, it is largely trusting (see Hardwig 1991). On some level it must also be much more instinctive than I have presented it. Perhaps A would have an instinctive tendency to abandon T, if B, C and D abandoned it, rather as one bird takes to the air if the rest of the flock do (see Haugeland 1990). But, whatever its shortcomings, the picture captures the strategic element in concept use, and gives us some idea of the inner structure of conventionalisation.

A recognisably more realistic account, and one which brilliantly brings out the positive and not just the negative aspects of conventionalisation, is Kuhn's well-known description of science as a paradigm-based activity. (See Kuhn 1962.) Clearly, being a paradigm is not an intrinsic property. It is not, as Hume would put it, a "natural" but an "artificial" property, something collectively accorded to an achievement by a group of practitioners. Being a paradigm is a social status. It will be sustained by cycles of strategic reasoning, or by behaviour having a corresponding structure, rather the way in which a currency is sustained. At the root of all these phenomena are self-reinforcing, self-referring, sets of assumptions and calculations about what everybody else is assuming and calculating. Another way to state the point is to say that something's being a paradigm means that it is an institution.[7]

III REALITY AND FINITISM

Where does "the world", that is, non-social reality, come into the sociological story that has just been sketched? It should be clear that this account is not a species of Berkeley-like idealism, but one to which any self-proclaimed materialist could assent. The taken-for-granted framework of the approach is one in which human beings, as biological organisms, are interacting with their material environment and trying to reach some manner of collective adaptation to it. Some stable adaptations must be possible, because if this were not so there would be no human organisms in the first place. What is not assumed by the picture and indeed, in keeping with the biological background, what is implicitly denied by it, is that there is any unique, privileged, best or final form of cognitive adaptation. If it is asked whether there can be progress on such a picture, the answer must be that there can be exactly the same kind of progress as there can be in Darwinian evolution. We can follow Kuhn (1962, 170) and connect the idea of progress to moves *away* from specific problems and maladaptations while divorcing it from the idea of moves *towards* a goal. This may be metaphysically unsatisfying, but arguably it suffices for all practical purposes and explains our strong intuitions that scientific knowledge is progressive.

How should this (non-idealistic) standpoint be expressed? Should we say that "facts" are "social", or that they are "real" or "physical"? In ordinary usage the word "fact" is equivocal between (i) a verbalised account or description of a state of affairs and (ii) the state of affairs, 'in itself', which is referred to in the verbalised account. For our purposes it is important to prise apart these two things and maintain a lively awareness of the difference between a state of affairs, and that same event or process understood, described or encoded in a certain way.[8] If a sociologist of knowledge asserts that facts are social this means, or should mean, that any account, description, classification, or theoretically formulated understanding of a state of affairs is social.

This claim must not be trivialised. It will not do to say, "Oh, but all that means is that concepts must have rules of use", and then treat rules as unproblematic or abstract and fixed things. Properly to contextualise concept application, and properly to contextualise rule following, means treating each and every act of concept application, and each and every instance of a rule, as sociologically problematic. Concept application, and rule following, must be thought of as a move from case to case, where every such move, in principle, calls for an exercise of the analyst's curiosity. Following Hesse, sociologists of knowledge call this position 'finitism'. The central claim of meaning finitism is that past usage does not explain the next case. Each act depends on local contingencies, of which past applications are only one factor. Furthermore, no two cases which fall under the same concept are identical. Or, to express the point more fully, whenever things are treated as identical that identity must be seen as (a) a theoretical claim, and hence (b) one which needs to be collectively sustained by a group of concept users. Alternative employments of the concept of identity always lurk in the background. Even if it is overwhelmingly natural or easy to see certain items as identical, that ease and naturalness is also but one contingent factor in the situation. Other considerations could (and often do) over-ride such impressions.[9]

The problematic character of concept application and the need to treat the move from case to case as a social phenomenon is clearly brought out by the historical study of scientific practice. This reminds us of the contingency, negotiation and changing interpretive climate that impinges on the move from past applications of a concept to new applications. Recall the history of the discovery of Neptune, in 1846, as it has been analysed by the Dutch astronomer Pannekoek, (See Pannekoek 1953, and Barnes 1982). The orbit of Uranus differed from that predicted by the application of Newtonian dynamics. Deploying Duhem's strategy to good effect, this was not read as a refutation of Newton but as the basis for predicting another massive, but hitherto undetected, body which caused the perturbations. A planet-like body was duly located close to the position calculated by Adams and Leverrier. This was the predicted, new planet "Neptune" – or was it? The subsequent track of the new body turned out to be somewhat different from that predicted by Adams and Leverrier. The American astronomer Peirce argued that the entity that had been found was therefore not the entity that had been predicted: it was something else that just happened to be in the same place at that moment. Clearly, Wittgenstein was right: sameness is problematic – even when we are dealing with huge pieces of matter like planets. In this case Pannekoek laid out for us an intriguing sketch of the social concerns, about the status of science, and the different interests of European and American astronomers, which contributed to the outcome, that is, the received understanding of the prediction and discovery.

As Hesse (1974) saw, Duhem's insight, into the connection between an hypothesis and its implications, points to a general fact about the connection between concepts and the instances falling under them. Duhem himself was clear on this, which is why he developed his account of hypothesis testing in conjunction with a sophisticated analysis of what he called "theoretical facts". Theoretical facts are verbal and conceptual encodings of real world states of affairs which – with some inevitable loss of information – bring them into a scheme of classification. Just as no hypothesis can be tested in isolation, so no concept can be applied in isolation. Just as a misfit between a prediction and an observation report can be removed by adjusting auxiliary hypotheses, so a misfit between a concept and its instances could be resolved by commentary drawn from the surrounding network of concepts. What works in one direction also works in the other. If misfits can be repaired in this way, satisfactory instances of concept application might be called into question. Understanding the depth of Duhem's argument means seeing that the processes he identified do not just come into play when things go wrong, when hypotheses yield wrong predictions or when instances don't fit concepts. They are equally in play when things go smoothly. They are always in play, and that is why, in principle, every act of concept application requires a sociological scrutiny of the context of use.

We are now at the heart of the sociological argument, having arrived at a sociological picture of concept application. Concepts which are meant to refer to objects in the material world must be such that they can be used rightly or wrongly. To make sense of conceptual content we must make sense of the normative aspects of concept application. Here the sociological picture can offer genuine illumination.

It is the only intelligible, and non-dogmatic, account of normativity that is available. The normative aspects of concept application are consensual. Standards of right and wrong derive from agreement in use. They only exist in and through that use. To exist at all standards must be invoked, cited, employed, referred to, challenged and defended. What has been said about the social status and identity of paradigms, in the Kuhnian sense, applies equally to the employment and past applications of a concept. Just as paradigms act as precedents, so the past applications of a concept act as a precedent for future applications. The correct use of that precedent, that is, applying a concept in a way that is consistent with its meaning, is correct because a group of users agree that it is correct. Without the sociological machinery of interaction there would be no normativity, and without normativity there would be no conceptual content, that is, no meaning. Meaning itself is unintelligible except as a sociological phenomenon.[10]

Duhem's work makes us aware of the holistic and 'distributed' character of knowledge and hence the interlocking (and potentially clashing) considerations to which every knowledge claim must be answerable. It gives us a model that can be applied again and again, at different levels, until we come down to the most basic of all, concept application itself. We are given a picture of knowledge as a structure with its own internal constraints, rather than something standing in a direct, isomorphic or picturing relation to the way the world is. There are, to be sure, very few thinkers who would openly subscribe to any such simple 'reflection of reality' viewpoint, but mere disclaimers cannot prevent this tempting image from silently re-asserting itself under the pressure of argument. The only way to root out its influence is to keep an alternative model firmly before our minds. But when such alternatives are put forward they often meet resistance which only makes sense as an expression of some form of naïve, unmediated realism. Let us look at such a case.

Social theories of knowledge are sometimes rejected on the grounds that if science were contingent on society, in the way sociologists have claimed, it would be impossible to understand how scientists could be as successful as they are in making predictions. This is the basis of an attack on 'social constructivism' made by the physicists Gottfried and Wilson (1997) in the pages of *Nature*. Although they do not articulate the basis of their claim in any detail, the underlying thought follows a path worn smooth by philosophical critics. It is this: if scientists are responding to society (i.e. to their interactions with one another) rather than responding to the (non-social) natural world which is supposed to be their subject matter, then how could they possibly gain a knowledge of the world sufficient to make successful predictions about it? It would be like taking a college class in history and then sitting the examination designed for those who had taken the course in chemistry. Any successes would be either chance or a miracle. It follows that sociologists of knowledge cannot explain how scientific prediction is possible. Prediction is not only possible, it is an impressive aspect of real science, therefore the sociologists must be wrong through and through.

Notice the assumption: either scientists are responding to society or scientists are responding to nature. If this were the choice before us it would indeed be difficult to see how the sociologist could make any sense of the successful adaptation of scientific knowledge to the structure of the non-social world. The error in the criticism is that sociologists are not proposing or assuming that any such either-or structure is applicable to the case in hand. Their claims are not premised on this disjunction. In fact the opposite is the case: their claim is that scientists are able to respond to nature in the way they do because they are responding to one another (to 'society') in the way they are. There is no choice between responding to society and responding to nature. The picture that is actually being advanced is one of responding to nature through society. Society mediates the response to nature because the response is a collective one. Society is not an alternative, it is the vehicle and channel. To repeat: we do not know the world in spite of society nor can we know it without society, rather, we know it (collectively) by means of society. Without society all we should have in the realm of cognition would be an atomised collection of individual efforts and opinions – something vastly weaker and qualitatively different from the social phenomenon we call 'science'. It would be weaker because it would not be cumulative and individual efforts would not be co-ordinated. It would be qualitatively different because the cognition involved would be idiosyncratic, divergent and subjective.[11] It is only through social organisation that it is possible to achieve the sophisticated adaptation to the details of the natural world which is rightly held to be the glory of the empirical sciences.

We need to think about the social structure of knowledge as an enabling and empowering device, rather than as a mere source of distraction. An analogy may help. Consider the eye, or the visual system, as a physiologist or a psychologist might study it. Just as the visual system is the organ of vision so, I want to argue, we should think of society as the organ of cognition. One is individual and the other collective, and one generates cognition with its own special experiential qualities, which contrasts with the absence of any collective mind to entertain a collective experience. These are important disanalogies, but they do not undermine the comparison that I want to make. In both cases a certain structure can be identified as the vehicle for a certain kind of cognition – in the one case vision, in the other case science. It would be an absurdity for the physiologist to trace out the visual system and then conclude that the organism must see in spite of it – as if it would see so much the better if all of this stuff did not get in the way. It would be equally foolish to think that, because the physiological structure of the visual system was made up of more or less solid material, the organism could not see through it, and hence must be seeing it, rather than seeing the world. Nobody makes these gross mistakes. The physiologist might legitimately wonder how the organism sees with this apparatus. It may not be obvious how it works, or how consciousness enters into the story. These puzzles, however, represent an entirely different order of response to the formulations I have been imagining. Remarkably, it is the latter which correspond to the critic's use of the data generated by sociologists. Sociologists delineate a structure of conventions and institutions and seek to show how scientific knowledge is embodied in them. These are, for the sociologist, the vehicle that carries our collective knowing of the world just as, for the physiologist, the visual system is the vehicle which carries the individual, visual experience of the world. Misconstruing the sociologist as saying that agents are responding to society, rather than responding to the world, is like misunderstanding the physiologist of vision as saying that an organism is looking at the back of its own eyes, rather than looking at the world with its eyes.

It is uncontroversial to say that society plays a role in science by directing attention towards certain areas of study. For example, the perceived military potential of aircraft helped channel funds and effort into the development of aerodynamics during the First World War. There were still significant differences of approach in different countries, e.g. the British concentrated research effort into achieving stability and control, while German research focused on improving the aerofoil and refining the mathematical theory of lift and drag.[12] Such differences, though clearly social phenomenon, are akin to the division of labour. They do not, in themselves, challenge any deeply held assumptions about the nature of knowledge. Nobody has much difficulty accepting that society may, as it were, shine a 'searchlight' in a particular direction, or that different social groups may shine their searchlights in different directions. Notice how the searchlight metaphor separates the role played by society (selecting the object of research) from the role played by the process of cognition once the object has been identified. Critics can use the metaphor whilst insisting on a qualitative distinction between the process of target selection and the process of cognitive engagement with the target once it has been located. This makes the metaphor acceptable to those with a non-social conception of the process of knowing and thus puts it at odds with the previous discussion. The point of saying that scientific cognition is a collective process, where society itself is the vehicle for knowledge, is precisely to deny the dualisms that make the searchlight metaphor unchallenging. Society does not merely light up or direct attention to some topic and then hand over to some 'purely rational', or 'non-social' process of knowledge gathering. Sociologists of knowledge reject such a dualism as untenable and insist, along the lines I have indicated, that social processes are integral to all aspects of knowing. There are, however, still many obstacles for a sociology of knowledge to overcome. A particularly significant source of resistance derives from yet another false dichotomy, namely, the 'rational' versus the 'social'.

IV INDUCTION AND CONVENTION

Duhem concentrated on the choice confronting a scientist if a prediction proves wrong. Does the error stem from taking this part of our network of knowledge for granted, or from that part? It may seem that I have jumped from Duhem's logical observation, about the role of choice in responding to anomaly, to the conclusion that the determinants of the choice are social. Are there not obvious non-social explanations? Perhaps we have faith in a theory because it is well confirmed, and we are rational creatures whose intellect is responsive to inductive considerations. This sounds eminently plausible, but I shall argue that it is completely consistent with the sociological claims previously made. This is not because some sources of credibility are social while others are "cognitive". The consistency does not reside in belief having sometimes the one and sometimes the other cause. An eclectic approach is quite wrong. The two aspects of the process are much more closely linked than this. They do not merely take turns, they actually depend on one another. Cognitive mechanisms, such as the tendency to make inductive inferences, do their work against a social base-line, and upon material furnished by prior social processes.

This is a vital point so I shall illustrate it with an historical example. Consider the dispute which took place between 1880 and 1905, between the oceanographers Carl Chun and Alexander Agassiz, over the intermediate oceanic fauna. The episode is described in detail by Mills 1980 and has been discussed by Barnes 1984. It was agreed by oceanographers that there was a layer of fauna specific to the surface of the ocean and another layer concentrated on the ocean floor. Was there a specific layer of fauna at intermediate levels? Chun said yes; Agassiz said no. Both these distinguished scientists launched a number of expeditions, trawled the oceans with carefully constructed nets and detecting devices at the relevant depths, and drew their conclusions on the basis of what they found. Their rival positions were empirical and their inferences inductive, which is why this case is of interest to us. Reducing a complicated and fascinating story to its simplest terms, the upshot of these expeditions was that Chun found evidence for the intermediate oceanic layer of fauna, while Agassiz found none. Clearly, "nature", in the form of the contents of the nets, played a vital role in sustaining the two rival positions. So where does sociology come into the story?

As Barnes points out, the answer to this question lies in two features of the situation. First, nobody simply responded to the totality of evidence, or putative evidence, on offer. In a manner typical of scientific disputes, part of the dispute was about what to count as evidence and what to discount. There are no rules for this, nor any self-evident, simple or 'natural' ways of deciding the matter. There is no higher court of appeal than the consensus of the scientific community itself so, ultimately, the status of the final sample must itself be conventional. Second, consider the categories in terms of which Chun and Agassiz conducted their dispute. For both of them the issue was the existence, or non-existence, of an entity called "the intermediate oceanic layer". They could have said that intermediate fauna exist in some places, where Chun happened to trawl, but not at other places, where Agassiz happened to trawl. They did not say this. They understood themselves to be gathering evidence for and against "the" intermediate layer. This was conceived as a more or less uniform thing, characteristic of oceans as such. While each knew about the other's work, Chun made inferences from the positive results to the probable existence of this layer; and Agassiz generalised from the negative results to its non-existence. Here we have nature and induction playing their allotted roles, but playing them against a background of thought which employed certain shared categories. Chun and Agassiz shared a conception of what they were about, and it was this shared conception which gave their findings their evidential status. Without it these findings could not have stood in a contradictory relation to one another and there could have been no controversy. To speak here of a shared conception means that the idea functioned as one of the conventions of their thinking. Or we could say that the concept of the intermediate oceanic layer functioned as an institution for the group of practitioners who took part in the dispute. This does not mean that they agreed with it on the level of their opinions. Obviously, this was not so: their opinions were in opposition to one another. It means that their opinions, both for and against, kept certain ideas in circulation as the currency of their thought.

Why not say that Chun and Agassiz were simply making a mistake to conduct their dispute in the terms they did? Having thus identified the conventions of their dispute as erroneous, we shall then be tempted to reaffirm another old idea. It looks as if the social is something that gets in the way of knowledge. These temptations should be resisted. First: no explanatory benefit is to be derived from adopting an evaluative stance. Evaluations are cheap and easily come by. They add nothing to the more challenging and interesting search for causes. Second: inductive inferences don't go wrong simply because they are conventionally structured. They could be neither right nor wrong without some conventional underpinning. If it had not been this one, it would have been another. As Barnes points out (p.120), it is vital to avoid the trap of thinking that Chun and Agassiz were being 'conventional' rather than 'rational', or 'conventional' rather than 'inductive'. There is, he says, a "tendency to see appeals to convention as denials of the reasoned character of scientific change" (p.120). There is no 'rather than' here. Inferences cannot be inductive without being conventional. I draw attention to the error because, amongst philosophical critics of the sociology of knowledge, it is pervasive. Later we shall encounter a philosopher making exactly this mistake.

Having illustrated the conventional aspect of induction concretely and in a particular case, let us address the issue more abstractly and generally. Just as Duhem's logical analysis of scientific testing helped highlight social processes, so a logical analysis of inductive reasoning can also serve to highlight its social character. Consider Rudolf Carnap's determined attempt to build a rigorous logic of inductive confirmation. Here, one might think, was an approach that would squeeze the social element down to a minimum. The result was quite the opposite.

Carnap's *Continuum of Inductive Methods* (1952) made it clear that there is no unique inductive method: there is in fact an infinity of them. Carnap's formalistic approach could not handle the complexities of real-life inductive practice, so he built simplified models of the process that were amenable to rigorous statement and mathematical development. He imagined simple "worlds" whose constituent objects had small numbers of properties that could be described in simple, formal languages. These simplifications helped to make it clear that there was no unique way to define the confirmation function c(h,e). This function is given by a formula which allows us to compute a value for c, the degree of confirmation, given a statement h (the hypothesis) and another statement e (the evidence). A number of plausible, candidate definitions of the confirmation relation stood out, but it became clear that the function c depended on more than h and e. It also depended on the parameter Carnap called lambda (λ). Lambda varies between zero and infinity. A value of zero gives Reichenbach's straight rule of induction, in which observed regularities are immediately assumed to be universal. A value of two gives a modified version of Laplace's rule of succession, and a value of infinity gives the rule used in Wittgenstein's *Tractatus* (5.15) which implies that new empirical input makes no difference to the probability of an hypothesis. The value of λ which gives the optimally efficient inductive method, argued Carnap, depends on the kind of world we live in. The more irregular the world, the higher the value of lambda that is appropriate. Unfortunately we can only know what kind of world we live in by a prior choice of inductive method. Carnap was very clear that, as a consequence of this, his investigation did not solve the general problem of justifying inductive inference. His investigation merely spelled out the predicament of the inductive reasoner with greater precision.

To make these ideas more precise, suppose we have a sample S of individuals, of which S_M possess the property M. How much does this evidence (e) confirm the hypothesis (h) that another individual, not in the sample, will have property M? Carnap arrived at the formula:

$$c(h,e)=\frac{S_M+\left(\frac{w}{k}\right)\lambda}{S+\lambda}$$

The symbol (w / k) refers to a property (called the "relative logical width") of the language in which the evidence and the hypothesis are couched. Notice how λ = 0 gives c = S_M / S whereas, when λ approaches infinity, c itself approaches w / k. Carnap expressed this by saying that λ gives the relative weight of the "logical factor" and the "empirical factor". The "logical factor", being a property of the language involved, refers to the conventions of the representational system and the classificatory categories in use. By "empirical factor" he meant such things as observed, relative frequencies. We must not forget, however, that even the numbers which specify a relative frequency depend on the underlying classificatory system. The two factors, then, are not wholly independent. Carnap was also quite explicit that a decision about the relative weight to be accorded to these two factors is a methodological choice. The issue is procedural rather than factual.

The different values of λ are to be thought of as specifying different learning strategies and different attitudes towards risk. They express different stances towards the dangers of too hasty, or too cautious, generalisation. Here we must begin to move beyond Carnap's individualistic formulations in which he asked simply, "which of the available methods a man X ought to chose" (p.53). These strategies and attitudes, when embodied in scientific practice, are not personal or subjective choices. They are educated, moulded, and collectively sanctioned. We all have individual feelings of risk or complacency, of the need for new approaches or trust in the old ways, but out of these comes a collective resultant, the strategy which prevails collectively. At this level we are dealing with the normative characteristics of the community, with its collective commitment to its traditions and its attitude towards innovation and novelty. The strategies defined by Carnap should not be read as psychological phenomena. He was intending to describe science, so the level of analysis should really have been social rather than individual. Carnap's logical endeavours in the realm of induction, therefore, succeeded in defining a parameter, λ, that is actually a property of social structures.

A practical application of these ideas about the social character of induction might be helpful. Consider the challenge issued to sociologists of knowledge by the late Donald Campbell (1989). Campbell was a philosophical sceptic but not, he insisted, a "nihilist". He wanted to know how to improve knowledge. Sociologists ought to be asking what form of social organisation was optimum for bringing a group of persons into contact with reality. What sort of "tribe," as Campbell put it, would learn most quickly? Would it be Type A, a small, egalitarian, democratic group? Or Type B, an hierarchical, authoritarian and traditional society? Campbell himself was inclined to favour Type A. It is clear that both the problem, and the suggested solution, has a certain analogy with Popper's (1945) discussion of science as an "open" society and with the theme of pluralism discussed earlier. (See also Merton's (1942) link between democracy and the normative structure of science.)

Campbell's intuition that Type A and Type B social structures would embody different cognitive strategies is a plausible one. An egalitarian, democratic group will provide a ready forum for new ideas. No sooner does a conception spring up in the mind of a creative individual than the inventor is in a position to present it for discussion. There are no significant entry costs. Others who disagree can argue against it, but they cannot silence it or stop its spread to others who might find it attractive. It seems that, as a consequence, the individual researchers must have before them a wider choice of intellectual resources than would be available in other socio-cognitive systems. Things seem quite different in the authoritarian, traditional and hierarchical society. Perhaps new ideas can be injected at the top and filter down, but even here tradition acts as a restraint. New ideas can disrupt existing patterns of deference so, in general, it seems against the interests of those at the top of the hierarchy to be too innovative. Feyerabend would note that, here, new ideas will only gain access on condition that they are consistent with what has gone before. In order not to de-stabilise what is already there they must be mere elaborations of the existing tradition.

The small group envisaged by Campbell as his Type A society would be a group whose implicit inductive strategy would express their inability to sustain a tradition. Collectively they could not generalise strongly from experience because any potential generalisation would be met by rival conjectures. For the members of such a group, their collective λ will have a high value.[13] If the world were a stable place the danger would be that the potential for rapid learning would be diminished. It would be undermined by individualism and subjectivity. Conversely a traditional group, Campbells's Type B society, would have a strong tendency to persevere with its traditional theories and would be at a disadvantage if its members were thereby rendered insensitive to the world's variability. Here the collective λ would be very low and the existing understanding of the world would be confidently projected forward. So how are we to answer Campbell's question? Carnap has shown us that we cannot furnish any general answer to the question about which is the best social arrangement for knowledge. The answer depends on how the world is, and that is exactly what we are trying to find out, using whatever social arrangements seem available and intuitively attractive. A general, prescriptive and normative sociology, of the kind Campbell was asking for, presupposes a prior solution to the problem of induction. We can be sure such a solution will not be forthcoming.[14] If we do have invincible intuitions as to what must be the best social form for epistemic purposes – a competitive pluralism, or an egalitarian participatory democracy, or an authoritarian hierarchy – we are almost certainly deluding ourselves. The social form will be attracting us for quite other reasons, and our theory of knowledge will be functioning as a mere ideological legitimation.[15]

V RELATIVISM AND SYMMETRY

Just as there is a variety of different forms of "realism" so there is a variety of different forms of relativism, linked by relations of family resemblance. Those who reject relativism sometimes fasten upon one special form of the doctrine, refute this to their satisfaction, and then allow themselves to proceed as if they had refuted relativism as such. The issue is also beset with other problems. The only real basis for identifying a position as relativist lies in its rejection of a corresponding form of absolutism. Moral relativism is the rejection of absolute standards of right and wrong, not the rejection of the notion of right or wrong as such. Relativism is consistent with accepting, say, local and contingent standards, and hence with the continued employment of the vocabulary of evaluation. It is even consistent with accepting universally held standards, as long as those standards are understood as being merely contingently universal. The issue is one of ultimate status, not mere scope or generality. The same considerations apply to epistemic standards.

Given that the essence of the matter is the contrast between the relative and the absolute, it is a sin against clarity routinely, and without qualification, to contrast relativism with rationalism or relativism with realism. These are not simple dichotomies. The opposite of rationalism is not relativism it is irrationalism; and the opposite of realism (or materialism) is not relativism, it is idealism. A rationalistic view may come into opposition to relativism if it involves a commitment to absolute rational standards, but this should not be assumed as a matter of course. The same applies to realism. If "realism" means, not just an affirmation of an independent reality, but also a commitment to some unique, absolute, and true description of that reality then, indeed, the idea stands in contradiction to relativism. Many self-professed "realists" do run these things together and assume both a form of materialism and a form of the correspondence theory of truth, but it is better to keep the issues of reality, truth and rationality separate from one another, and from the discussion of relativism.[16]

There are a variety of ways of denying that there are any absolute epistemic standards, and hence a variety of ways of defining relativism. It might be done by asserting that all knowledge claims are of equal worth, that they are all true or all false, or all equally justified, or all equally unjustified. Some critics, and some proponents, of relativism take themselves to be dealing with claims of this kind. In this spirit it has been said that relativism is the ideology of multiculturalism in that it does not discriminate between the value or truth status of different perspectives but places them all on a par with one another. It undercuts the basis on which one culture might justify its domination or rejection of another. This form of relativism may recommend itself to those engaged in political struggles on behalf of minority or marginal groups whose voice has difficulty making itself heard. But those engaged in such struggles are also acutely aware of the dangers of this stance. If there is already domination by one form of culture where is the leverage and justification for the demand to be heard? Relativism, it seems, suffers from an epistemic version of the paradox of toleration. Should toleration extend to those who are themselves advocating intolerance? If the answer is negative this, in itself, is an exercise in intolerance; if the answer is positive, it represents a capitulation to intolerance. One way or another intolerance is victorious. Similarly, some formulations of relativism are well known to entail logical traps for an unwary advocate.

Considerations such as these have led some feminist writers to reject relativism, seeing it as standing in opposition to the demands of emancipation. One example of this stance is Alison Wylie's well-documented account of feminist criticisms within archaeology (Wylie 1996). The feminist critics suspected that the contribution of women to life in ancient societies had been systematically played down. She describes a determined effort made by a group of archaeologists, both feminists and non-feminists, to examine the issue of how the role of women had been treated in the discipline. This exercise revealed a number of assumptions in the literature, for example, the automatic attribution of an active role to men and a passive role to women. It became clear that these assumptions could not stand up to critical scrutiny, having neither positive evidence nor *a priori* probability to recommend them. It is this evidential aspect that Wylie, understandably, wants to emphasise. Those who have worked to uncover weaknesses in a field, and who see their efforts as improving the existing level of understanding, do not want their claims simply put on a par with those they are criticising. They want to say that their ideas are better grounded. For these reasons Wylie rejects what she calls "strong constructivism", and its associated relativism, in favour of a stress on the local differences in credibility that can attach to competing claims and theories. In the context of a given argument some claims, she wants to say, just are better than others, and (as judged by hard-nosed evidential criteria) feminist archaeology is just better archaeology than its male-biased predecessor.

The relativism here being rejected is the doctrine that all claims are equally justified or unjustified. At no point in this otherwise convincing piece, does Wylie consider that her conclusions could also count as, or be consistent with, another form of relativism. There is no hint that the opposition of the "evidential" with the "socially constructed", within which she frames the discussion, may itself be defective. Suppose that relativism were not to be defined in terms of epistemic value at all. Evidential evaluations, of the kind proper within a discipline, are then not denied, or challenged, but instead made the object of explanation. The question to be posed by the sociologist would then concern the causes of credibility. Instead of all beliefs being treated as if they were, or should be, equally evaluated, they could all be treated as equally problematic in terms of why they are believed or believed with a certain intensity. The idea is that all beliefs, however the sociologist or the actors themselves might evaluate them, stand in need of causal explanation. That too is a form of relativism, and one that has been promoted since the earliest days of the field. There is no inconsistency between this stance and accepting that, in another context, and for other purposes, these (equally problematic) beliefs will be differentially credible to the actors concerned. Differential credibility, which is what Wylie is looking for, and which she grounds in local coherence, is precisely the thing to be explained. Wylie's rejection of relativism, therefore, only applies to one possible (and unacceptable) form of that doctrine. In order to avoid clashing with the reasonable considerations here at issue we should stick to the following definition. For the purpose of the sociology of knowledge relativism is the thesis that the credibility of all beliefs calls for explanation in terms of local, contingent causes (Barnes and Bloor, 1982).

One way to express the form of relativism introduced in the previous section is by use of the metaphor of symmetry. Sociologists of scientific knowledge, particularly those associated with the so-called "Edinburgh school", or the "strong programme", have said that the explanation of both true and false beliefs should be "symmetrical" in the sense of tracing back their credibility to the same kinds of cause. (Not, of course, to identical causes but to the same kinds of cause. Identical causes produce identical effects and what are in question here are different bodies of belief.) The position is also sometimes expressed in terms of "bracketing off" evaluation for the purposes of conducting a causal explanation. This way of speaking is acceptable provided certain cautions are heeded. To bracket off evaluation does not mean that evaluation "plays no role" in the proceedings, it means that it is part of the topic to be explained, rather than one of the resources to be used in the explanation. In what follows I shall use the label "strong programme" to refer to this commitment to symmetry. In the literature the programme also involves a commitment to causality and reflexivity (see Bloor 1991, ch.1) – but for our purposes it is symmetry that stands out in importance.

One of the difficulties in the idea that all bodies of belief stand in need of an explanation is that it cuts across the common-sense structure of curiosity. In everyday life we ask for explanations for things that stand out against a background of taken for granted routine. We ask for motives for crimes, not motives for honesty. We ask for the causes of rail accidents, not the causes of routinely safe journeys. We want to know the reasons for a suicide, not the reasons why people continue to cope with life. To ask for explanations carries the implication that something has gone wrong. To ask why scientists believe what they do can therefore sound to many ears as if it carries the implication that perhaps they should not believe it. There seems to be an implicit suggestion that all is not as it should be, and that something disreputable is to be revealed. A symmetrical, or completely general distribution of curiosity, thus comes to look like a completely general attitude of criticism. This may explain the false but seemingly ineradicable conviction, in certain quarters, that sociologists of knowledge are anti-scientific.

Sociologists do indeed need to structure their curiosity in a different way to that encouraged by common sense. Others must learn that the new structure does not carry the evaluative implications they fear. Such new structures of curiosity are possible to achieve, and the feat is frequently accomplished by other groups of specialists. While the farmer, hill walker or sailor, tend to focus on "bad" weather, treating it as qualitatively different from "good" weather, meteorologists have a wider, more disinterested concern with the causes of all manner of conditions. The doctor's concern with "disease" and its causes has often closely followed that of the layperson's, but the physiologist or biochemist has a structure of curiosity that frequently cuts across this divide. For the car driver, machine operator, or radar screen scanner, "mistakes" represent causes for concern quite different from proper exercises of the relevant skills, but for the psychologist they are both expressions of the same underlying mechanisms under slightly different causal conditions. Here is how one eminent experimental psychologist expressed the standpoint of the professional. "Despite years of psychological effort, it is still not widely realised in our culture that a man can see something which did not happen, and that he does so precisely through the workings of the system which in other cases makes him perceive accurately". This passage comes from p. 63 of Donald Broadbent's book, *In Defence of Empirical Psychology* (1973). His point is clear. Whilst philosophers have sometimes been tempted to talk in terms of two kinds of perception, veridical and illusory, for the naturalistically inclined experimental psychologist all perception comes from the same set of perceptual mechanisms. Whilst common sense inclines us to ask for the causes of illusions, but does not raise the question of the causes of veridical perception, professional psychologists cannot and should not proceed in this way. Their approach shows what I have called 'symmetry' in operation at the level of individual psychology. There is no question of postulating one piece of machinery to produce veridical outcomes and another piece to explain illusion. The same types of cause are in operation in both cases. The analogy that I drew between the visual system and the social system suggests that the same structure of professional curiosity can be adopted by the sociologist.

It should be clear from these examples, particularly the quotation from Broadbent, that what I have called the "symmetrical" or "relativist" stance is just a version of what, elsewhere, we have no difficulty in identifying as the "scientific attitude". The scientific attitude will not, for sure, be a unitary or simple thing, it will be a family resemblance grouping, but the point still stands. The suggestion is that what the sociologist of knowledge has been cultivating is nothing less than the scientific attitude – directed at science itself.[17]

There are a number of standard criticisms of the methodological requirement of symmetry. They typically revolve around the mistaken conviction that symmetry, relativism, and the strong programme in general, depend on the claim that the only causes are social causes. Pointing to the (indisputable) role of non-social causes, such as sensory experience, is taken to be sufficient to refute the approach. It does no such thing. The fact that an object encountered in experience can prompt quite different beliefs about it or accounts of it, and that in so far as it is a common factor it cannot explain the differences, should be enough to expose the weakness of such criticisms. The causal impact of objects on our sense organs underdetermines the conclusions we draw. The causal powers of the objects in our environment are certainly a necessary part of the story of belief formation, but they are not sufficient.

We must also remember what has been said previously about the ambiguity of talk about "facts", and the importance of keeping separate the "fact" as a state of affairs and the "fact" as a verbal account. We must not assume that the way we, or contemporary scientists, designate the state of the world, or the objects encountered, amounts to a more natural response than others provided by ancestors or aliens. Our current practices should have no implications for how problematic or how unproblematic we find the responses or designations of other groups. All such responses and designations call for explanation. That is the point of the symmetry postulate, and it has nothing to do with any claim – which would be quite absurd – to the effect that the only causes are social causes.

In order to illustrate this accusation – that the only causes are social causes – I shall examine an argument put forward by Stephen Cole, a student of the influential sociologist Robert Merton, and a critic of the strong programme. His case is worth examining because it will expose a certain, wide-spread, misunderstanding. In his book *Making Science: Between Nature and Society* (1992) Cole begins by declaring himself to be a "constructivist", but a "realist-constructivist" as distinct from a "relativist- constructivist" (p. x).

A realist-constructivist believes that science is socially constructed both in the laboratory and in the wider community, but that this construction is influenced or constrained to a greater or lesser extent by input from the empirical world. Instead of saying that nature has no influence on the cognitive content of science, the realist constructivist says that nature has some influence and that the relative importance of this influence as compared with social processes is a variable which must be empirically studied (p x).

This is offered as a point of contrast to the strong programme and other, allegedly more 'radical', forms of constructivism, all of which are said to proceed on the assumption that the input from the natural world is zero or negligible.[18] The task of the sociologist, notice, is to solve an empirical problem: measuring a certain "variable". This variable specifies the relative influence of "nature" as compared with "social processes". Its value ranges between zero (no influence) through "some influence" to, one must presume, a real or hypothetical case of total influence.

Clearly there are some things which are right about Cole's stance. First, the material world does play a significant role, and if "realism" is simply the label for acknowledging this, then realists we should be. (Properly understood, the strong programme has always been quite explicitly realist in this sense and it is a mystery why Cole should think otherwise.[19]) Second, there is an informal sense in which we might sometimes want to say that one body of belief was more in touch with the real world than another one. Not every group has the same intensity of involvement with the world, as do today's professional scientists. Verbal and theoretical accounts or stories do, after all, perform different functions.

Nevertheless, there are problems with Cole's statement when it is construed as a research programme. There is no well-defined empirical problem having the structure that Cole suggests. It may sound like an empirical question to decide on the proportion of the two ingredients that make up knowledge, just as it is an empirical problem to decide on the proportions of salt to water in samples of a solution. But what if it isn't like this at all? What if both ingredients are equally necessary, and both factors are fully engaged and taxed to their limit in all cases? That there is something wrong with Cole's idea of what constitutes an empirical question becomes clear if we return to the analogy I introduced earlier. Recall the physiologist examining the structure of the visual system. Ask yourself: is it an "empirical question" to decide on the proportion of the visual experience that is contributed by the object perceived, as compared to the proportion contributed by the perceptual system itself? Surely not. Visual experience does not have the kind of relation to its causal preconditions that would allow us to make sense of the question of 'how much' each contributes. One can ask *what* each contributes, but not *how much.* The issue of 'how much' could not be settled by experiment, any more than it could by philosophical reflection. The reason is simple. Confused questions do not have sensible answers, of any kind.[20]

A better way to frame the problems of the sociology of knowledge is not in terms of relative influence, but in terms of the different kinds of conventional and institutional machinery to be found in knowledge. As I have insisted, the conventionality of knowledge is not an optional ingredient so much as a vehicle for, and facilitator of, our access to reality. It isn't a question of more or less access but of the manner, quality and kind of access. Notice in this connection, that Carnap's approach, though superficially resembling Cole's, is really profoundly different. Carnap indeed spoke of λ as a measure of the relative contribution of the 'empirical' and 'logical' factors, but this does not address Cole's desire to find out how much the world, as distinct from society, influences knowledge. The difference is that, on a sociological reading of Carnap, any relative weight, and any value of λ, is no more nor less a social influence than any other. It is true that, at $\lambda = \infty$, the influence of the world is zero. But the different values of λ do not signify different degrees of responsiveness to the world. They signalise different ways of being responsive. A group following the straight rule, where $\lambda = 0$, is not more influenced by the world than one following an inductive rule where $\lambda = 2$, it is merely more confident in its response.

VI HYPERSYMMETRY

The misunderstanding described above (that, in the strong programme, the only causes are social causes) can be found in numerous books and papers by Bruno Latour. (See for example, Latour 1987, 1992 and 1993. For advocates of the strong programme, he says, "Society was supposed to explain Nature!" (1992, 278). His own conception of how science should be analysed is, at least in intent, quite different from that proposed in the sociology of knowledge. It is meant to be a qualitatively new approach that he calls 'anthropological' – though this is not going to be the anthropology of a Durkheim or an Evans-Pritchard. Unfortunately, just like Cole's, Latour's position is premised on a radical misconception about what has gone before in the field. Latour thinks sociologists are locked into the assumption that causes of belief are either social or natural and that the more they are of one the less they can be of the other. Sociologists, we are assured, are moving along a single dimension between two poles, the subject and the object, or society and nature, and their position along this dimension defines the proportion they assign to one or the other of the two ingredients in the mixture – the only ingredients that are possible. Latour interprets the "strength" of the claim made by the strong programme in terms of this zero-sum game allegedly being played with nature and society (1992, 283). What can a claim to strength be, when made by supporters of a sociological programme, but a claim to have eliminated all causes but social ones? After taking the reader through the Kantian dialectic of the subject and the object, not to mention the travails of modernity, we are back with the old charge that sociologists play down or ignore the role of nature.

Latour accepts the methodological virtues of addressing both true and false beliefs with the same degree of analytical curiosity, that is, he accepts the old symmetry principle of the strong programme. He accepts it, but wants to go beyond it. He declares himself in favour of a new, generalised version of symmetry. The new version is meant to overcome a crippling, residual asymmetry which, he thinks, still lurks in the old version. This is the "asymmetry" of (allegedly) assigning all causal power to society and none to the things in nature, which are mere social constructs. The new, generalised principle of symmetry is designed to restore parity of esteem, to give activity back to nature, and to produce accounts which are, at last, truly symmetrical. It soon becomes clear, however, that this is going to involve some very radical conceptual revisions:

> it is crucial to treat nature and society symmetrically and to suspend our belief in a distinction between natural and social actors. (1988, 260)

Latour is not just saying that objects in the world, such as electrons, exert an influence and certain conventions in society, like the definitions of geometry, also exert an influence, and both should be taken into account. From his point of view that would be mere eclecticism, it would be to stay within the very polarity from which he wants to escape. That would be Cole's solution. How, then, does Latour recommend we escape from mere eclecticism? The crucial move is to stop seeing society and nature as two causes and see them as two effects, or two products of some single underlying process. The new, generalised symmetry principle enjoins us to think of nature and society as "co-produced". Unfortunately, though hardly surprisingly, Latour never manages to make this metaphysical vision even remotely clear. Metaphors are introduced, monads and entelechies are invoked, a new terminology of "quasi-objects" and "actants" is suggested, and mental exercises recommended, such as using purposive language to talk about inanimate things and mechanistic language to talk about people – but all to no avail. Embarrassingly, the project is mired in impenetrable obscurity. When Latour tries to put his ideas into action and use them to analyse historical episodes, such as the reception of Pasteur's ideas, all that we find, beneath the rhetoric, is a dilute version of the standard methods of the sociology of knowledge, the very thing denounced as hopeless and outdated.

There is, however, one very important difference which remains, and which marks a sharp and enduring divergence from the materialistic tendencies which inform the strong programme. Latour repeatedly and systematically runs together the idea of nature and the idea of an account or description of nature. The difference between a thing and how it is called finds little or no recognition in his writings. Recall how Latour glossed the strong programme as the extraordinary claim that society explains nature. In reality, of course, the programme concerns the role of society in explaining the knowledge of nature, not nature itself. The fact is lost on this critic, as it is on so many. In Latour's case this is not an accident. Believing that sociologists have been imprisoned in some form of the subject-object distinction, he demonstrates his own freedom in this regard by consistently refusing to draw that distinction in the course of his own writings. It is routine for him to run together reality itself and some verbal account or scientific description of that reality. Here is a typical passage that repays study:

> As long as the social sciences did not apply their tools to Nature and to Society at once, the identity of the two transcendences and its common constructed character were left in the dark. Even when established science and stable society were studied together, their common production was still not visible. (1992, 282.)

The two 'transcendences' in this passage are society and non-social nature. In reality, says Latour, these two things are co-produced, so there is really only one transcendent thing, namely the source of both nature and society. So some of the language of this passage is provisional. But notice that one of the poles of this, soon to be transcended, schema is first of all called "Nature" and then it is called "science". But nature, in our ordinary way of thinking, is the object of knowledge, the thing that is known, while science is the knowledge we have of it, our theories about it and our description of it. So this 'overcoming' of the subject – object distinction runs together the two senses of 'fact' I have been trying to keep apart. Our science and the world we know are fused.

If ever there were a formula designed to diminish the value of sociological study, it would be this fusing together of what the world is really like and what, at any given time, it is thought to be like. It might have been invented to destroy critical self-awareness. Ultimately this is exactly what it does in Latour's own thinking. In *Science in Action*, rather than trying to explain anything, the sociologist is simply told to follow scientists around and, it seems, agree with everything they say. When the scientist is being sceptical, the follower is permitted to evince scepticism; when the scientist is confident and unqualified in an expression of belief, so is the follower (1988, 100). What else would one expect from a conception of knowledge, which has abandoned the apparatus for sustaining critical distance?[21]

VII Interests and Practices

One example of this gap between aspiration and achievement in Latour's work concerns the role of social interests. Interest explanations are a standard part of the armementorium of the sociologist. They have played a fundamental role ever since the great Edinburgh sociologist David Hume shocked his Scottish contemporaries, such as Thomas Reid, with his sociological construction of the concepts of justice and obligation (see Bloor 1997a, ch.9).[22] Today's critics of interest explanations respond roughly as Reid responded: they find them "reductionist", and offensively causal. They are a vulgar intrusion into a realm where they wish to celebrate more spiritual values such as freedom, spontaneity, purpose and striving. In *The Pasteurisation of France* (1988, 260), Latour calls on sociologists to give up the appeal to interests, but it is impossible to read his account of the differential response to the germ theory of disease given by hygienists, surgeons, physicians, and the medical authorities in the military, without seeing it as an appeal to interests. The point is denied, the terminology is changed and metaphors introduced, but the basic picture cannot be suppressed: the issue is one of interests.

Similar hostility to the "reductive" character of interests is to be found in other writers. The common theme of their complaint is that interests are rigid and pre-formed. No explanation based on them can do justice to the changing and emerging properties of a situation. These critics have a certain conception of what an interest explanation must consist in. It must (they think) assume a pre-existing set of social relations, a social given, which the sociologist privileges above all other aspects of the situation. These pre-given structures then determine social agents (who, of course, thereby lose their real nature as agents and become what ethnomethodologists call "social dopes"). When the interest has exerted its influence the situation is fundamentally unchanged, the interest is still there and the same static configuration holds sway.

Such themes have been developed at length by Andrew Pickering in *The Mangle of Practice* (1995). The metaphor in the book's title is meant to emphasise the remorseless interaction of all the different components of scientific knowledge. His aim is to produce a "post-humanist" and "post or anti-disciplinary" conception of knowledge. The label "anti-disciplinary" is to show that he is not simply advocating eclecticism. He wants to transcend the allegiances that tie down sociologists of knowledge to a merely social perspective. Sociologists, he says, treat social processes in a static way, as external causes standing outside the situation, rather than as things which are themselves part of an ongoing interaction and change. The social, too is "mangled" in the plane of practice, but sociologists are unwilling to see this.

Is it true that sociologists treat social causes as static and "non-emergent"? I think not. Let us stay with interests. Ever since Kuhn developed his account of paradigm-based "normal science", it has been apparent that the notion of an interest must have a prominent role to play. The activity of the practitioners within every paradigm will generate an interest in that paradigm's maintenance and development. This interest will be grounded in the interaction of the scientists and will thus be, in Pickering's own phrase, in the plane of practice. As we work through the sequence that Kuhn sketched (of normal science, extra-ordinary science and revolution) the operation and structure of this pattern of interests will change. It is not a static and unhistorical thing at all. Think of a paradigm becoming progressively consolidated. A divergence of interests is likely to emerge between the circle of scientists who initially developed the paradigm and those who came along later to conduct the normal, day-to-day articulation of the achievement. It will be in the interests of the founders that their accomplishment continues to be treated as a source of guidance and insight. At the same time others may well become aware of how it would be in their interests if a radically new approach were called for. It would yield opportunities absent under the present dispensation. So there will unfold over time a predictable divergence of interests.

My point doesn't depend on agreeing with Kuhn, or treating his account as beyond criticism. The work merely functions as an illustration. It serves to make the point that the conceptual apparatus, which has long been current in the field, carries the opposite connotations to those claimed by Pickering. It is not "static" or such that time is deemed unimportant. Far from being static, or external to the plane of practice, the routine ways of thinking in sociology have exactly the properties that he calls for.[23] The real difference between Pickering and the sociologists of knowledge he criticises is that Pickering seems disinclined to believe that anything can be explained; "there is no substantive explanation to be given for the extension of scientific culture" (1995, 146-7). Meanwhile sociologists, with scientistic optimism and humanist hubris, keep trying to explain things. They think they can illuminate Pearson's conceptual innovations in statistics (MacKenzie 1981) or the reception of Einstein's work in Cambridge (Warwick (1992-3)) or Hamilton's mathematical differences with his formalist contemporaries (Bloor 1981) or Boyle's dispute with Hobbe's (Shapin and Schaffer 1985) or the differences between German and American genetics (Harwood 1993) or the fight between Wundt and Külpe over the reality of imageless thoughts (Kusch 1999).

VIII CASE STUDIES

A discussion devoted to the general principles of a field is not the place to describe the details of its empirical basis. Nevertheless, some comments need to be made. The grounding of its principles in detailed case-studies has always been a matter of some pride and yet, on occasion, this entire aspect of the field has been swept aside and dismissed. It has been said that not a single convincing case- study has ever been provided which lends support to the claims of the sociology of scientific knowledge, see Cole (1996, 278). This is a fascinating claim. Either the practitioners of an entire field are deluding themselves, or the critics are blind to something that is before their eyes.

How can such a situation arise? Incompetence by one or the other parties may be the answer, but we should first try out another hypothesis. *Gestalt* psychologists have shown how some line-drawings can present radically different aspects to different observers or to the same observer at different times. A well-known example is a drawing that can be seen as either a young woman or an old woman. The lines in the picture perform a different role depending on the *Gestalt.* A demure profile in the one is a hooked nose in the other; the young woman's jaw-line is the old woman's nostril, and so on (see Osgood 1953, 207). Perhaps the data that looks so alluring to the sociologists of knowledge looks different to their critics because of some comparable process.

To test this hypothesis I shall examine a case-study in the history of science by Hon (1995). It should be explained that Hon does not mean his study to be added to the list of those cited by sociologists of knowledge. He is a critic and uses his case study against the sociology of knowledge. With this in mind, here are the details.

Hon looks at an important series of experiments conducted by Walter Kaufmann between 1898 and 1906. The experiments were seen as a test of Einstein's ideas though, at the time, what later came to be known as relativity theory tended to be assimilated to Lorenz's work. The results ran counter to the predictions of Einstein and Lorenz, but by around 1915 it was generally concluded that it was Kaufmann who was in error. The author of the case-study is interested in the identification of error; however, he sees the changing status of Kaufmann's results as a phenomenon that does *not* require a sociological interpretation and that is *not* relativist in its implications. I think the phenomena described in the study do need a sociological interpretation and do support a relativist analysis. I see a different picture to the author of the study.

Kaufmann developed an apparatus that would allow him to measure how the ratio e/m for electrons, that is, the ratio of the charge to the mass, varied with velocity. It involved deflecting beams of particles emitted by a radioactive source as they passed through electric and magnetic fields. His original intention was to test ideas derived from Thomson and Heavyside who predicted that the mass of a charged particle would appear to increase with velocity because of the interaction of the charge with the ether. The experiment seemed sufficiently accurate to discriminate between different models of the electron. Initially it supported the ideas of Max Abraham who suggested that the electron could be represented as a rigid sphere with a uniform distribution of charge spread through it. To this extent it told against rival models such as those associated with Lorenz and Einstein which treated the electron as subject to the Fitgerald-Lorenz contraction.

Hon describes the detailed reasoning behind the experiment and the great interest surrounding the theoretical work of Abraham and others. The theme which gave it excitement was that it all pointed to the conclusion that the mass of the electron was entirely electromagnetic in origin, and hence that its mechanical mass was zero. Hon's main concern, however, is with the varied assessment of the soundness of Kaufmann's results as a discriminator between different models of the electron. Accordingly he documents the responses to it of Poincaré, Planck, Einstein and Lorenz. Poincaré noted certain respects in which the functioning of the apparatus might be open to question. Was the vacuum as high as claimed? If the assumptions used in calculating the results were not accurate this would render them unreliable. Nevertheless, on balance, Poincaré accepted the outcome. As Hon puts it, "Poincaré submitted to Kaufmann's expertise on matters of experimentation and accepted the results" (p.204).

Planck, by contrast, who was one of the early supporters of relativity theory, did not accept the results. Like Poincaré he had doubts about the integrity of the vacuum and the consequences for the experiment if the electrons were ionising some residual gas as they passed through the electromagnetic field. The field would no longer be uniform as the analysis assumed. Planck also drew attention to the role played in the calculation of the result by the assumed value of e/m for the rest mass of the electron. By substituting some newer estimates that were different from those used by Kaufmann he brought the result somewhat closer to that predicted by Einstein and Lorenz.

For his own part, Einstein, despite knowing about Kaufmann's experiments of 1903, chose to ignore them in his famous 1905 paper. Later, in 1907, he took the line that the entire issue had been framed in terms that were too narrow. Should physicists really be constructing highly detailed models of the electron in the way that Abraham and others, including Lorenz, were doing? Isn't this premature and somewhat *ad hoc*? Einstein did not see himself in the business of matter theory at all and sought to recommend a different, methodological and epistemological, perspective. In the meantime, he declared, he would need a wider range of experiments to convince him that his work did not fit the facts. As Hon glosses his position, Einstein was more impressed by theoretical beauty than inherently problematic experiments.

At first, in 1904, Lorenz offered an interpretation of Kaufmann's experiment, which brought it, in Lorenz's opinion, into reasonable agreement with his theory. Kaufmann duly refined his procedure, increased the accuracy of the experiment, and by 1906 Lorenz accepted that the result refuted his model of the deformable electron. (This did not, however, incline him to stop work on the approach.) By 1915, and the emergence of other experimental results, Lorenz came to feel that the weight of evidence had swung the other way and now told against Kaufmann. By 1922 Lorenz was following Poincaré and Planck in saying that the result was probably an artefact of the defective vacuum and the breakdown of uniformity of the electromagnetic fields within the apparatus.

Now for the interpretation of the case-study. Hon says, "My intention is neither to render the concept of error relative nor to explain the occurrence of error in sociological terms; rather, I wish to argue for a close connection between ... epistemological framework and methodological approach and ...detection of error." (p.171) One might have thought that even in these terms Hon was indeed rendering the concept of error relative, namely, relative to an epistemological framework and to methodology. I take it, therefore, that the denial of relativism must rest on an implied distinction between sociological relativism and relativisms that are taken to be of a more intellectual and rational (and acceptable) character. There is no doubt that Hon wants, as he puts it, to contextualise error, but the context is identified as philosophical rather than social. Hon's positive position is this: to count a proposition as error means judging it against other propositions which are taken to be true. This judgement involves an assumed vantage point. "The scientist must ultimately seek this vantage point by himself or herself. He or she must decide what weight to assign to a particular set of evidence with respect to the correctness or otherwise of a certain statement." Using a phrase from Polanyi, Hon concludes, "This 'residue of personal judgement,' which depends on one's philosophical make up and methodological approach, determines the way experimental results are assessed." (p.171)

We can see at once that Hon's approach may reasonably be called "individualistic". The scientist must find the vantage point of judgement "by himself or herself". As well as being individualistic Hon's approach is also highly rationalistic. Judgements are tracked back to philosophical dispositions, "These epistemological and methodological elements constitute the context in which a failure is determined and identified." (p.171) Accordingly Hon relates Poincaré's acceptance of Kaufmann's work to a philosophy of science which accepts the priority of experiment over theory, and Einstein's rejection of it to his philosophy of science which accords priority, or more priority, to theoretical considerations and non-empirical virtues such as generality. (Planck's response is noted and described but not given any corresponding explanation.) Lorenz is described as vacillating between the "two poles" provided by Poincaré and Einstein.

The falsity of these claims should be clear. It is simply not true that Poincaré, Einstein, Planck and Lorenz reached their conclusions by themselves. They were working in contact with, and in the knowledge of, the work of numerous other scientists, including one another. The idea that this continuing dimension of interaction can be put aside and the actors treated as if, for crucial scientific purposes, they operated as independent individuals lacks even the remotest plausibility. The judgements and stances they adopted will have been sensitive in numerous ways to the judgements and stances of others. Of course there will be a residual element of individuality in their judgements; Einstein was not Lorenz and Lorenz was not Poincaré. They each had their individual history and personal trajectory, but that does not make them independent agents.

Hon's own statements readily suggest how individuality can be acknowledged in a common-sense way without lapsing into individualism. Recall his mention of the "two poles" defined by Poincaré's tendency to prioritise experiment and Einstein's tendency to prioritise theory. Here we have different personal orientations to the two great dimensions of scientific practice – experiment and theory. The orientations are individual, but the dimensions are institutions. We know that some scientists tend to specialise in one and some in the other. We know that around these activities there have crystallised different roles: there is the role of "experimenter" and the role of "theorist". These roles are not timeless abstractions; they have an historical origin and are subject to a changing understanding of what they involve. Nor are they necessarily exclusive, some persons can be adepts at both or shift the emphasis of their work from one to the other. But it is clear that in the current organisation of the discipline there is a significant division of labour. Thus we hear that "Poincaré submitted to Kaufmann's expertise on matters of experimentation" (p.204). Individualism and the picture of scientists acting by themselves cannot even make sense of a simple fact of this nature – the very fact, or kind of fact, that is central to Hon's entire study. Acts of deference and respect are inherently social acts and only make sense within a social system.

A corresponding point applies to Einstein's actions as Hon documents them. Einstein chose to orient himself more positively to the claims of theoreticians and to draw on the widespread understanding of the importance of their contributions and their special virtues and authority. Had this understanding not been available and well- dispersed throughout the physics community, his personal judgement would have lacked any credibility. Rather than being a piece of intelligible, intellectual risk-taking it would have seemed merely idiosyncratic.

In a footnote Hon sheds some further light on Einstein's personal strategy in this regard. He says, "It is worth noting, from a sociological point of view, that Einstein's reputation was at that time newly minted and the relativity theory was one of its central pillars." (p.223) Unfortunately he does not go on to explain why this is sociologically interesting, or what the connection is between this sociological observation and the rest of his analysis. The link seems to be as follows: because Einstein's emerging reputation depended on his theoretical contribution he presumably had a vested interest in not seeing Kaufmann's results gain wide acceptance. But if this is the point at which Hon is hinting it is hardly compatible with the proposition that the residue of personal judgement involved is to be contextualised in purely philosophical and methodological terms. It undermines the picture of the scientists seeking the vantage point for their judgements by themselves. It points, rather, to the idea that the credibility of the reasons advanced, the preferences involved, and the authorities invoked, are themselves things that need to be explained – and explained by reference to a social context. By consigning this observation to a footnote Hon has allowed himself to evade the responsibility of integrating the fact in question into his historical account and subsequent analysis.

There are two further respects in which Hon identifies the operation of social processes which are significant to his story but which are passed over and left dangling. First, consider the focus of interest of Kaufmann's work, namely, the relation between the mechanical and electromagnetic mass of the electron. Hon quotes Max Born who, some fifty years after the event, wrote: "as a matter of fact, the velocity dependence of energy and of mass has nothing ... to do with the structure of the body considered, but is a general relativistic effect. Before this became clear, many theoreticians wrote voluminous, not to say monstrous, papers on the electromagnetic self-energy of the rigid electron.."(Born, quoted p.197) Today, Born adds, all this effort seems wasted because the theoretical point of view has changed. The present tendency "is to circumvent the problem of self-energy rather than solve it. But one day it will return to the centre of the scene". Clearly something is a focus of interest for a group because sufficient numbers of members treat it as a focus of interest. They do not do this randomly or irresponsibly but neither is it determined by any absolute methodological or rational principles. It may be argued that Einstein's position was, in a sense, ahead of the game in this respect. His position foreshadowed the stance that Born identified as today's but, as Born also suggested, this shift itself can involve marginalising potentially important issues that, one day, may again be placed at the centre of attention. So Hon's quotation from Born identifies an element in the story that is clearly social.

Here is another unacknowledged social dimension to the story. Hon concludes his study by saying, "A consensus eventually emerged that Kaufmann's experimental conclusion was false, and the experiment was pronounced erroneous, but then there was no agreement as to the cause of the failure" (p.223). In such cases, he adds, the "alleged empirical findings simply fade away" (p.223). There is "no need for the concerned scientists to reach an agreement as to the characteristics of the error" (p.223). Given this account of consensus, what has happened to Hon's individualism and rationalism? We seem now to be dealing with an essentially collective phenomenon – the fading away of the finding – which cannot be identified with, explained by, or tracked back to, any of the diverse, individual lines of reasoning. Hon's advertised emphasis on scientists reaching decisions by themselves, guided by their philosophy of science, has ceased to do any work. It has given away to a wholly different story. Individual lines of reasoning, it appears, have been transcended in the consensus. The fact of Kaufmann's result fading away is not because any of the individuals, taken separately, have the opinions they do; rather, it is a fact because their individual stances interact in the way they do. In itself the point is very simple. All the more strange, then, that this should be the culmination of a story explicitly designed *not* "to explain the occurrence of error in sociological terms" (p.171)

Recall that the reason for looking at this case-study was to throw light on the startling claim, made by Cole, that sociologists of knowledge, despite their own confident beliefs to the contrary, have produced no convincing examples of social processes influencing the content of knowledge. I suggested that this clash of opinions may be because the same material was being viewed in different ways so that it formed different *Gestalten*. Has the present study born this out? No, the situation is not like the picture of the old woman and the young woman. We do not have *every* line of one conception of the situation being given a different but corresponding role in the other. We have seen Hon produce a picture with many salient social features, but then draw conclusions in which these are ignored. He has not provided an alternative interpretation of them; he has simply failed to integrate them into his conclusion. Judged by local standards of empirical adequacy, the attempt to use this case as a counter-example to the sociology of knowledge misfires. If technically accomplished analysts of science, such as Hon, can produce sociologically relevant case-studies, and then flatly deny it, no wonder that the field has difficulty making headway against critics. If we imagine this process applied to each case in turn, it is all too easy to see how critics can conclude that there are no cases at all to support the sociologist of knowledge.

IX IN VINO VERITAS

The sociology of science is currently the object of particularly vehement attack. It is appropriate to end the discussion by looking at this phenomenon. Indeed, not to produce some manner of response to it would risk being seen as evasive though, in truth, there is nothing novel or significant in the content of the criticisms. The only thing that is new is the set of associations existing in the minds of some of the critics, and the tone they feel justified in adopting. Sociologists of science now find themselves grouped alongside post-modernists, radical feminists, multiculturalists, and Afro-centrists. The basic charge, predictably but wrongly, is that sociologists of science are anti-scientific. One location of these attacks is a collection called *The Flight from Science and Reason* (Gross *et. al.* 1996). I have already addressed one contributor to this volume when discussing the work of the sociologist Cole. I shall end by looking at another of the contributors: the logician and philosopher Susan Haack. Haack's paper in the above book is called, "Towards a Sober Sociology of Science". Sober sociologists are contrasted with those who are, "intoxicated by one or another of various misunderstandings of the thesis that science is social" (p. 262). Given that the position I have been defending here is amongst those cited by Haack, as an example of bad sociology, her paper will provide a useful testing ground for what I have been saying. It will allow me to make a direct comparison between the perceptions and claims of one of the current critics and the argument itself – as it has been consistently developed over some twenty years.

Haack begins her article by saying:

> I don't believe that sociology of knowledge must, in the nature of the case, be the 'stupid and discreditable business' it has of late too often been. So my purpose in what follows is to articulate what distinguishes good from bad sociology of knowledge... (p. 259)

The words in quotation marks in the passage are an endorsement of the writings of another philosopher who has, says Haack, previously developed a position similar to her own. Now let us look at the substance of the claim. Good and sober sociologists of knowledge, Haack explains, do not believe that science is purely social. In particular, they accept a distinction between a theory being *accepted* as true by a group of scientists and the theory being genuinely *warranted.* Acceptance is a social phenomenon, while being warranted is a matter of being supported by good evidence. Acceptance is the more general category. Some acceptances are rationally explicable in terms of scientists recognising good evidence when they see it, while others are more purely social in the mechanism at work. Thus: "In any instance in which acceptance and warrant were quite disconnected, a purely sociological account of the acceptance of the theory would be appropriate..." (p. 262). The proper aim of the good sociologist is to delineate the social conditions under which good, creative, honest, scrupulous enquiry is possible (p.261). The internal organisation of science and its social environment will determine to what extent this ideal is realised.[24]

The traditional character of this position will be evident. It is the uncontroversial claim that society can facilitate or inhibit the rational pursuit of knowledge. As in earlier statements of this position, such as those by Merton, it is accompanied by references to cautionary literature about the corruption of science under totalitarian regimes. We can also detect some familiar dichotomies. Like Cole, Haack proceeds as if it is a contingent matter to identify the force accorded to evidence as distinct from the force exerted by social processes, or to decide the degree to which the two things are, in her terms, "correlated" with one another. The only minor variation that may be novel is that Haack insists that a proper analysis of how a belief becomes warranted cannot be entirely logical. It must also allow for causal processes to deal with the impact of experience on an individual's beliefs and, Haack allows, this analysis must even allow for a social dimension.

> A social dimension is also necessary because, in view of the role of experiential evidence, "how warranted theory T is" must be taken as elliptical for "how justified a scientific community is, at a time, in accepting T"; which depends in a complex way on how justified an individual who possessed all the evidence known to each member of the community would be in accepting T, discounted by some index of how justified each member of the community is in believing the others to be reliable. (p.260-261)

So knowledge is social in the sense that the total evidence involves the summation of what all the individual members experience, when it has been weighed according to reliability.

Before going into any further detail it is important to notice a significant feature of Haack's distinction between the social and the evidential. Haack takes these to be like chalk and cheese. They have different natures; they are different kinds of ingredient, and their analysis belongs to different spheres of competence. The philosopher is to produce theories of warrantability, leaving mere acceptance to the sociologist.[25] Here we have the false opposition of the conventional and the inductive I warned against earlier.

At no point in the paper is there any hint that the distinction between the social and the rational, and the social and the evidential, may not be secure. The "bad sociologists" are simply criticised for failing to make the distinction. That they may actually have challenged the distinction, and done so quite explicitly, rather than simply failed to grasp it, is never brought up for consideration. And yet, as we have earlier seen in this discussion, the very essence of the sociologist's position is that this is a false dichotomy. This has been illustrated in the present treatment both by historical example and by reference to Carnap's logical analysis of confirmation. That Carnap did indeed try to produce a purely logical treatment, while Haack acknowledges the socially distributed character of the evidence, makes no difference for the present purposes. No significant alteration would be needed to allow Carnap's model to take cognisance of Haack's "social dimension". Carnap's essential point would still follow: there is no unique measure of evidential support or warrant. Whatever measure is used will have the character of a collective choice, and it will have to be sustained as a convention. The social, in other words, is right in there, in the midst of the rational process of warranting. Warranting is not acceptance minus the social, it is itself a process whose structure and content cannot be properly analysed without identifying its conventional and social dimension. This dimension goes far deeper than the summative process accepted by Haack because, as Carnap's work reveals, it is implicated in the very content of the items to be summed.

Haack's failure to allow for this possibility shows up subsequently in her attempt to dismiss the argument from underdetermination. Given that the causal impact of experience from the world does not serve to fix belief, because there is never a unique interpretation, sociologists have sought to locate social determinants. These social determinants do not, of course, exert their influence instead of sensory input, but always in conjunction with it. Haack's response to this argument contains two errors. First, it would be better here to talk of underdetermination by *experience* than by *evidence*.[26] Remember that, for the sociologist, evidence itself, the degree of confirmation, is already a social category in that it cannot be defined without reference to a conventional component. Second, Haack tries to take issue with the underdetermination argument by insisting that some beliefs are tentative and provisional.[27] This cannot possibly meet the point because the underdetermination argument applies just as much to probabilistic conclusions as it does to categorical and unqualified claims to belief. This should be obvious by reference, once again, to what has been said about the meaning of Carnap's work on confirmation theory. The social and conventional elements identified here are precisely applicable to beliefs of the form Haack has in mind, e.g. when we might say that we believe T1 is better warranted or confirmed than T2, though neither are warranted or confirmed very much.

We have now seen that the central thrust of Haack's argument, her dualism of the social and the evidential, is untenable. Nothing else that is put forward in her paper can rescue it from this fatal flaw. There is the usual litany of attacks against the allegedly debunking character of the sociology of knowledge and the alleged denial of the role played by causal inputs from the material world. All of these have been anticipated and answered in advance, as far as the strong programme is concerned, in the previous discussion.[28] Nevertheless, it is worth noting two further features of the argument.

First, Haack does not address a single, clear and well-defined target. Instead of distinguishing between the different positions that are mentioned in her discussion, they are all lumped together. We are given a list (on p.262) of very diverse stances, all of which the reader is invited to assume must embody the sins attributed to "some recently dominant trends" (p.259) or what is "invariably" (p.260) done. The strong programme is thrown into the list alongside (amongst others) Longino's feminism and Latour's idiosyncratic brand of ontology. I have pointed out that Latour is actually a critic of the strong programme. He is not a sociologists of knowledge in any ordinary sense, having, in his own words, "written three books to show the impossibility of a social explanation of science" (1992 p.284). I have also drawn attention to differences between Longino's thinking, and perhaps the concerns of feminism in general, and the preoccupations which are central to the sociology of knowledge as such. Some of these differences arise from the predictable divergence between political activism and academic enquiry; others have touched more specific methodological questions. Unfortunately, all such discriminations are lost in Haack's category of bad, intoxicated sociology.

Imagine if sociologists were to attack an entity called "the philosophy of science" and took as a target an amalgam of Popper, Carnap, Kuhn and late

Feyerabend, all of whom were dismissed because, say, of their alleged "positivism," "scientism" and "inductivism". The perpetrators, rightly, would be laughed out of court for their confusion and naiveté. And yet, logically, this is equivalent to what has gone on here. The most that can be said in mitigation of Haack's procedure is that others have done exactly the same, for example Cole 1992 and Fine 1996. It is high time that real positions were identified, delineated with some care, and dealt with separately. Crude stereotypes are not good enough.

Second, Haack actually gets the symmetry postulate of the strong programme crucially wrong. In one of the few places in the article in which we have a piece of direct exposition, rather than long-range imputation, she describes the strong programme as, "treating true and false theories strictly alike" (p.264). Where does the word "strictly" come from? Such a reading produces nonsense. There is certainly no justification for this word in the one brief formulation of the programme that is given as a reference. Indeed there has never been any textual warrant for Haack's wording, except perhaps in the writings of other critics who make the same error.[29] Would any philosopher casually accuse a reputable, working psychologist, such as Broadbent, of treating illusory and veridical perception as "strictly alike"? I doubt it. They would take the trouble to notice that his point was that both come from the workings of the same mechanisms and that they stem from the same general kind of process. (Obviously not from strictly identical processes – otherwise they would be identical.) As I have explained above this is, and always has been, the position of sociologists who advocate the strong programme. The point has been made clearly and explicitly in the literature for twenty years but still, it seems, some philosophers cannot extract it from the page in front of them. That critics should then feel able to use phrases such as "stupid and discreditable," to describe those whose views they have just traduced, is deeply regrettable.

X CONCLUSION

I began by saying that the sociology of knowledge constituted a challenge to more traditional formulations of epistemology. I shall end by restating that challenge and presenting it in its sharpest possible form. Philosophers have charged sociologists of knowledge with being relativists. In return I shall charge philosophers with being absolutists. I have embraced the role of relativist and explained what it should mean and why it is desirable and defensible. I should like to see philosophers make their absolutism equally explicit and, if they can, plausible.

The history of philosophy is widely understood as the history of the separation of the different specialised sciences. From the trunk of the tree there has branched out mathematics, physics, biology and, around 1900, psychology (see Kusch 1995). The sociology of knowledge is the latest offspring. It may be the last. The long, historical task of philosophy may be close to its end. Why is this? Because there is only one role left: to be the self-proclaimed guardian of a residue of absolute values – and that role is unsustainable. Tasks like "conceptual clarification" cannot sustain a disciplinary identity not, at least, without the prop of a tacit absolutism. Without this, clarification is best left to a diversity of specialists who are immersed in the projects of their respective fields. Clarity is a pragmatic category and must always be clarity for a certain purpose and clarity for a specific group of thinkers. Without the practices and paradigms of a specialist field of empirical enquiry, what ground is there to stand on? Philosophy is the guardian of absolute values, or it is nothing.

A moral relativist denies that there is any absolute basis or justification for moral obligation. This does not mean, "anything goes". Social life never permits such a principle because it would be the negation of social order itself. But social life also functions, and always has functioned, without any genuine, absolute justification for its demands and imperatives. Frequently its local and relative standards will be presented as absolute and perceived as absolute. For example, they will be seen as the decrees of God. On the level of philosophical reflection, any critic of moral relativism should be prepared to give a clear justification of their claim to have available to them some non-relative, absolute moral truths. The same applies to a critic of epistemological relativism for, at some point, they must lay claim to absolute standards.

There are bound to be those who believe they can evade this responsibility. They will think they can reject relativism without, at some point, embracing absolutism. There will, no doubt, be talk of a "third-way", and of going "beyond" the choice between relativism and absolutism. But those who claim they are both non-relativists and non-absolutists are deluding themselves. Critics of the relativism of the sociology of knowledge should not prevaricate. They should have the courage of their convictions, and the clarity of mind, to declare their absolutism and to show the world the absolute values they have been vouchsafed. Having done this, they can then explain to the ever curious sociologist just how they accomplished this epistemological miracle.

ACKNOWLEDGEMENTS

I should like to thank Celia Bloor and Martin Kusch for kindly reading and criticising an earlier draft of this paper. They are not, of course, to be held responsible for the arguments and opinions advanced above.

David Bloor
University of Edinburgh

NOTES

[1] An important general survey of historical case-studies, analyzed from a sociological perspective, is provided by Shapin (1982). Empirical, and again mainly historical, monographs which have appeared since Shapin's review include: Collins (1985), Desmond (1989), Harwood (1993), Kusch (1995), Kusch (1999), Pickering (1984), Pinch (1986), Richards (1988), Rudwick (1985), Shapin (1994), Shapin and Schaffer (1985). It should go without saying that not all of these authors represent the same theoretical standpoint. Theoretical and philosophical discussions include Barnes (1988), Barnes, Bloor and Henry (1996) and Bloor (1997a). For an exchange between sociologists of knowledge, of the strong

program persuasion, and philosophers of science, see Brown (1984). Two significant methodological papers which are to be recommended are Barnes (1991) and Barnes (1993).

[2] There is room for discussion over where Mannheim's work stands in relation to the current sociology of scientific knowledge. The picture that I have just given has been challenged. In an interesting article Kaiser (1998) describes Mannheim's neo-Kantian background and his agreement with Dilthey's claim that there is a qualitative difference between the methods of humanistic scholarship and the natural sciences. Mannheim was opposed to scientism in sociology and saw science and technology (or, at least, the contemporary employment of them) as a moral and cultural threat. Kaiser goes on to express doubts about the way Mannheim's position has been assimilated into current discussions by Merton and the present writer. In Bloor (1991) I said Mannheim had shown a failure of nerve in accepting that sociological explanation could not be applied to the content of scientific and logical thinking. Kaiser says this gets Mannheim back to front. He did not fail in this respect because he was not trying to do this: science was only of marginal concern to him. The point is a reasonable one and reminds us of the remorseless tendency to modify a tradition of work as it becomes assimilated to new circumstances and new goals. Two points, however, need to be made about Kaiser's discussion. First, the central question is: has any actual error has been committed by using Mannheim in this way (i.e. as a symbol of the reluctance to press the case for an unrestricted sociology of knowledge)? If it has, Kaiser does not put his finger on what it is. Mannheim's position may have been taken out of context but it has not been misrepresented. Secondly, whatever Mannheim meant or wanted to do, the fact remains that he did not give a sociological analysis of the immanent development of ideas and implied that such an exercise was not viable. Kaiser concedes this vital point in footnote 22, p.76, towards the end of the paper.

[3] It should perhaps be noted that Stark's (1938) argument, at least as it appeared in the article cited by Merton, does not actually contain any such crude violation of the norm of universalism. It is presented as a claim about the relative virtues of abstract theorizing versus concrete experimentation, and as a condemnation of the "spirit of dogmatism" that, allegedly, all too frequently attends the former. The sinister aspect of the argument lies in the claim, guarded though it is, that these cognitive styles, the dogmatic spirit and the pragmatic spirit, have a racial grounding: "I wish solely to make a statement as to the frequency of occurrence of the natural tendency to pragmatic or dogmatic ways of thinking" (p.772). Some of Stark's argument amounts to an affirmation of the norm of universalism: "I acknowledge scientific achievement in new discoveries irrespective of the nationality of the discoverer..." (p.772). He speaks of the "inherent laws" of nature which are "independent of human existence", and concludes that "the object of physical science is international" (p.770). His claim is that the way research is carried out and described is a function of the character and mentality of those who do the research. Should the extent to which Stark endorses the norm be dismissed as mere hypocrisy? This is tempting, but if we left the matter there we should miss an important point. That Stark is able to argue in this way tells us something, not just about Stark, but about Merton's norms. It reminds us that the application of verbalized maxims, principles, rules and norms is endlessly problematic and negotiable. Behavior is underdetermined by such formulations. This applies to any and every use of them, not just when they come from the mouth of the disingenuous or corrupt. Here we have an important limitation in the project of explaining behavior by reference to norms. This is the weak point in the program of sociological explanation as Merton and his followers conceive it. It can also serve to make another important point. The alternative to explaining behavior by reference to the internalization of general and abstract norms is to see it as modeled on concrete and particular examples. This was Kuhn's (1962) approach – see, in particular, what he said about "the priority of paradigms", i.e. the priority of instances over general rules. (It was also the later Wittgenstein's approach.) Those such as Restivo (1995) p.100, who assert or intimate that

there is some sort of unity of approach between Merton and Kuhn are in danger of missing this fundamental difference.

[4] Arguably Duhem did have an ideological axe to grind derived from his religious faith. He was also not above nationalistic polemics against German scientists in the context of the First World War. He also made some mordant comments on the relative strengths and weaknesses of the English and French minds. I do not, therefore, want to pretend that he was the realization of some impossible ideal of free-floating detachment. Indeed, Duhem's polemical purposes may well have added to the sharpness of his observations on the nature of scientific thinking. For present purposes all that I need is that, human limitations accepted, his work can be acknowledged as a high-quality source of insight.

[5]I wonder if, on some level, Longino does know that there is a price to be paid? Perhaps this is why she prepares the ground by saying that she, "offers an account of knowledge as partial, fragmentary, and ultimately constituted from the interaction of opposed styles and/or points of view.“ (p. 199) This is identified as "postmodernist in spirit“ (p.199). Whatever it is called, it recommends a form of knowledge that lacks systematic cognitive virtues which, rightly or wrongly, are widely taken for granted.

[6] Here is a typical and crucial passage from Longino's paper: "What we are looking for in the account of objectivity is a way to block the influence of subjective preference (read: ideology) at the level of the background assumptions involved in observation and inference, and of individual variation in perception at the level of observation. The possibility of criticism does not totally eliminate subjective preference either from an individual's or from a community's practice of science. It does, however, provide a means of checking its influence in the formation of knowledge, for as long as background assumptions can be articulated and subjected to criticism from the scientific community, they can be defended, modified, or abandoned, in response to such criticism.“ (p. 208) Notice how the introduction of competing voices is assumed to work in the direction of diminishing subjective preference. There is no hint that such competition itself amounts to an ideological preference, or represents a substantial commitment to a certain style of knowing the world. This is like saying that you are playing party politics while I, of course, simply act for the good of the country.

[7]Wittgenstein also emphasized the importance of 'paradigms,' and meant by that term something similar to Kuhn. The significance of these ideas (in connection with what Anscombe called Wittgenstein's "linguistic idealism“) is analyzed in Bloor (1996a).

[8]It has to be admitted that not all sociologists of knowledge, and certainly not all those who are perceived as sociologists of knowledge, are as careful as they should be in this respect. I shall discuss a specific case of this, though one that arises through strategy not carelessness, later in the paper in connection with the work of Latour. The widespread conviction that sociologists subscribe to some species of idealism may have been reinforced by the habit of running together reality and the description of reality. Lest philosopher think this is a specific weakness of sociologists, it is worth remembering that many so called 'realists' commit this sin as well when they assume that reality has a privileged description which yields 'the truth' about it. I have also lost count of the number of times philosophers of science have slid, seemingly unaware, between 'observation' and 'observation report'.

[9]For a discussion of finitism see chapters 3,4 and 5 of Barnes, Bloor and Henry (1996). For the case of rule-following in Wittgenstein's philosophy, and its finitistic character, see chapter 2 of Bloor (1997a). The central point of finitism is that all meaning must be grounded in ostensive training and ostensive definition, and ostensive definition depends on an exposure to a finite number of examples. Verbal definitions simply presuppose words which have themselves been given an ostensive definition, so they too take us back to a finite number of examples. The problematic character of the move to the next case is therefore inescapable.

[10] This is what the Wittgensteinian slogan 'meaning is use' really amounts to.

[11] The crucial point about subjectivity is that for a would-be concept user who is socially isolated, there is no difference between an act of concept application being right and the user merely thinking it is right. Normativity becomes impossible to explain, and without normativity there is no genuine content to a (would-be) concept. See Bloor (1996)

[12] This case is documented in Hashimoto (1990). It should be emphasized that his argument and data go more deeply into the social structuring of knowledge than is indicated by my selection of this particular example.

[13] It is important to keep in mind that these comments apply to the inductive strategy of the group as a whole, not to the psychology of individual members of the group. An individualistic group may well contain individuals who make bold, subjective generalizations. Nevertheless, the end result could well be that the shared knowledge of the group will be low-level, empirical and lacking in bold extrapolations. (Think of Kuhn's pre-paradigm science.) There is a comparison to be drawn here with moral knowledge. Individualistic societies do not lack individuals with strong, subjective intuitions. Quite the contrary, intensity of subjective feeling may be taken to be the essence of morality. The point is that such a group will not be able to mobilise itself as a collective moral entity. Coming back to inductive knowledge of physical reality, there are thus circumstances in which the λ associated with a group will be at the opposite end of the scale to that which suggests itself if the parameter is interpreted psychologically and individualistically.

[14] An answer might be forthcoming if Campbell had meant his question to refer to local organizational preferences and short-term decisions, but the thrust of the paper is towards something with greater epistemological sweep than these pragmatic concerns. If one is looking for guidance on questions of practical judgement there are studies of research groups with very different patterns of organisation and interaction. These could be examined to see if they provide models that one wished to emulate or avoid. Edge and Mulkay's (1976) study of the Cambridge and Manchester groups who pioneered radio-astronomy could be read in this light, though this was not why it was written. Campbell would, perhaps, have found some evidence here to support his intuitive preference for small, democratic groups.

[15] For a discussion of the Popper-Kuhn debate in these terms see Bloor (1991).

[16] J. R. Searle (1995) gives a good analysis of why the question of realism should be kept separate from that of truth. Ch. 9 of Searle's book gives a defense of the correspondence theory which seems to me convincing, but to represent a very weak form of the theory. Other discussions of the correspondence theory have reached similar conclusions, but construed them as criticisms, see e.g. Bloor (1991) ch.2.

[17] I have said more about the relation between the strong program and Cambridge experimental psychology, represented here by Broadbent, in Bloor (1997b). The fuller discussion indicates the significance of the work of Sir F.C. Bartlett.

[18] Collins (1992) offers, as a methodological argument, the suggestion that sociologists should proceed as if the input from the material world were minimal. His methodological idealism tends to be read by critics as if it were a full, ontological idealism. It has also been treated, wrongly, as emblematic of the entire field of work in the sociology of knowledge. Even to point out the distinction between the methodological and ontological has been treated as if it were a mere evasion. In reality Collin's suggestion has always been rejected by supporters of the strong programme. For a discussion of its methodological weaknesses see Barnes, Bloor and Henry (1996) ch.1.

[19] On the realism and materialism of the strong programme see Barnes' (1992) paper 'Realism, Relativism and Finitism,' and Bloor (1991) ch. 2 'Sense Experience, Materialism and Truth' (first published 1976). Further discussions which may help offset this inexplicable, and seemingly invincible, misunderstanding are Barnes (1991) 'How Not To Do the

Sociology of Knowledge', Barnes (1993) 'How To Do the Sociology of Knowledge' and Bloor (1996b) 'Idealism and the Sociology of Knowledge'.

[20]There are certain sorts of psychological experiment where, at first glance, it might look as if the psychologist is able to discover how much of our experience is contributed by the object seen and how much by the seeing subject. In these cases it may seem as if the study of the visual system follows a Cole-like pattern of the kind I dismissed. Consider the famous Muller-Lyer illusion. This involves two lines of equal length with arrow-heads at the ends. On one line the arrow heads point outwards, on the other they point inwards. This makes the lines look unequal, the one with arrow heads pointing inwards typically looking longer. It is possible to measure the extent of the illusion. This is done by giving the subject an adjustable comparison line, without arrow-heads. The subject then judges the lines in the illusion against the comparison line, adjusting the comparison line until it matches their apparent length. The apparatus is described in Gregory (1966) p.158. The psychologist can then say that the illusion alters the length of the line by such and such a percent. Does not this answer a question of the kind I said was too confused to admit of an answer? No. The psychologist's result, which is perfectly legitimate, is telling us exactly what it says, namely, that the illusion alters the perceived length of the line by such and such an extent. This answer does not have the form required to be an answer to Cole's question. It does not say that such and such a proportion of the experience of the illusion was contributed by the world and such and such by the visual system. The proportionality revealed in the experiment refers to the length of the line that is experienced, not the relative contribution of the two basic factors involved in generating the experience. Suppose the line is really x centimetres long but appears to be x+d centimetres. We cannot conclude that "nature" contributed the experience of the x centimetres and "the brain" contributed the experience of the d centimetres. The ratio d/x does not tell us how much the experience is "influenced or constrained" by reality and how much by the brain. That formulation still lacks a clear sense and does not correctly capture the methodology or experimental rationale of the psychologist.

[21] On the theme of critical distance, and the practical difficulties in maintaining it, see the special issue on 'The politics of SSK', Social Studies of Science 26,no.2, 1996. Sociologists who have tried to produce a symmetrical analysis of highly charged controversies, such as the dispute over the efficacy of large doses of vitamin C in the treatment of cancer, find themselves inexorably drawn into the conflict. One side or the other will often seek to capture the support of the sociologist. Scott, Richards and Martin (1990) suggest that, typically, marginal or minority groups, such as those around Linus Pauling in the vitamin C case, will try to "capture" the sociologist. Central, high-status groups will typically reject the sociologist's analysis. Outsiders, it seems, feel they have more to gain from being the subject of a symmetrical study than do insiders. But if the reception of a symmetrical analysis will typically be asymmetrical, why bother to be impartial at all? Why not bow to the inevitable, throw symmetry aside, and become an advocate? Three points need to be made in response. First, the asymmetrical reception of a symmetrical analysis is entirely predictable. Who ever expected anything else? Second, as Collins has pointed out, the laws governing the reaction to a symmetrical analysis are almost certainly more complicated than the principle suggested above, that outsider groups will be positive and insider groups negative. Collins (1996) presents a counter-example to this from his gravity-wave research. Third, the symmetry postulate is a methodological postulate. It states a condition of adequacy for an explanation, not an injunction about the researcher's attitudes or sympathies. Those who advocate commitment and conclude, on this basis, that the symmetry principle must be rejected are running together questions which should be kept apart. Martin (1996) states the point correctly at the end of an interesting account of his growing involvement in the debate over the origin of Aids. Impartiality and symmetry, he says, "apply to *explanations of beliefs*. The method of the SP [strong programme, DB] analyst should be to use the same sorts of

explanations to explain (what are taken to be) different sorts of belief. These tenets say nothing about the personal beliefs or engagement of the analyst" (p.266).

[22] Hume (1739) wanted to introduce the experimental method of reasoning into the study of moral subjects, as the title page of the *Treatise* explicitly states. He did not use the language of 'social construction', but his terminology was very close: he spoke of moral categories such as property and the obligation to keep promises as 'artificial virtues', so they were artifacts rather than constructs. They were also social artifacts as his account of convention, in Bk.III, makes clear..

[23] The social process of paradigm consolidation is structurally identical to a number of other important transitions, such as the "lock-in" of one technology and the "lock-out" of a competing technology, or the strongly skewed geographical location of industry, say, in the north rather than the south of a country. The mathematical economist Brian Arthur has shown that these processes, which all involve positive feedback loops, can be represented in a precise mathematical model, see Arthur (1984). The important feature of the model, for the present discussion, is that it displays the growth and change of an interest over time. It deals with 'emergence', and yet emergence is precisely what Pickering believes is missing from the disciplinary perspective of the sociologist. This work, and the entire class of cases with which it deals, thus provides further evidence of the falsity of his charge.

[24]Haack gives a list of "potential hindrances" to good science. These include "pressure... to solve problems perceived as socially urgent" and pressure "to ignore questions perceived as socially disruptive" (p.261). The qualification "potentially" presumably means that these may, but will not necessarily, be hindrances. But couldn't they be a positive stimulus under some circumstances? It is worth noting just how difficult it is to produce general answers to questions about what will be good and what will be bad for science. Taking the first of the two potential hindrances just mentioned, it would be difficult to think that the enormous stimulus given to scientific research by the demands of two World Wars was not accompanied by pressure which was consequent on the felt urgency of the situation. As to the second hindrance, consider the consequences of the Allied refusal to let Germany build powered aircraft in the years immediately after World War I. Here the socially disruptive consequences that were at issue concerned the ability of Germany to wage any future war. The pressure in question amounted to a ban imposed under the terms of the Versailles treaty. The result was that brilliant aerodynamicists such as Theodore von Kármán turned their skills to designing and studying gliders. This may well have helped the science of aerodynamics forward: it certainly did not seem to hold it back, and it also provided a stimulus to the study of meteorology. See von Kármán (1967). On the enormous symbolic and ideological significance assumed by the glider see Fritzsch (1992) ch.3.

[25]This is a variant of a familiar theme to be found in the writings of, amongst others, Lakatos (1971) and Laudan (1977) For some observations on the theological precedents of this manichean division of labor see Bloor (1989) and the Afterword to Bloor (1991). For a devastating criticism of Laudan see Barnes (1979).

[26] "Evidence never *obliges* us to accept this claim rather than that, the thought is, and we have to accept something; so acceptance is always affected by something besides the evidence." Haack (1996) p. 263.

[27] "Not all scientific claims are accepted as definitely true or rejected as definitely false, nor should they be; indeed, keeping warrant and acceptance appropriately related requires, inter alia, that, when the evidence is insufficient, we acknowledge that we don't know." Haack (1996) p. 263.

[28]Haack prefaces her paper by a quotation from Stove (1991) who is the source of the words "stupid and discreditable". The prefatory quotation contains a version of the claim that sociologists of knowledge are guilty of self-contradiction and special pleading because (allegedly) they are saying that everybody is determined by their class situation except the

sociologist who has miraculously transcended it. Haack calls this a "shrewd observation" which "identifies exactly what is wrong" with the sociology of knowledge (p.259). For a demolition of this familiar and over-worked ploy see Hesse (1980) ch.2 and Herrnstein Smith (1997) ch. 5.

[29] Laudan made a similar move when he glossed the symmetry postulate in terms of an alleged claim about "completely homogeneous" causes. See Laudan (1984 p. 52). For a reply see Bloor (1984).

REFERENCES

Arthur, W.: 1984, 'Competing Technologies and Economic Prediction', *Options* **2**, 10-13.

Barnes, B.: 1979, 'Vicissitudes of Belief', *Social Studies of Science* **9**, 247-263.

Barnes, B.: 1982, *T. S. Kuhn and Social Science*, Macmillan, London.

Barnes, B.: 1984, 'Problems of Intelligibility and Paradigm Instances', in J. Brown (ed.) *Scientific Rationality: The Sociological Turn*, Reidel, Dordrecht, pp. 113-125.

Barnes, B.: 1988, *The Nature of Power*, Polity Press, Cambridge

Barnes, B.: 1991, 'How Not To Do the Sociology of Knowledge', *Annals of Scholarship* **8**, no.3, 321-335.

Barnes, B.: 1992, 'Realism, Relativism and Finitism', in D. Raven, L.V. Thyssen and J. de Wolf (eds.), *Cognitive Relativism and Social Science*, Transaction Publications, London, pp. 131-147.

Barnes, B.: 1993, 'How To Do the Sociology of Knowledge', *Danish Yearbook of Philosophy* **28**, 7-23.

Barnes, B. and D. Bloor: 1982 'Relativism, Rationalism and the Sociology of Knowledge' in M. Hollis and S. Lukes (eds.), *Rationality and Relativism*, Blackwell, Oxford, pp. 21-47.

Barnes, B., D. Bloor and J. Henry: 1996, *Scientific Knowledge: A Sociological Analysis*, Athlone, London and Chicago University Press, Chicago.

Bloor, D.: 1981, 'Hamilton and Peacock on the Essence of Algebra', in H. Mehrtens, H. Bos and I. Schneider (eds.), *Social History of Nineteenth Century Mathematics*, Birkhauser, Boston, pp. 262-232.

Bloor, D.: 1989, 'Rationalism, Supernaturalism and the Sociology of Knowledge', in I. Hronsky, M. Féhér, B. Dajka, *Scientific Knowledge Socialised*, Reidel, Dordrecht, pp. 59-74.

Bloor, D.: 1991, *Knowledge and Social Imagery*, University of Chicago Press, Chicago.

Bloor, D.: 1996, 'The Question of Linguistic Idealism Revisited', in H. Sluga and D. Stern (eds.), *The Cambridge Companion to Wittgenstein*, Cambridge University Press, Cambridge, pp. 354-382.

Bloor, D.: 1997a, *Wittgenstein: Rules and Institutions*, Routledge, London.

Bloor, D.: 1997b, 'Remember the Strong Programme?', *Science, Technology and Human Values* **22**, 373-385.

Broadbent, D.: 1973, *In Defence of Empirical Psychology*, Methuen, London.

Brown, J.(ed.): 1984, *Scientific Rationality: The Sociological Turn*, Reidel, Dordrecht.

Campbell, D.: 1989, 'Models of Language Learning and their Implications for Social Constructionist Analyses of Scientific Belief,' in S. Fuller, M. de Mey, T. Shinn and S. Woolgar (eds.), *The Cognitive Turn: Sociological and Psychological Perspectives on Science*, Kluwer, Dordrecht, pp. 153-158.

Carnap, R.: 1952, *The Continuum of Inductive Methods*, University of Chicago Press, Chicago.

Cole, S.: 1992, *Making Science: Between Nature and Society*, Harvard University Press, Cambridge, Mass.

Cole, S.: 1996, 'Voodoo Sociology: Recent Developments in the Sociology of Science' in P. Gross, N. Levitt and M. Lewis (eds.), *The Flight from Science and Reason*, New York Academy of Sciences, New York, pp. 274-287.

Collins, H.: 1985, *Changing Order: Replication and Induction in Scientific Practice*, Sage, London.

Collins, H.: 1996, 'In Praise of Futile Gestures: How Scientific is the Sociology of Scientific Knowledge?', *Social Studies of Science* **26**, 229-244.

Conant, J.: 1966, 'The Overthrow of the Phlogiston Theory', in J. Conant and L. Nash (eds.), *Harvard Case Histories in Experimental Science*, vol. I, pp. 67-115.

Desmond, A.: 1989, *The Politics of Evolution: Morphology, Medicine and Reform in Radical London*, Chicago University Press, Chicago.

Duhem, P.: 1906, *The Aim and Structure of Physical Theory*, trans. P.P. Wiener, Athenium, New York, 1962.

Edge, D. and M. Mulkay: 1976, *Astronomy Transformed: The Emergence of Radio Astronomy in Britain*, Wiley, New York.

Feyerabend, P.: 1963, 'How to be a Good Empiricist – A Plea for Tolerance in Matters Epistemological', in B. Baumrin (ed.) *Philosophy of Science. The Delaware Seminar* vol. II, Interscience, New York, pp. 3-39.

Feyerabend, P.: 1988, *Against Method*, Verso, New York.

Fine, A.: 1996, 'Knowledge Made Up; Constructivists Sociology of Scientific Knowledge', in P. Galison and D. Stump (eds.), *The Disunity of Science*, Stanford University Press, Stanford, Cal., pp. 231-254.

Fritzsche, P.: 1992, *A Nation of Flyers: German Aviation and the Popular Imagination*, Harvard University Press, Cambridge, Mass..

Gottfried, K. and K. Wilson, K: 1997, 'Science a Cultural Construct', *Nature* **386**, 545-547.

Gregory, R.: 1966, *Eye and Brain. The Psychology of Seeing*, Weidenfeld and Nicolson, London.

Haack, S.: 1996, 'Towards a Sober Sociology of Science', in P. Gross, N. Levitt, and M. Lewis (eds.), *The Flight from Science and Reason*, New York Academy of Sciences, New York, pp. 259-265.

Hardwig, J.: 1991, 'The Role of Trust in Knowledge', *Journal of Philosophy* **88**, 693-708.

Hashimoto, T.: 1990, *Theory, Experiment, and Design Practice. The Formation of Aeronautical Research, 1909-1930*. Unpublished Ph.D. thesis, University of Baltimore, Maryland.

Haugeland, J.: 1990, 'The Intentionality All-Stars', in J. Tomberlin (ed.) *Philosophical Perspectives 4, Action Theory and Philosophy of Mind*, Ridgeview, Ataxadero, Cal, pp. 383-427.

Hon, G.: 1995, 'Is the Identification of Experimental Error Contextually Dependent? The Case of Kaufmann's Experiment and its Varied Reception', in J. Z.Buchwald (ed.), *Scientific Practice: Theories and Stories of Doing Physics*, University of Chicago Press, Chicago, pp. 170-223.

Harwood, J.: 1993, *Styles of Scientific Thought; The German Genetics Community 1900-1933,* University of Chicago Press, Chicago.

Herrnstein Smith, B.: 1997, *Belief and Resistance. Dynamics of ContemporaryControversy*, Harvard University Press, Cambridge, Mass.

Hesse, M.: 1974, *The Structure of Scientific Inference*, Macmillan, London.

Hesse, M.: 1980, *Revolutions and Reconstructions in the Philosophy of Science*, Harvester, Brighton.

Hollinger, D.: 1995 'Science as a Weapon in Kulturkampfe in the United States During and After World War II', *Isis*, vol. **86**, 440-454.

Hume, D.: 1739, *A Treatise of Human Nature*, ed. L.A. Selby-Bigge, Clarendon Press, Oxford (1960).

Kaiser, D.: 1998, 'A Mannheim for All Seasons: Bloor, Merton, and the Roots of the Sociology of Scientific Knowledge', *Science in Context*, vol. **11**, 51-87.

Kuhn, T.: 1962, *The Structure of Scientific Revolutions*, Chicago University Press, Chicago.

Kusch, M.: 1995, *Psychologism: A Case Study in the Sociology of Philosophical Knowledge*, Routledge, London.

Kusch, M.: 1999, *Psychological Knowledge. A Social History and Philosophy*, Routledge, London.

Lakatos, I.: 1971, 'History of Science and its Rational Reconstructions', in R. Buck and R. Cohen (eds.) *Boston Studies in the Philosophy of Science*, vol.**VIII**, 91-136.

Latour, B.: 1987, *Science in Action*, Harvard University Press, Cambridge, Mass.

Latour, B.: 1988, *The Pasteurization of France*, Harvard University Press, Cambridge, Mass.

Latour, B.: 1992, 'One More Turn After the Social Turn', in E. McMullin (ed.), *The Social Dimension of Science*, University of Notre Dame Press, Notre Dame, Ind., pp. 272-294.

Latour, B.: 1993, *We Have Never Been Modern*, Harvester Press, New York.

Laudan, L.: 1977, *Progress and its Problems: Towards a Theory of Scientific Growth*, Routledge and Kegan Paul, London.

Laudan, L.: 1984, 'The Pseudo-Science of Science?', in J. Brown (ed.), *Scientific Rationality: The Sociological Turn*, Reidel, Dordrecht, 41-73.

Longino, H.: 1992, 'Essential Tensions – Phase Two ; Feminist, Philosophical, and Social Studies of Science', in E. McMullin (ed.), *The Social Dimension of Science*, University of Notre Dame Press, Notre Dame, Ind., pp. 198-216.

MacKenzie, D.: 1981, *Statistics in Britain 1865-1930. The Social Construction of Scientific Knowledge*, Edinburgh University Press, Edinburgh.

Mannheim, K.: 1936, *Ideology and Utopia* (trans. L. Wirth and E. Shils), Routledge and Kegan Paul, London.

Martin, B.: 1996, 'Sticking a Needle into Science: The Case of Polio Vaccines and the Origin of AIDS', *Social Studies of Science* **26**, 245-276.

Merton, R.: 1942, 'Science and Technology in a Democratic Order', *Journal of Legal and Political Sociology* **1**, 115-126, reprinted as ch. 13 in *The Sociology of Science*, 1973, University of Chicago Press, Chicago.

Merton, R.: 1973 *The Sociology of Science. Theoretical and Empirical Investigations*, University of Chicago Press, Chicago.

Mills, E.: 1980, 'Alexander Agassiz, Carl Chun and the Problem of the Intermediate Fauna', in M. Sears and D. Merriman (eds.), *Oceanography: The Past, Proceedings of the Third International Congress on the History of Oceanography*, New York, pp. 360-372.

Osgood, C.: 1953, *Method and Theory in Experimental Psychology*, Oxford University Press, New York.

Pannekoek, A.: 1953, 'The Discovery of Neptune', *Centaurus* **3**, 126-137.

Pickering, A.: 1995, *Constructing Quarks: A Sociological History of Particle Physics*, Edinburgh University Press, Edinburgh.

Pickering, A.: 1995, *The Mangle of Practice: Time, Agency and Science*, University of Chicago Press, Chicago.

Pinch, T.: 1986, *Confronting Nature: The Sociology of Solar Neutrino Detection*, Reidel, Dordrecht.

Popper, K.: 1945, *The Open Society and its Enemies*, Routledge and Kegan Paul, London.

Popper, K.: 1959, *The Logic of Scientific Discovery*, Hutchinson, London.

Restivo, S.: 1995, *The Theory Landscape in Science Studies*, in S. Jasanoff, G. Markle, J. Petersen, T. Pinch (eds.), *Handbook of Science and Technology Studies*, Sage, London.

Richards, J.: 1988, *Mathematical Visions: The Pursuit of Geometry in Victorian England*, Academic Press, London.

Rudwick, M.: 1985, *The Great Devonian Controversy. The Shaping of Scientific Knowledge among Gentlemanly Specialists*, Chicago University Press, Chicago.

Scott, P., Richards, E. and Martin, B.: 1990, 'Captives of Controversy: The Myth of the Neutral Social Researcher in Contemporary Scientific Controversies', *Science, Technology and Human Values* **15**, 474-494.

Searle, J.: 1995, *The Construction of Social Reality*, Allen Lane, Harmondsworth, Middlesex.

Shapin, S.: 1982, 'History of Science and its Social Reconstructions', *History of Science* **20**, 157-211.

Shapin, S.: 1994, *A Social History of Truth: Civility and Science in Seventeenth-Century England*, University of Chicago Press, Chicago.

Shapin, S. and Schaffer, S.: 1985, *Leviathan and the Air-Pump. Hobbes, Boyle, and the Experimental Life*, Princeton University Press, Princeton.

Stark, J.: 1938, 'The Pragmatic and Dogmatic Spirit in Science', *Nature*, vol.**141**, 770-772.

Stove, D.: 1991, *The Plato Cult*, Blackwell, Oxford.

von Kármán, T.: 1967, (with Lee Edson), *The Wind and Beyond: Theodore von Kármán, Pioneer in Aviation and Pathfinder in Space*, Little Brown, Boston.

Warwick, A.: 1992-3, 'Cambridge Mathematics and Cavendish Physics: Cunningham, Campbell, and Einstein's Relativity, Pt.I The Uses of Theory, Pt. II Comparing Traditions in Cambridge Physics', *Studies in History and Philosophy of Science* **23**, 625-656, **24**, 1-25.

Wylie, A.: 1996, 'The Constitution of Archaeological Evidence: Gender Politics and Science', in P. Galison and D. Stump (eds.), *The Disunity of Science*, Stanford University Press, Stanford, Cal., pp. 311-343.

Zuckerman, H.: 1988, 'The Sociology of Science', in N.J. Smelser (ed.) *Handbook of Sociology*, Sage, London.

WOLFGANG LENZEN

EPISTEMIC LOGIC

INTRODUCTION

0.1 History of epistemic logic

The core meaning of the Greek word *episteme* is *knowledge*. Thus, taken literally, epistemic logic represents the logic of knowledge. In modern philosophy, however, *epistemic logic* is used as a technical term not only for the logic of knowledge but also for the logic of belief, (although the latter might more appropriately be referred to as *doxastic logic* from the Greek *doxa* to mean *belief*).

Like logic in general, also epistemic logic in particular may be said to have been founded by Aristotle. This is true at least in the sense that several passages in *De Sophisticis elenchis* and in the *Prior* and *Posterior Analytics* deal with basic issues of what is nowadays conceived of as epistemic logic. More detailed investigations of principles of epistemic logic may be found in the manuals of Medieval authors such as Buridanus, Burleigh, Ockham, and Duns Scotus (cf., e.g., Chisholm 1963 and Boh 1986). However, systematic calculi of epistemic logic have only been developed after the elaboration of possible-worlds-semantics in the mid of our century. The most important works to be mentioned here comprise Carnap 1947, Kripke 1959, and Hintikka's pioneering *Knowledge and Belief* of 1962. Further steps towards the establishment of epistemic logic as a particular branch of modal logic have been taken by Kutschera 1976 and by Lenzen 1980a.

What is common to these approaches is that they remain *static* in character, i.e. they only describe the "logical" structure of the belief- or knowledge-system of a certain subject a at a certain time t. The basic principles for the *dynamics* of epistemic systems have been investigated esp. by Gärdenfors 1988 (cf., e.g., the contribution "Revision of belief systems" in section C III of this Handbook). Another generalization of epistemic logic has recently been attempted in the field of computer science (cf. Fagin et al. 1994) where one tries to model in particular the effects of communication between n subjects a_i for the joint knowledge of a "distributed system" $S=\{a_1,...,a_n\}$. Such considerations, however, fall outside the scope of this paper which only aims at describing, in barest outlines, the basic laws for propositional logics of *belief*, *knowledge*, and *conviction* and at discussing some selected issues related to "quantifying in" epistemic contexts.

0.2 Methodology of epistemic logic

Although epistemic logic exists as a branch of philosophical logic for quite a long time, it remains to be explained in which sense of the word *logic* epistemic logic

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 963-983.
© 2004 *Kluwer Academic Publishers.*

constitutes a logic at all, or – to put it in the form of the sceptical question of Hocutt 1972 – "Is epistemic logic possible?". The general problem behind this question may be illustrated as follows. Take any propositional attitude, $\phi(a,p)$, which a certain subject a bears towards a proposition (or a state of affairs expressed by the proposition) p; let another proposition q be logically equivalent to p, $\vdash p \leftrightarrow q$. Then there appears to be no "logical" guarantee that a bears the same attitude ϕ also towards proposition q, for it seems always possible that a does not "see" (and hence doesn't know) that p and q are logically equivalent. Thus, in a certain sense, the following situation always seems possible: $\vdash p \leftrightarrow q$, but $\phi(a,p) \wedge \neg\phi(a,q)$, i.e. not $\phi(a,p) \leftrightarrow \phi(a,q)$. But then even most elementary "laws" such as, e.g.,

(CLOS1) $\phi(a,p\wedge q) \leftrightarrow \phi(a,q\wedge p)$

(CLOS2) $\phi(a,p\vee p) \leftrightarrow \phi(a,p)$

or

(CLOS3) $\phi(a,p) \leftrightarrow \phi(a,\neg\neg p)$

would not be valid, and one could hardly find *any* epistemic logical law which adequately describes the factual knowledge- or belief-system of an arbitrary subject, a.

However, this sceptical conclusion rests on a very narrow conception of our everyday's attribution of propositional attitudes. When in the preceding paragraph the possibility was granted that a person a might not "see" that two logically equivalent propositions p and q are in fact logically equivalent, the ascription of $\phi(a,p)$ and the non-ascription of $\phi(a,q)$ will usually be based on a's *verbal behaviour*. When asked whether (she believes that) p is true, a answers in the affirmative, while when asked whether (she believes that) q, a happens to answer in the negative. Now, even if one assumes that the answers were intended quite sincerely, there remain several sources for a possible clash between what *a said* and what she really *believed*. She may have misunderstood one or the other question; one of the answers may be the result of a slip of tongue; etc. In any case, the very fact that p and q are logically equivalent and hence "mean the same thing" strongly suggests that a did not fully *understand* the meaning of p and/or q.

In everyday's discourse, however, we standardly presuppose that the people with which we talk have an adequate understanding of what is said. Therefore we assume that their belief- or knowledge-systems satisfy certain conditions of *rationality*, in particular a certain amount of logical consistency and deductive closure.[1] In this sense one may consider the task of epistemic logic to consist (1) in elaborating the "logical" laws which one may *rationally expect* the belief- and knowledge-system of a subject a to obey and (2) in clarifying the analytical relations that exist between these epistemic attitudes. In the following section the former laws will be represented by sets of axiomatic principles **B1-B7**, **C1-C11**, and **K1-K8** (for the logic of Belief, Conviction, and Knowledge, respectively), while the epistemic laws interrelating these notions will be denoted as **E1-E12**. A more systematic exposition of the syntax and semantics of corresponding formal calculi may be found in Lenzen 1980a.

I THE LOGIC OF BELIEF

In the vast majority of publications on epistemic logic it is tacitly presupposed that only one unique concept of belief has to be investigated. However, as was first argued in Lenzen 1978, at least two different concepts of belief – which display a quite distinct logical behaviour – must be carefully distinguished: "strong" and "weak" belief.

1.1 The logic of "strong belief"

Let '$C(a,p)$' abbreviate the fact that person *a* is firmly convinced that *p*, i.e. that *a* considers the proposition *p* (or, equivalently, the state of affairs expressed by that proposition) as absolutely certain; in other words, *p* has maximal likelihood or probability for *a*. Using 'Prob' as a symbol for subjective probability functions, this idea can be formalized by the requirement:

(PROB-C) $C(a,p) \leftrightarrow \mathrm{Prob}(a,p)=1$.

Within the framework of standard possible-worlds semantics $<I,R,V>$, $C(a,p)$ would have to be interpreted by the following condition:

(POSS-C) $V(i,C(a,p))=\mathbf{t} \leftrightarrow \forall j(iRj \rightarrow V(j,p)=\mathbf{t})$.

Here *I* is a non-empty set of (indices of) possible worlds; *R* is a binary relation on *I* such that *iRj* holds if and only if (or, for short, iff) in world *i*, *a* considers world *j* as possible; *V* is a valuation-function assigning to each proposition *p* relative to each world *i* a truth-value $V(i,p)\in\{\mathbf{t},\mathbf{f}\}$. Thus $C(a,p)$ is true (in world $i\in I$) iff *p* itself is true in every possible world *j* which is considered by *a* as possible (relative to *i*).

The probabilistic definition **POSS-C** together with some elementary theorems of the theory of subjective probability immediately entails the validity of the subsequent laws of conjunction and non-contradiction. If *a* is convinced both of *p* and of *q*, then *a* must also be convinced that *p* and *q*:

(C1) $C(a,p) \wedge C(a,q) \rightarrow C(a,p\wedge q)$.

For if both $\mathrm{Prob}(a,p)$ and $\mathrm{Prob}(a,q)$ are equal to 1, then it follows that $\mathrm{Prob}(a,p\wedge q)=1$, too. Furthermore, if *a* is convinced that *p* (is true), *a* cannot be convinced that $\neg p$, i.e. that *p* is false:

(C2) $C(a,p) \rightarrow \neg C(a,\neg p)$.

For if $\mathrm{Prob}(a,p)=1$, then $\mathrm{Prob}(a,\neg p)=0$, and hence $\mathrm{Prob}(a,\neg p)\neq 1$. Just like the alethic modal operators of possibility, $\Diamond$, and necessity, $\Box$, are linked by the relation $\Diamond p \leftrightarrow \neg\Box\neg p$, so also the doxastic modalities of thinking *p* to be possible – formally: $P(a,p)$ – and of being convinced that *p* satisfy the relation

(Def. P) $P(a,p) \leftrightarrow \neg C(a,\neg p)$.

Thus, from the probabilistic point of view, $P(a,p)$ holds iff a assigns to the proposition p (or to the event expressed by that proposition) a likelihood greater than 0:

(PROB-P) $V(P(a,p))=\mathbf{t} \leftrightarrow \mathrm{Prob}(a,p)>0$.

Within the framework of possible-worlds semantics, one obtains the following condition:

(POSS-P) $V(i,P(a,p))=\mathbf{t} \leftrightarrow \exists j(iRj \wedge V(j,p)=\mathbf{t})$,

according to which $P(a,p)$ is true in world i iff there is at least one possible world j – i.e. a world j which a considers as possible relative to i – in which p is true.

In view of **Def P**, the former principle of consistency, **C2**, can be paraphrased by saying that whenever a is firmly convinced that p, a will a fortiori consider p as possible. However, considering p as possible does not conversely entail being convinced that p. In general there will be many propositions p such that a considers both p and $\neg p$ as possible. Such a situation, where $P(a,p) \wedge P(a,\neg p)$, makes clear that unlike the operator C, P will not in general satisfy a principle of conjunction analogous to **C1**. However, the converse entailment

(C3) $P(a,p\wedge q) \rightarrow P(a,p) \wedge P(a,q)$

and its counterpart

(C4) $C(a,p\wedge q) \rightarrow C(a,p) \wedge C(a,q)$

clearly are valid, because the probabilities of the single propositions p or q always are at least as high as the probability of the conjunction $(p\wedge q)$. Similarly, since the probability of a disjunction $(p\vee q)$ is always at least as high as the probabilities of the single disjuncts p and q, it follows that both operators C and P satisfy a corresponding principle of disjunction:

(C5) $C(a,p) \vee C(a,q) \rightarrow C(a,p\vee q)$
(C6) $P(a,p) \vee P(a,q) \rightarrow P(a,p\vee q)$.

Now the probabilistic "proofs" of such principles are not without problems. Since its early foundations by de Finetti 1964, the theory of subjective probability has always been formulated in terms of *events,* while in the framework of philosophical logic attitudes like $C(a,p)$ are traditionally formulated in terms of *sentences.* So if one wants to apply the laws of the theory of subjective probability to the field of cognitive attitudes, one has to presuppose (i) that for every event X there corresponds exactly one proposition p, and (ii) that the cognitive attitudes really are "propositional" attitudes in the sense that their truth is independent of the specific linguistic representation of the event *X*. That is, whenever two sentences *p* and *q* are logically equivalent and thus describe one and the same event *X*, then *C*(*a*,*p*) holds iff *C*(*a*,*q*) holds as well. This requirement can be formalized by the following rule:

(C7) $p \leftrightarrow q \vdash C(a,p) \leftrightarrow C(a,q).$

This principle further entails that everybody must be convinced of everything that logically follows from his own convictions:

(C8) $p \rightarrow q \vdash C(a,p) \rightarrow C(a,q).$

For if *p* logically implies *q*, then *p* is logically equivalent to $p \wedge q$; thus *C*(*a*,*p*) entails $C(a,p \wedge q)$ (by **C7**) which in turn entails *C*(*a*,*q*) by **C4**.

As was already stressed in section 0.2 above, there has been a long discussion whether and to which extent the epistemic attitudes of real subjects are *deductively closed.* In view of man's almost unlimited fallibility in matters of logic, some authors have come to argue that **C8** should be restricted to very elementary instances like **C4** or **C5** or to some other so-called 'surface tautologies' (cf., e.g., Hintikka 1970a). Which option one favours will strongly depend on the methodological role that one wants to assign to epistemic logic. If epistemic logic is conceived of as a *descriptive system* of people's factual beliefs, then not even the validity of the most elementary principles like **C4** seems warranted. If, on the other hand, epistemic logic is viewed as a *normative* system of *rational* belief, then even the strong condition of full deductive closure, **C8**, appears perfectly acceptable. Incidentally, if one presupposes that everybody has at least one conviction – an assumption which is logically guaranteed by some of the subsequent iteration-principles[2] – **C8** entails the further rule

(C9) $p \vdash C(a,p),$

according to which everybody is convinced of every tautological proposition (or state of affairs) *p*.

To round off our exposition of the logic of conviction, let us consider some laws for *iterated* epistemic attitudes. According to the thesis of the "privileged access" to our own mental states, whenever some person *a* is convinced of *p*, *a* knows that she has this conviction. Similarly, if *a* is not convinced that *p*, i.e. if she considers *p* as possible, then again she knows that she considers *p* as possible:

(E1) $C(a,p) \rightarrow K(a,C(a,p))$

(E2) $\neg C(a,p) \rightarrow K(a,\neg C(a,p)).$

Here '*K*(*a*,*q*)' abbreviates the fact that *a knows* that *q*. Now, clearly *a* knows that *q* only if in particular *a* is convinced that *q*:

(E3) $K(a,p) \rightarrow C(a,p).$

Hence one immediately obtains the following purely doxastic iteration-principles

(C10) $C(a,p) \rightarrow C(a,C(a,p))$
(C11) $\neg C(a,p) \rightarrow C(a,\neg C(a,p))$.

It is easy to verify that the implications **C10** and **C11** may be strengthened into equivalencies. Generally speaking, iterated doxastic operators or "modalities" are always reducible to simple "modalities" of the types $C(a,p)$ and $\neg C(a,q)$, where p and q contain no further doxastic expressions. As a matter of fact, iterated doxastic *propositions* of arbitrary complexity can be reduced to simple, non-iterated propositions. In the end, then, the logic of conviction turns out to be structurally isomorphic to the "deontic" calculus **DE4** of Lemmon 1977 which differs from the better-known alethic calculus **S5** only in that it does not contain the "truth-axiom" $\Box p \rightarrow p$. Given the intended doxastic interpretation of "necessity" as *subjective* necessity or certainty, the failure of $C(a,p) \rightarrow p$ comes as no surprise. After all, humans are not infallible; therefore someone's conviction that p – however firm it may be – can never logically guarantee that p is in fact the case.

1.2 The logic of "weak belief"

While the concept of conviction, $C(a,p)$, has been defined above to obtain iff person a is absolutely *certain* that p, the more general concept of "weak" *belief*, $B(a,p)$, will be satisfied by the much more liberal requirement that person a only considers p as *likely* or as *probable*. Here the lower bound of (subjective) probability may reasonably be taken to be .5. In other words, person a believes that p iff she considers p as more likely than not:

(PROB-B) $B(a,p) \leftrightarrow \text{Prob}(a,p)>1/2$.

This "weak" notion of belief also satisfies the principle of non-contradiction analogous to **C2**:

(B1) $B(a,p) \rightarrow \neg B(a,\neg p)$.

Clearly, if p has a probability greater than 1/2, then $\neg p$ must have a probability less than 1/2. On the other hand, $B(a,p)$ does not satisfy the counterpart of conjunction principle **C1**, because even if two single propositions p and q both have a probability $> .5$, it may well happen that $\text{Prob}(a,p\wedge q)$ is $< .5$. For instance, let an urn contain two black balls and one white ball where one of the black balls is made of metal while the white ball and the other black ball is made of wood. Now if just one ball is drawn from the urn at random, the probability of p = 'The ball is black' equals 2/3 and is thus > 1/2; also the probability of q = 'The ball is made of wood' is 2/3 > 1/2. But the probability of the joint proposition $(p\wedge q)$ = 'The ball is made of wood and is black' only is 1/3.

It follows from the theory of probability that conjunctivity of belief is warranted only in the special case where one of the two propositions is *certain*:

(E4) $B(a,p) \wedge C(a,q) \rightarrow B(a,p \wedge q)$.

Here certainty may be said to represent a special instance of belief in the sense of:

(E5) $C(a,p) \rightarrow B(a,p)$.

The validity of this principle derives from the fact that each proposition *p* with maximal probability 1 *a fortiori* has a probability greater than .5! Thus, *semantically* speaking, *a*'s believing that *p* is entirely compatible with *a*'s being absolutely certain that *p*, although from a *pragmatic* point of view when person *a* says 'I believe that *p*', she thereby expresses that she is *not convinced* that *p*.[3]

The epistemological thesis of the privileged access to (or the privileged knowledge of) our own mental states mentioned earlier in connection with principles **E1** and **E2** evidently applies not only to the particular doxastic attitude $C(a,p)$, but to the more general notion $B(a,p)$ as well. Thus, whenever person *a* believes that *p*, *a* knows that she believes that *p*; and, conversely, if she does not believe that *p*, she knows that she does not believe that *p*:

(E6) $B(a,p) \rightarrow K(a,B(a,p))$
(E7) $\neg B(a,p) \rightarrow K(a,\neg B(a,p))$.

In view of **E3** and **E5,** one immediately obtains the following "pure“ iteration-laws:

(B2) $B(a,p) \rightarrow B(a,B(a,p))$
(B3) $\neg B(a,p) \rightarrow B(a,\neg B(a,p))$.

Furthermore the rules of deductive closure of belief:

(B4) $p \leftrightarrow q \vdash B(a,p) \leftrightarrow B(a,q)$
(B5) $p \rightarrow q \vdash B(a,p) \rightarrow B(a,q)$
(B6) $p \vdash B(a,p)$

can be justified in strictly the same way as the corresponding principles for conviction.

In order to obtain a complete axiomatization of the logic of "weak" belief, one has to introduce the somewhat unfamiliar relation of "strict implication" between *sets of propositions* $\{p_1,...,p_n\}$ and $\{q_1,...,q_n\}$ ($n \geq 2$). Let this generalization of the ordinary relation of logical implication be symbolized by $\{p_1,...,p_n\} \Rightarrow \{q_1,...,q_n\}$. This relation has been defined by Segerberg 1971 to hold iff, for logical reasons, at least as many propositions from the set $\{q_1,...,q_n\}$ must be true as there are true propositions in the set $\{p_1,...,p_n\}$. Now, just like the logical implication between *p* and *q* guarantees that the probability of *q* is at least as great as the probability of *p*, so also the strict implication between $\{p_1,...,p_n\}$ and $\{q_1,...,q_n\}$ entails that the *sum* of the probabilities of the q_1 is at least as great as the corresponding sum $\sum_{i\leq n}$ Prob(a, p_i). Therefore, if at least one proposition from $\{p_1,\ldots,p_n\}$ is believed by a to be true (and hence has a probability > .5) and if all the other p_i are not believed by a to be false (and hence have a probability ≥ .5), so that in sum $\sum_{i\leq n}$ Prob(a, p_i) > n•1/2, it follows that also $\sum_{i\leq n}$ Prob(a,q_i) > n/2, and thus at least one of the q_i must be believed by a to be true:

(B7) $$\{p_1,\ldots,p_n\} \Rightarrow \{q_1,\ldots,q_n\} \vdash B(a,p_1)\wedge\neg B(a,\neg p_2)\wedge\ldots\wedge\neg B(a,\neg p_n) \rightarrow B(a,q_1)\vee\ldots\vee B(a,q_n).$$

II THE LOGIC OF KNOWLEDGE

2.1 In search of a "definition" of knowledge

Although a's firm belief that p is true is logically compatible with p's actually being false, it is a truism since Plato's early epistemological investigations in the *Theaitetos* that a cannot *know* that p unless p is in fact true. This first, "objective" condition of knowledge can be formalized as:

(K1) $K(a,p) \rightarrow p$.

Another "subjective" condition of knowledge has already been stated in the preceding section: **E3** says that person a cannot know that p unless she is convinced that p. This is a refinement of Plato's insight that knowledge requires belief – *viz.*, belief of the strongest form possible. Plato had discussed yet a third condition of knowledge which is somewhat harder to grasp. In order to constitute an item of *knowledge*, a's true belief must be "justified" or "well-founded". One might think of explicating this requirement by postulating the existence of certain propositions $q_1,\ldots,q_n$ which "justify" a's belief that p by logically entailing p. But which epistemological status should be accorded to these "justifying" propositions? If it were only required that the q_i must all be true and that a is convinced of their truth, then the "third" condition of knowledge would become redundant and each true belief would by itself be "justified".[4] On the other hand one cannot require that the q_i are *known* by a to be true, because then Plato's definition of knowledge as "justified" true belief would become circular.[5]

For the present purpose of investigating the *logic* of knowledge, two alternatives offer themselves. Either one treats 'knowledge' as a primitive, undefinable notion which is only partially characterized by the necessary conditions **K1** and **E3**. Or one takes the conjunction of these two conditions as already *sufficient* for a's knowing that p – an option favoured by Kutschera 1982 and, more recently, by Sartwell 1991.[6] Let us refer to this simple concept of knowledge as 'knowledge*' or 'K^*'. If one thus defines:

(Def. K*) $K^*(a,p) \leftrightarrow C(a,p) \wedge p,$

then the logic of knowledge* can easily be derived from the logic of conviction. This will be briefly carried out in section 2.2. The logic of a more demanding primitive notion of knowledge, $K(a,p)$, will afterwards be investigated in section 2.3.

2.2 The logic of knowledge as true, strong belief*

The first basic principle

(K*1) $K^*(a,p) \rightarrow p$

is an immediate corollary of Def. K*. Furthermore, the former conjunction-principle **C1** for strong belief directly entails a corresponding principle for knowledge*,

(K*2) $K^*(a,p) \wedge K^*(a,q) \rightarrow K^*(a,p \wedge q),$

and the rules of deductive closure of conviction, **C7** – **C9**, analogously entail the following rules for K^*

(K*3) $p \leftrightarrow q \vdash K^*(a,p) \leftrightarrow K^*(a,q)$

(K*4) $p \rightarrow q \vdash K^*(a,p) \rightarrow K^*(a,q)$

(K*5) $p \vdash K^*(a,p).$

It is easy to verify that **Def. K*** together with **C10** entails the iteration law

(K*6) $K^*(a,p) \rightarrow K^*(a,K^*(a,p)).$

As regards the "converse" iteration principle $\neg K^*(a,p) \rightarrow K^*(a,\neg K^*(a,p))$, two subcases must be distinguished. If *a*'s failure to know that *p* is due to *a*'s not sufficiently *believing* that *p*, then the conclusion $K^*(a,\neg K^*(a,p)$ is warranted; for in view of **C11** also

(E8) $\neg C(a,p) \rightarrow K^*(a,\neg C(a,p))$

becomes provable. If, however, $\neg K^*(a,p)$ results from a failure of the "objective" condition of knowledge*, i.e. if *p* is false although *a* is strongly convinced that *p*, then *a* will evidently *not* know that she does not know that p.[7] Hence the logic of K^* is at least as strong as the well-known modal system **S4** but definitely weaker than **S5**. A closer characterization will be given towards the end of the next section.

2.3 The logic of a more demanding concept of knowledge

The basic principle **K1** was already dealt with in section 2.1. Second, in analogy to **K*2**, also the more sophisticated concept of knowledge along Platonian lines should be taken to satisfy the principle of conjunctivity:

(K2) $K(a,p) \wedge K(a,q) \rightarrow K(a,p\wedge q)$.

For if one assumes that *a*'s single convictions that *p* and that *q* are justified, then *a*'s combined conviction that $(p\wedge q)$ would be justified as well. Third, the methodological position outlined in the introduction of this paper validate the following rules of deductive closure also for the more ambitious concept *K*:

(K3) $p \leftrightarrow q \vdash K(a,p) \leftrightarrow K(a,q)$
(K4) $p \rightarrow q \vdash K(a,p) \rightarrow K(a,q)$
(K5) $p \vdash K(a,p)$.

Since epistemic logic is here taken as a normative theory of *rational* (or "implicit") attitudes, these rules are just as acceptable as their doxastic counterparts **C7** – **C9** plus their corrolaries **K*3** – **K*5**.

The *K*-analogue of the iteration law **K*6**, i.e. so-called "KK-thesis", says that whenever a person *a* knows that *p*, *a* knows that she knows that *p*:

(K6) $K(a,p) \rightarrow K(a,K(a,p))$.

In the literature surveyed in Lenzen 1978, several "counter-examples" have been constructed to show that a person *a* may know something without knowing that she knows. For instance, assume that during an examination student *a* answers the question in which year Leibniz was born by replying 'In 1646'. The very fact that *a* managed to give the correct answer usually is taken as sufficient evidence to conclude that *a knew the correct answer.* But *a* may *not have known* at all *that she knew* the correct answer; in fact she may have thought she was just *guessing*.

Such examples typically play on the ambiguity of the English verb 'to know' which has the meaning both of the German 'wissen' and of 'kennen'. In the former case, 'to know' is followed by a that-clause and then expresses a *propositional attitude*; while in the latter case, 'to know' is part of a direct object construction ('to know the answer'; 'to know the way'; 'to know the city of London'; etc) and then expresses no such attitude. Therefore the above "counter-example" fails to refute **K6** since *a*'s "knowing" the correct answer, i.e. her knowing the year in which Leibniz was born, does not represent a propositional attitude as would be required by **K6**. According to the premises of the story, *a* did *not* know *that* Leibniz was born in 1646 because she was not at all certain of the date. If someone really *knows* that Leibniz was born in 1646, i.e., by **E3**, if *a* is *a fortiori* convinced that Leibniz was born in 1646, then *a* can never believe that he does not know that Leibniz was born in 1646.

The argument contained in the preceding passage contains an application of another important principle which establishes an epistemic logical connection between all the three basic notions of knowledge, belief, and conviction. In its general form, it would have to be put as follows: Whenever person a is *convinced* that p, she will *believe* that she *knows* that p:

(E9) $C(a,p) \rightarrow B(a,K(a,p)).$

In view of certain iteration laws discussed earlier in this paper, **E9** can be strengthened into the statement that when a is convinced that p, she must be convinced that she knows that p.

(E10) $C(a,p) \rightarrow C(a,K(a,p)).$

Incidentally, the implications **E9** and **E10** might further be strengthened into equivalencies, and because of **C10** also the following law becomes provable:

(E11) $C(a,C(a,p)) \leftrightarrow C(a,K(a,p)).$

E11 shows that knowledge and conviction are *subjectively indiscriminable* in the sense that person a cannot tell apart whether she is "only" convinced that p or whether she really knows that p. This observation does not remove, however, the *objective* difference between a's being convinced that p and a's knowing that p; only the latter but not the former attitude entails the truth of p. Therefore it is always ("objectively") possible that a is convinced of something which as a matter of fact is not true; but person a herself can never think this to be possible.[8]

Because of the objective possibility of $C(a,p) \wedge \neg p$, the K-analogue of the doxastic iteration principle **C11**, i.e. $\neg K(a,p) \rightarrow K(a,\neg K(a,p))$, fails to hold. From the assumption that person a *does not* know that p one cannot infer that she *knows* that she does not know that p. For if a *mistakenly* believes that she knows that p, i.e. if $C(a,p) \wedge \neg p$, one has $\neg K(a,p)$ (because of **K1**) and yet a does not know of her mistake, because in view of **E9** a believes that she *does* know that p; hence she is far from believing (or even knowing) that she does *not* know that p.

Summing up, then, no matter whether 'knowledge' is taken in the simply sense of K^* or in the more demanding sense of K, the logic of knowledge is (isomorphic to a modal calculus) at least as strong as **S4** but weaker than **S5**. Now there is a very large – indeed, as shown in Fine 1974, an *infinite* – variety of modal systems "between" **S4** and **S5**. E.g., so-called system **S4.2** is characterized by an axiom which (when the alethic operator $\Box$ is interpreted as 'knowledge') takes the form:

(K7) $\neg K(a,\neg K(a,p)) \rightarrow K(a,\neg K(a,\neg K(a,p))).$

Another calculus **S4.4** is axiomatized by (the $\Box$-counterpart of):

(K8) $p \wedge \neg K(a,\neg K(a,p)) \rightarrow K(a,p)).$

However, the meaning of these principles is not at all evident because common sense says little or nothing about the epistemic counterpart of the alethic modality $\Diamond\Box p$, i.e. $\neg K(a,\neg K(a,p))$. Fortunately, the laws of epistemic logic developed earlier in this paper give us a clue how to understand this complex term. It is easy to prove that person a is convinced that p iff she does not know that she does not know that p:

(E12) $\neg K(a,\neg K(a,p)) \leftrightarrow C(a,p)$.

One the one hand, $C(a,p)$ entails $C(a,K(a,p))$ (by **E10**) and a fortiori $\neg C(a,\neg K(a,p))$ (by **C2**) which in turn entails $\neg K(a,\neg K(a,p))$ by **E3**; on the other hand $\neg C(a,p)$ implies $K(a,\neg C(a,p))$ (by **E2**) and hence also $K(a,\neg K(a,p))$ by the rule **K4** in conjunction with **E3**.

In view of **E12**, then, the **S4.2**-like principle **K7** amounts to saying that when person a is convinced that p, she knows that she is convinced that p – this is exactly the content of our earlier principle **E1**. Similarly, **S4.4**-like principle **K8** states that when p is true and when a is convinced that p, then a already knows that p.

As the reader may easily verify, on the basis of **Def. K*** both

(K*7) $\neg K^*(a,\neg K^*(a,p)) \rightarrow K^*(a,\neg K^*(a,\neg K^*(a,p)))$

and

(K*8) $p \wedge \neg K^*(a,\neg K^*(a,p)) \rightarrow K^*(a,p))$

become theorems of the logic of strong belief. Hence the logic of K^* actually is (isomorphic to) **S4.4**. As regards the logical structure of the more demanding concept of knowledge, K, all that can be asserted here is that it is (isomorphic to an alethic modal system) *at least as strong as* **S4.2** *but weaker than* **S4.4**.[9]

To conclude our discussion of the propositional logic of knowledge, let it just be pointed out that a possible-worlds semantics for K can be given along the following lines:

(POSS-K) $V(i,K(a,p))=\mathbf{t} \leftrightarrow \forall j(iRj \rightarrow V(j,p)=\mathbf{t})$.

Here 'R' denotes an accessibility relation between worlds which obtains iff world j is compatible with (or "possible" according to) all that a knows in world i.

III "QUANTIFYING IN" AND OTHER PROBLEMS IN FIRST ORDER EPISTEMIC LOGIC

During the late 50ies and 60ies a large controversy concerning the very possibility of quantified modal logic took place among such prominent philosophers as, e.g., W.V. Quine, J. Hintikka, and D. Kaplan. In what follows, only the most fundamental issues will be touched while the historical development of the discussion must remain out of consideration.[10] The main source of the problem of "quantifying in" is the failure of substitutivity of co-referential singular terms within modal contexts:

3.1 Referential opacity

According to a by now familiar terminology, a context ϕ is said to be *referentially transparent* with respect to a term t iff t may be replaced in ϕ, *salva veritate*, by any coreferential term t':

(SUB-ϕ) $\forall tt'(t= t' \rightarrow (\phi(t) \leftrightarrow \phi(t')))$.

If **SUB-ϕ** does not hold, ϕ is said to be *referentially opaque*. Now, epistemic operators such as $B(a,p)$, $C(a,p)$, or $K(a,p)$ evidently generate referentially opaque contexts. For example, in Sophocles' famous drama, although Iocaste was (identical with) Oedipus' mother – $i = \iota xM(x,o)$ – the fact that Oedipus knew he was in love with Iocaste did not at all entail that Oedipus knew he was in love with his mother, i.e., making use of some straightforward abbreviations, one has $K(o,L(o,i))$ but $\neg K(o,L(o,\iota xM(x,o)))$. In general, the inference from $K(a,\phi(t))$ to $K(a,\phi(t'))$ seems warranted only if, instead of the mere identity $t=t'$, one has the stronger premise that this identity is *known* by subject a to hold:

(SUB1) $\forall tt'(K(a,t=t') \rightarrow (K(a,\phi(t)) \leftrightarrow K(a,\phi(t'))))$.

In the case of the other epistemic operators $C(a,p)$ and $B(a,p)$, one obtains analogously:

(SUB2) $\forall tt'(C(a,t=t') \rightarrow (C(a,\phi(t)) \leftrightarrow C(a,\phi(t'))))$
(SUB3) $\forall tt'(C(a,t=t') \rightarrow (B(a,\phi(t)) \leftrightarrow B(a,\phi(t'))))$.[11]

Now, the referential opacity of epistemic contexts appears to render any quantification into these contexts dubious. Consider, e.g., the elementary law of existential generalization:

(EX1) $\phi(t) \rightarrow \exists x\phi(x)$,

and let ϕ be some epistemic statement such as, e.g., 'Oedipus believes that his mother is dead', $B(o,D(\iota xM(x,o)))$. Because of his ignorance concerning the identity of Iocaste and his mother, $\neg K(o,i=\iota xM(x,o))$, Oedipus certainly does *not* believe that *Iocaste* is dead: $\neg B(o,D(i))$. But then, one might argue, the premise $B(o,D(\iota xM(x,o)))$ does not entail the existential proposition $\exists xB(o,D(x))$ asserting that there exists someone, x, such that Oedipus believes x to be dead. For, according to Quine, $\exists x\phi(x)$ is *true* only if the open sentence $\phi(x)$ expresses a property which is true *of* some individual x, no matter which way we happen to refer to this individual. But it is evidently *not* true *of* Oedipus' mother, i.e. *of* Iocaste, that Oedipus would believe *her* to be dead since Oedipus does not believe *that* Iocaste is dead. Thus the inference from $B(o,D(\iota xM(x,o)))$ to $\exists xB(o,D(x))$ should not be considered as logically valid (although, with respect to the particular predicate $D(x)$ chosen in our example, the truth of the conclusion $\exists xB(o,D(x))$ would most likely seem to be warranted by Oedipus' other beliefs).

Closely related to this *logical* objection is a *linguistic* objection of Quine's pertaining to the meaningfulness of quantified epistemic expressions in general. The content of someone's epistemic attitude usually is a *state of affairs* which can be expressed by some proposition p. Accordingly epistemic operators such as 'a believes that' or 'a knows that' (or, for that matter, also other modal operators such as 'it is necessarily true that') have to be followed by a *full*, "closed" sentence p, e.g. $p = F(t)$. The propositional operators "seal off" the subsequent that-clause in a way that the replacement of the singular term t by a variable x as, e.g., in 'a believes that $F(x)$' produces an syntactically ill-formed expression which, in contrast to the open sentence $F(x)$, cannot be taken to express a real *property*. For, a linguistic expression $\phi(x)$ denotes a property only if, for every individual x, ϕ either applies to – or fails to apply to – x regardless of the way in which we happen to refer to x. As was argued in connection with principles **SUB1** – **SUB3**, however, in the case of epistemic expressions this condition is not fulfilled. Anyway, according to Quine, a quantified "sentence" like $\exists xB(a,F(x))$ – or its "ordinary language"-counterpart 'There exists some individual x such that a believes that x [or *it*] is F' – is devoid of a sound interpretation and hence, strictly speaking, *meaningless*. In the next section it will be shown how these objections can be overcome once an important distinction between two different kinds of epistemic expressions is taken into account:

3.2 De dicto and de re

Epistemic phrases such as 'a believes t to be F' or 'a knows t to be F' admit of two quite distinct interpretations: first, the more common *de dicto* reading where the content of a's belief or knowledge is the "dictum", i.e. the sentence or proposition, that t is F; second, a somewhat less common *de re* interpretation according to which the complex property of being believed or known by a to be F is attributed to the individual (or "res") t. While *de dicto* sentences can be represented by means of our standard operators in the usual manner:

$B(a,F(t))$ – a believes that t has the property F
$C(a,F(t))$ – a is convinced that t has the property F
$K(a,F(t))$ – a knows that t has the property F,

de re sentences appear to require a new formalism. Let $B(a,F)$, $C(a,F)$, and $K(a,F)$ abbreviate, for any epistemic subject a and for any "normal" predicate F, the complex properties of being (weakly or strongly) believed or known by a to be F. Then *de re* sentences will take the following symbolic form:

$B(a,F)(t)$ – t is (weakly) believed by a to be F
$C(a,F)(t)$ – t is strongly believed by a to be F[12]
$K(a,F)(t)$ – t is known by a to be F.

This type of formal representation – and the characterization of the epistemic predicates $B(a,F)$, $C(a,F)$, and $K(a,F)$ as expressing complex *properties* – is meant to suggest that *de re* sentences are referentially *transparent*. If some "res" t has the property of being believed or known by a to be F, then it does not matter in which way we refer to that individual; i.e. if t' is identical with t, then t' also has this property. Thus we may assume that the following principles hold:

(SUB4) $t{=}t' \rightarrow (B(a,F)(t) \leftrightarrow B(a,F)(t'))$
(SUB5) $t{=}t' \rightarrow (C(a,F)(t) \leftrightarrow C(a,F)(t'))$
(SUB6) $t{=}t' \rightarrow (K(a,F)(t) \leftrightarrow K(a,F)(t'))$.

Furthermore, given the intended interpretation of our *de re* sentences, they evidently admit of existential generalization. Clearly, if t has the property of being believed or known by a to be F, then there exists some individual x which has this property:

(EX2) $B(a,F)(t) \rightarrow \exists x B(a,F)(x)$
(EX3) $C(a,F)(t) \rightarrow \exists x C(a,F)(x)$
(EX4) $K(a,F)(t) \rightarrow \exists x K(a,F)(x)$.

Similarly, if every "res" x has the property of being believed or known by a to be F, then t must have this property, too:

(UN1) $\forall x B(a,F)(x) \rightarrow B(a,F)(t)$
(UN2) $\forall x C(a,F)(x) \rightarrow C(a,F)(t)$
(UN3) $\forall x K(a,F)(x) \rightarrow K(a,F)(t)$.

Note that all quantified expressions in **EX2- EX4** and **UN1 – UN3** are *de re* constructions which – unlike their *de dicto* counterparts discussed in the previous section – do not fall under Quine's verdict of being ungrammatical.

Next it remains to be investigated which logical relations exist between epistemic propositions *de dicto* and *de re*. For convenience we will set aside the attitudes of weak and strong belief and concentrate instead on knowledge. Under which circumstances will it be allowed to "export" the singular term t occurring within the *de dicto* construction 'a knows that t is F' so as to infer that t has the property of being known by a to be F, and *vice versa*? To answer these questions one first has to state precise truth conditions for knowledge-sentences *de dicto* and *de re*. Unfortunately, there is little agreement concerning the general framework within which such a semantics should best be developed. In particular it is still somewhat controversial in which sense one and the same individual t can be assumed to exist in (or to be identifiable across) different possible worlds. E.g., according to the "counterpart-theory" developed in Lewis 1968, the domains of two such worlds should always be taken to be set-theoretically disjoint: If t exists in a certain world i, then not t himself but at best one of his "counterparts" t^* can exist in another world $j{\neq}i$. In what follows, however, we will rather adopt an approach suggested by Kripke 1972 according to which a possible-worlds model $<U,I,R,V>$

should always be based on a common universe of discourse, U, i.e., for every world $i\varepsilon I$, the domain of i is one and the same set U.[13]

Within this Kripkean framework our general condition **POSS-K** mentioned in section 2.3 immediately combines with the usual interpretation of the first-order formula $p = F(t)$ to yield the following truth condition for *de dicto* knowledge statements: $K(a,F(t))$ is true under the interpretation V in a world i iff in every world j which is "accessible" from i (i.e. which is possible according to all that a knows in i) V makes $F(t)$ true in j; and the latter condition, $V(j,F(t))=\mathbf{t}$, means more specifically that the object assigned by the interpretation V to t in world j, $V(j,t)$, belongs to the extension of the predicate F in world j, $V(j,F)$:

(POSS-K-DICTO) $V(i,K(a,F(t))) = \mathbf{t} \leftrightarrow \forall j(iRj \rightarrow V(j,t)\varepsilon V(j,F))$.

Since a valuation function V can in general assign different objects x, x', x'', ... to a singular term t in different worlds i, i', i'', ..., the above truth condition amounts to the rather weak requirement that in every world j the object denoted by t in j has the property F in j. In contrast, the truth of the *de re* statement 't is known by a to be F' shall be taken to require more strictly that in each relevant world j (such that iRj) *one and the same object* x is denoted by t in j and this object x has the property F in j:

(POSS-K-RE) $V(i,K(a,F)(t)) = \mathbf{t} \leftrightarrow \exists x(\forall j(iRj \rightarrow V(j,t) = x \,\&\, x\varepsilon V(j,F)))$.

According to this analysis every *de re* knowledge entails a corresponding *de dicto* knowledge:

(RE-DICTO) $K(a,F)(t) \rightarrow K(a,F(t))$,

while the converse implication does not generally hold. In the next section we will discuss the extra premises that must be satisfied in order to infer a *de re* statement $K(a,F)(t)$ (or the existential corollary $\exists xK(a,F)(x)$) from the *de dicto* statement $K(a,F(t))$. To conclude this section let it just be mentioned that in the case of *belief* things are yet a little bit more complicated. In analogy to $K(a,F)(t)$, the truth of $C(a,F)(t)$ also does require that in each relevant world j one and the same object x is denoted by t in j and this object x has the property F in j. This furthermore warrants that for some singular term t' such that (in the actual world i) $t'=t$, $C(a,F(t'))$ will be true (in i). However, subject a may perhaps not know of this identity and may therefore fail to believe that t himself has property F: Remember Quine's famous 1956 scenario of Ralph's beliefs concerning t="Bernard J. Ortcutt" and t'="a certain man in a brown hat"!

3.3 Rigid designators and 'Knowing who t is'

When it comes to designing a formal calculus of first order epistemic logic, it seems very convenient to interpret (at least a subset of) the singular terms as *rigid designators* where t designates an object x *rigidly* iff $V(i,t)=x$ for every $i\varepsilon I$, i.e. iff t refers to one and the same individual x in each possible world. Kripke 1972 argued that the proper names of our ordinary discourse actually are used as rigid designators while other singular terms, in particular definite descriptions, do not always designate their referents in a rigid way. Without entering into the philosophical discussion of this issue here, let us simply postulate that the *names* b, b', b'', ... of our formal language are interpreted (by the respective valuation function V) as rigid designators:

(RIGID) $V(i,b)=V(i',b)$ for all $i,i'\varepsilon I$,

while the denotation of a definite description $\iota x\phi x$ in world j, $V(j,\iota x\phi x)$, is logically determined by $V(j,\phi)$ and may thus vary from world to world. It then easily follows that the crucial inference

(DICTO-RE) $K(a,F(b)) \rightarrow K(a,F)(b)$

and hence – in view of **EX4** – also "quantifying in"

(QUANT-IN1) $K(a,F(b)) \rightarrow \exists xK(a,F)(x)$.

is valid for any rigid designator b.

If, however, t is a non-rigid singular term, then the corresponding inferences require an extra premise to guarantee that t refers to one and the same individual in at least all *relevant* worlds, i.e. in every $j\varepsilon I$ such that iRj. According to Hintikka 1962, this premise should be paraphrased as 'a knows who t is'. Unfortunately, the truth conditions for this informal requirement are rather vague. Consider, e.g., t = 'the 1998 President of the United States'. What kinds of facts must a subject a know in order to know who the 1998 US-President is? Does it suffice that a just knows his *name*, or will a also have to know certain facts about the *person* Bill Clinton; must a furthermore be able to *identify* Bill Clinton under "normal" circumstances; etc.? In view of these indeterminacies one better forgets the informal reading 'a knows who t is' and considers instead its formal counterpart which Hintikka represents as $\exists xK(a,x=t)$. Again, however, this condition is not without problems. As was rightly stressed by Quine, any quantified "sentence" of the type $\exists xK(a,\Phi(x))$ involving an epistemic *de dicto* operator would have to be "translated" as 'There exists some individual x such that a knows that x satisfies condition Φ'. But any such locution is grammatically ill-formed. The only meaningful interpretation of quantified epistemic sentences consists of the *de re* construction 'There exists some individual x such that x is known by a to satisfy Φ' – $\exists xK(a,\Phi)(x)$. Hence the crucial prerequisite for "quantifying in" the singular *de dicto* statement $K(a,F(t))$ has to be formalized more exactly by the condition $\exists xK(a,=t)(x)$ which says that there exists some individual x such that x is known by a to be (identical to) t:

(QUANT-IN2) $K(a,F(t)) \wedge \exists xK(a,=t)(x) \rightarrow \exists xK(a,F)(x)$.

Note, incidentally, that the only individual which may ever satisfy the condition $K(a,=t)$ is, of course, t itself. For, in view of the truth condition of knowledge, if some x is *known* by a to be (identical to) t, then *a fortiori* x has to be (identical to) t. Thus the crucial premise in **QUANT-IN2** might as well be formulated by requiring that t itself has the property of being known by a to be (identical to) t! Unlike the *de dicto* formula $K(a,t=t)$, the somewhat queer-looking *de re* sentence $K(a,=t)(t)$ is not trivially satisfied by arbitrary subjects a.[14] In view of the semantic principle **POSS-K-RE** stated in section 3.2 above, $V(i,K(a,=t)(t))=\mathbf{t}$ requires that there exists some individual x such that in every relevant world j $V(j,t)=x$ and $x \varepsilon V(j,=t)=\{V(j,t)\}$, i.e. in every world j such that iRj the singular term t has to be interpreted by V as designating one and the same individual x (*viz*, "the" t in the real word i).

Thus the equivalence $K(a,=t)(t) \leftrightarrow \exists x K(a,=t)(x)$, or also $K(a,=t)(t) \leftrightarrow \exists x(x=t \wedge K(a,=t)(x))$, turns out to be valid. More generally, just like in ordinary first order logic with identity any singular statement $\Psi(t)$ is provably equivalent to $\exists x(x=t \wedge \Psi(x))$, so also every singular *epistemic de re* sentence $K(a,\phi)(t)$ turns out to be equivalent to the quantified formula $\exists x(x=t \wedge K(a,\phi)(x))$:

(RE-QUAN) $$K(a,\phi)(t) \leftrightarrow \exists x(x=t \wedge K(a,\phi)(x)).$$

This equivalence provides the basis for a possible simplification of our formalism. In order to distinguish *de re* from *de dicto* sentences, the ordinary propositional operator $K(a,p)$ had been supplemented in section 3.2 by a predicate-forming operator $K(a,\phi)$ which, for any predicate ϕ, yields the epistemic predicate 'is known by a to be ϕ'. Within the realm of *quantified* epistemic sentences, however, the *de dicto/de re* distinction is superfluous. As was stressed above, there is no meaningful way to formulate quantified *de dicto* sentences; every quantified epistemic sentence always has to be understood *de re*! Therefore we might for convenience retain the ordinary *de dicto* operator to formally represent quantified (de re) sentences according to the subsequent

(CONVENTION) $$\exists x K(a,\phi(x)) \leftrightarrow \exists x K(a,\phi)(x)$$ $$\forall x K(a,\phi(x)) \leftrightarrow \forall x K(a,\phi)(x).$$

In particular, the condition $\exists x(x=t \wedge K(a,\phi)(x))$ as it occurs in **RE-QUAN** might be rewritten as $\exists x(x=t \wedge K(a,\phi(x))$ and we would thus obtain the following formal representation of the singular *de re* knowledge sentence 't is known by a to be F': $\exists x(x=t \wedge K(a,\phi(x))$ (and similarly for the other epistemic operator of strong and weak belief). In sum, then, we would obtain a first order calculus with only one type of epistemic operator $K(a,\phi)$, $C(a,\phi)$, and $B(a,\phi)$. These have to be interpreted *de dicto* whenever ϕ is a "closed" sentence or proposition p, but they have to be interpreted *de re* when $\phi(x)$ is a "open" sentence with the variable x being bound by a quantifier $\exists x$ or $\forall x$ outside the epistemic operator.

To conclude, let it be mentioned that the general semantic approach advocated here – i.e. the choice of possible-worlds models $<U,I,R,V>$ with a common universe of discourse for each possible world i – validates the following epistemic counterparts of the so-called "Barcan formula" and "converse Barcan formula" (of alethic modal logic):

(UN4) $B(a,\forall xF(x)) \to \forall xB(a,F)(x)$

(UN5) $C(a,\forall xF(x)) \to \forall xC(a,F)(x)$

(UN6) $K(a,\forall xF(x)) \to \forall xK(a,F)(x)$

(UN7) $\forall xC(a,F)(x) \to C(a,\forall xF(x))$

(UN8) $\forall xK(a,F)(x) \to K(a,\forall xF(x))$.

The invalidity of the B-counterpart of **UN7** is due to the fact that the operator of "weak belief" does not satisfy a conjunction principle analogous to **C1** or **K2**. In the simplified calculus based on the above **CONVENTION**, the *de re* components of the laws **UN4** – **UN6** might, of course, be symbolized by means of the apparently *de re* formulae $\forall xB(a,F(x))$, $\forall xC(a,F(x))$, and $\forall xK(a,F(x))$, respectively. Yet this convenient formalization should not seduce anyone to overlook the important difference between these two kinds of propositions which corresponds to the Medieval distinction between propositions "in sensu composito", e.g., $K(a,\forall xF(x))$, and propositions "in sensu diviso", e.g., $\forall xK(a,F(x))$.

Wolfgang Lenzen
University of Osnabrück

NOTES

[1] For a recent defense of this view cf., e.g., Meyer 1998.

[2] Clearly, since $C(a,p)\vee\neg C(a,p)$ holds tautologically, **C10** and **C11** entail that $\{C(a,C(a,p)) \vee C(a,\neg C(a,p))\}$ is epistemic-logically true. So either way there exists a q such that $C(a,q)$.

[3] Cf. Lenzen 1995 for a closer discussion of the differences between (and the dependency of) the semantics and the pragmatics of epistemic utterances.

[4] Clearly, if $C(a,p)\wedge p$, then there exist some $q_1,...,q_n$ such that the q_i are true and $C(a,q_i)$ and $\{q_1, ...,q_n\}$ logically entail p, *viz.*, $q_1=...=q_n=p$!

[5] For a closer discussion of this problem the reader is referred to part D of this Handbook, esp. to the contribution on the "Analysis of Knowledge".

[6] Cf. for a closer discussion Beckermann 1997.

[7] Otherwise the assumption $C(a,p)\wedge\neg p$ would entail a contradiction, i.e. $C(a,p) \to p$ would become a theorem of the logic of strong belief.

[8] This observation not only represents the key for the resolution of several epistemic "paradoxes" but also helps to clarify the problems that prominent philosophers encountered during their epistemological reflections on the nature of knowledge and belief. For a more detailed discussion of the "surprise examination paradox" cf. Lenzen 1976. Lenzen 1980b offers an analysis of Wittgenstein's sometimes confused discussion of "Moore's paradox" in his late booklet 1962.

[9] Cf. Lenzen 1979 for a closer discussion of further candidates for the logic of knowledge.

[10] Cf. Quine 1956, Hintikka 1961, and Kaplan 1969; Hintikka 1975 tries to summarize the controversy and he also mentions various other writers who had contributed to the discussion of "quantifying in".

[11] In view of the non-conjunctivity of the operator $B(a,p)$, the premise $B(a,t{=}t')$ is too weak to warrant the inference from $B(a,\phi(t))$ to $B(a,\phi(t'))$. One here needs a stronger premise such as $C(a,t{=}t')$ or $K(a,t{=}t')$; cf. principle **E4** stated in section 1.2 above.

[12] Interestingly, neither in English nor in German does there exist an idiomatic locution expressing such a strong *de re* belief in terms of 'being convinced' or 'being certain'.

[13] Let it be noted in passing that this does not entail that every individual "existing" in the actual world also "exists" in every other possible world (and hence "exists necessarily"). Real existence can be regarded as an empirical, contingent property which does not automatically apply to every individual in the domain of world *i*! Another position concerning the issue of "trans-world-identity"" has been defended by Hintikka (1969, 1970b).

[14] The reason being that the first occurrence of 't' as part of the complex epistemic predicate $K(a,=t)$ is referentially *opaque*, i.e. $t=t'$ does not entail that x has property $K(a,=t)$ iff x has property $K(a,=t')$. The second occurrence of 't' in $K(a,=t)(t)$, however, is referentially transparent, i.e. $t=t'$ and $K(a,=t)(t)$ entail that $K(a,=t)(t')$.

REFERENCES

Beckermann, A.: 1997, 'Wissen und wahre Meinung', in W. Lenzen (ed.), *Das weite Spektrum der Analytischen Philosophie*, de Gruyter, Berlin, pp. 24-43.

Boh, I.: 1986, 'Elements of epistemic logic in later Middle Ages', in *L'Homme et son Univers au Moyen Age – Philosophes Médiévaux* **27**, 530-543.

Carnap, R.: 1947, *Meaning and Necessity*, University of Chicago Press.

Chisholm, R.M.: 1963, 'The logic of knowing', in *Journal of Philosophy* **60**, 773-795.

Fagin, R., J. Y. Halpern, Y. Moses and M. Y. Vardi: 1994, *Reasoning about Knowledge*, MIT Press, Cambridge, Mass.

Fine, K.: 1974, 'An ascending chain of **S4**-logics', in *Theoria* **40**, 110-116.

Finetti, B. de: 1964, 'Foresight: Its logical laws, its subjective sources', in H.E. Kyburg and H.E. Smokler (eds.), *Studies in Subjective Probability*, R.E. Krieger, Huntington, N.Y., pp. 53-118.

Gärdenfors, P.: 1988, *Knowledge in Flux: Modelling the Dynamics of Epistemic States*, MIT Press, Cambridge, Mass.

Hintikka, J.: 1961, 'Modality and Quantification', in *Theoria* **27**, 119-128; revised version reprinted in J. Hintikka, *Models for Modalities*, D. Reidel, Dordrecht, pp. 57-70.

Hintikka, J.: 1962, *Knowledge and Belief*, Cornell University Press, Ithaca, N.Y.

Hintikka, J.: 1969, 'Semantics for propositional attitudes', in J. W. Davis et al. (eds.), *Philosophical Logic*, D. Reidel, Dordrecht, pp. 21-45; reprinted in J. Hintikka, *Models for Modalities*, D. Reidel, Dordrecht, pp. 87-111.

Hintikka, J.: 1970a, 'Surface information and depth information', in J. Hintikka & P. Suppes (eds.), *Information and Inference*, D. Reidel, Dordrecht, pp. 263-297.

Hintikka, J.: 1970b, 'The Semantics of Modal Notions and the Indeterminacy of Ontology', in *Synthese* **21**, 408-424.

Hintikka, J.: 1975, 'Quine on Quantifying in: A Dialogue', in J. Hintikka, *The Intentions of Intentionality and other New Models for Modalities*, D. Reidel, Dordrecht, pp. 102-136.

Hocutt, M.O.: 1972, 'Is epistemic logic possible?', in *Notre Dame Journal of Formal Logic* **13**, 433-453.

Kaplan, D.: 1969, 'Quantifying In', in *Synthese* **19**, 178-214.

Kripke, S.: 1959, 'A completeness theorem in modal logic', in *Journal of Symbolic Logic* **24**, 1-14.

Kripke, S.: 1972, 'Naming and Necessity', in D. Davidson & G. Harman (eds.), *Semantics of Natural Language*, D. Reidel, Dordrecht, pp. 253-355 + 763-769.

Kutschera, F.v.: 1976, *Einführung in die intensionale Semantik*, de Gruyter, Berlin.

Kutschera, F.v.: 1982, *Grundfragen der Erkenntnistheorie*, de Gruyter, Berlin.

Lemmon, E.J.: 1977, *An Introduction to Modal Logic*, Blackwell, Oxford.

Lenzen, W.: 1976, 'Die Paradoxie der überraschenden Übung: Logische, epistemologische und pragmatische Aspekte', in *Logique et Analyse* **19**, 267-284.

Lenzen, W.: 1978, *Recent Work in Epistemic Logic*, North Holland Publ. Company, Amsterdam.

Lenzen, W.: 1979, 'Epistemologische Betrachtungen zu [**S4**, **S5**]', in *Erkenntnis* **14**, 33-56.

Lenzen, W.: 1980a, *Glauben, Wissen und Wahrscheinlichkeit*, Springer Verlag, Vienna.

Lenzen, W.: 1980b, 'Wittgensteins Zweifel über Wissen und Gewißheit', in *Grazer Philosophische Studien* **10**, 43-52.

Lenzen, W.: 1995, 'On the semantics and pragmatics of epistemic attitudes', in A. Laux and H. Wansing (eds.), *Knowledge and Belief in Philosophy and Artificial Intelligence*, Akademie-Verlag, Berlin, pp. 181-197.

Lewis, D.: 1968, 'Counterpart Theory and Quantified Modal Logic', in *Journal of Philosophy* **65**, 113-126.

Meyer, U.: 1998, *Glaube und Notwendigkeit*, Schöningh Verlag, Paderborn.

Quine, W.V.O.: 1956, 'Quantifiers and Propositional Attitudes', in *Journal of Philosophy* **53**, 177-187.

Sartwell, C.: 1991, 'Knowledge is merely true belief', in *American Philosophical Quarterly* **28**, 157-165.

Segerberg, K.: 1971, 'Qualitative probability in a modal setting', in J. Fenstad (ed.), *Proceedings of the 2nd Scandinavian Logic Symposium*, North Holland Publ. Company, Amsterdam, pp. 341-352.

Wittgenstein, L.: 1962, *Über Gewißheit*, Suhrkamp, Frankfurt.

DAVID NOVITZ

KNOWLEDGE AND ART

Raised eyebrows are one common response to the claim that art is a valuable source of knowledge and understanding about the world. Some philosophers prefer to see the idea as a fond and self-deluded notion; rather like believing that one's lap-dog is wise and knowledgeable, and that his cuddly warmth, sympathetic licks and doggy breath indicate a fund of wisdom and compassionate understanding.

Philosophical scepticism about art as a source of knowledge has well-established roots in the history of Western philosophy. Its earlier manifestations are to be found in Plato and, to a lesser extent, Aristotle; it pervades the philosophy of the Enlightenment, is a fulsome part of the positivism of the nineteenth and twentieth centuries, and is to be found alive and well – although under a different guise – in the work of postmodernists everywhere.

I A BRIEF LOOK BACKWARDS

Plato was adamant that the dramatic arts, poetry, and the painting of his day were not, and could not be, the vehicles of genuine knowledge. As is well known, his criticisms of the instructive value of these art forms were expressed through the voice of Socrates, who (especially in the *Republic*) maintained that since art merely imitates appearances and plays on the emotions it is bound to mislead the intellect – not just by imparting false views of the gods but also, and more particularly, by corrupting our understanding.

And so, in Book 10 of the *Republic*, Socrates launched his famous attack on all of the representative or imitative arts insofar as they are valued as representations of reality. Because the poet and the dramatist only imitate the appearances of reality, they may be technically brilliant imitators without knowing anything at all about that which is imitated. Hence, the fact that Homer gives a convincing portrayal of human nature and the world of the gods should not be taken to show that he knows very much about either. Like painting, literary imitation is considered by Plato to be little more than a form of amusement that tells us very little about how things really are, and does not provide us with genuine insight or knowledge. For this reason, the amusement that it offers is likely to be harmful since it encourages people to adopt false opinions and demonstrably misleading ideas.

Artistic skills (*techné*; *poeisis*) are regarded by both Plato and Aristotle as the skills that circumscribe craftsmanship – skills of *making* rather than of the theoretical or intellectual skills (*theoria*) that are an integral part of coming to know. Hence one finds in Aristotle as well as in Plato the idea that the dramatic arts are not primarily concerned with the generation of knowledge and understanding, but with crafting or making. Even so, we find much more room in Aristotle for the view that

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 985-1012.
© 2004 *Kluwer Academic Publishers.*

we may achieve genuine understanding through art than we do in Plato, for he is of the view that the pity and fear that we experience on viewing a good tragedy enable us to recognize that a person with certain flaws of character cannot live a fulfilled or a flourishing life; indeed that such a person must be unhappy and must cause unhappiness in the lives of others.

Hence on Aristotle's view, a large part of what tragedy does is bring us to dwell on and consider the sorts of character deficiencies and actions that prevent a person from flourishing during the course of a life-time, and so from achieving happiness or *eudaimonia*. Hence, it is by skillfully showing us how human beings of a certain character live, and by demonstrating the outcome of living in that way, that tragedy affords insights into how life may best be lived. Hence, a generous construal of Aristotle in the *Poetics* and the *Politics* would suggest that tragedy has both a cognitive and a motivational dimension; it teaches us something about human character and the emotions, and it teaches us how we ought and ought not to behave. For the pity and fear that the play arouses in its viewers indirectly motivate us to avoid certain forms of behaviour and ways of living, and to embrace others (Nussbaum 1986, 378-394) – although whether this counts as practical *knowledge* rather than opinion is never entirely clear in Aristotle.

According to Socrates, we have seen, the poet does not rely on a creative or originative imagination in order to produce poetry. Instead, he contends that the poet's creative powers come from without, that they are the product of divine inspiration. In the *Phaedrus* we are told that "if any man come to the gates of poetry without the madness of the Muses, persuaded that skill alone will make him a good poet, then shall he and his works of sanity be brought to nought by the poetry of madness" (Plato 1952, 245a). Poetry, like all art, is not always rational and may even approach the point of insanity. The vexing thing for Socrates, however, is that poetry is not entirely without value, and it is for this reason that the artist's madness must be laid at the feet of the gods. It is they who are responsible for the artist's madness.

Similar views of artistic creativity seems to have prevailed right up until the time of the Renaissance. Christians of the Middle Ages, influenced as they were by Plato, preferred to regard works of art as theophanies – that is, as manifestations of a transcendent deity who both inspired the work and whose glory was expressed by means of it. St Augustine, however, seems to have been an exception, for even though a Platonist, he refused to explain works of art as simple imitations; still less as a product of divine inspiration. In a letter strongly reminiscent of John Locke and David Hume, he writes:

> it is possible for the mind, by taking away ... some things which the senses have brought within its knowledge, and by adding some things, to produce in the exercise of the imagination that which, as a whole, was never within the observation of any of the senses. (Augustine 1956, Vol.1, 255-256).

Still, it is only really at the time of the Renaissance, with its increased emphasis on the powers and worth of the individual, that artistic creation comes to be treated less as a product of divine inspiration and more as the result of the imaginative powers of the individual. Philosophers waited until the seventeenth century before writing about the imagination, and their comments were far from flattering. Cartesians were dismissive of the imagination and flatly denied that it could have any role at all to play in the acquisition of genuine knowledge. René Descartes writes of "the misleading judgement that proceeds from the blundering constructions of imagination", (Descartes 1931, 7.) In the same vein, Nicholas Malebranche devoted the whole of the second book of his *Recherche de la Verite* to establishing that the imagination is the source of all sorts of deceptions and must be severely constrained (Malebranche 1980, Book 2).

Francis Bacon was equally hostile to the poetic imagination and denied that it could have any part to play in the acquisition of knowledge. The imagination, on his view, may "sever that which nature hath joined, and so make unlawful matches and divorces of things" (Bacon 1864-74, Vol.3, 343). But it "hardly produces sciences"; only poetry, which is "to be accounted rather as a pleasure or play of wit than a science" (Bacon 1864-74, Vol.4, 406). Likewise, Thomas Hobbes maintained that "Fancy without the help of Judgement is not commended as a Vertue", and where it does so function "Fancy is one kind of Madnesse" (Hobbes 1962, 33). It was this that led John Locke to maintain that "if the fancy be allowed the place of judgement at first in sport, it afterwards comes to usurp it.... There are so many ways of fallacy, such arts of giving colours, appearances, and resemblances by this court dresser, the fancy, that he who is not wary to admit nothing but truth itself ..." cannot but be caught (Locke 1890, Section 33, 75). The products of fancy are epistemically and morally suspect, and this is why metaphors, allusions and figures of speech generally are described as "perfect cheats" when "we would speak of things as they are" (Locke 1961, Vol.2, 105-6).

Similar views are found in Hume and Kant. Hume regards fancy as a possible source of the destruction of human nature since, if we depend on it, "human nature must immediately perish and go to ruin" (Hume 1978, 225). It is the origin of "the loose and indolent reveries of a castle builder" (Hume 1978, 624) and the inventions of poets – each of which is "to be regarded as an idle fiction" (Hume 1978, 494). However, the fanciful imagination is "neither unavoidable to mankind, nor necessary, or so much as useful in the conduct of life..." (Hume 1978, 225).

Like Hume, Kant has harsh things to say about the fanciful imagination. In *Anthropologie* (31, VII, 174) he tells us that "phantasy" is nothing other than the uncontrolled spatial imagination which runs riot in day dreams or nightmares (Paton 1961, Vol.2, 227). However, it can be controlled, and when it is, we speak of composition. The most obvious example of this is to be found in artistic composition but even so, whether controlled or uncontrolled, the imagination thus conceived is altogether incapable of enhancing our understanding. Fancy, in Kant's words, is "the mere play of the imagination": it is an "unruly" imagination, an imagination ungoverned by rules, which can help us to make sense of our experience.

II SCIENTIFIC AND POSTMODERNIST SCEPTICISM

No set of ideas has done more to undermine the authority of science in the twentieth century than what loosely passes as postmodernism. To some – myself included – these attempts at subversion seem largely misplaced, for, whatever its shortcomings, empirical science has been remarkably successful in furnishing people with an understanding of the natural environment. Until quite recently, though, this same success encouraged an uncritical adulation of science and the scientific method. Overwhelmingly many people, encouraged by scientists and philosophers alike, were brought to believe that science alone could furnish us with useful knowledge about the world; that there are no other secure sources of knowledge and understanding.

This view, bred originally of the seventeenth century Enlightenment, has persisted well into the twentieth century. Indeed, during the first three or four decades of this century, the lure and promise of positivism was everywhere apparent. Only science was the source of reliable insight, and any claim to knowledge not based on or amenable to scientific enquiry was dismissed as bogus – as a kind of charlatanism strictly to be avoided. These are all to be found in David Hume, who persuaded successive generations of philosophers and laymen alike that useful knowledge could only be derived from sense experience; that *a priori* claims to knowledge about the world or the cosmos, whether based on idle superstition or deductive systems of thought, were and would remain entirely uninformative.

The result was an increasingly firm alliance between epistemology and the empirical sciences that has survived in one form or another for over two hundred years. In this century, and until quite recently, analytic philosophers theorised about the growth of knowledge only in a scientific context. As a result, there has been a marked tendency to concentrate on propositional knowledge, as if to suggest that it is the only knowledge that really matters. In effect, as I shall show, this meant that traditional epistemology was needlessly confined in its endeavour since it attended only to a relatively small, albeit a vitally important, area of human knowledge and understanding.

A growing sense of unease with the so-called hegemony of science, together with a barrage of claims about the epistemic value of art, led in the second half of the twentieth century to a wide range of counter-claims – either about alternative sources of knowledge, or about the unavailability of knowledge altogether. Not only has there been considerable growth in new-age therapies and methods of divination – indeed, a growth in indefensible, often harmful, quack-therapies of all kinds – but there has been, as well, a somewhat more reputable philosophical movement that has striven, in its own way, to dismantle the authority of science, and in so doing, to "empower" others who occupy different "realities".

Postmodernism has taken the academe by storm. Many postmodernists, from Jacques Derrida to Michel Foucault encourage the view that knowledge (as traditionally construed) is not properly attainable (Derrida 1974, 1978; Foucault 1980). Rather, following Friedrich Nietzsche, the suggestion is that there are "perspectives" – in their case imparted by texts or the "play of signs" – that we dignify as the truth or as knowledge (Cf. Nietzsche 1911, Vol.II, 180). But these claims to truth and to knowledge can always be "deconstructed" by showing that there are other defensible accounts of the world we inhabit, no one of which can be shown to be more adequate than any other. Consequently, what we claim as true and as knowledge is not based on how things are in themselves. Rather, if Nietzsche and Foucault are to be believed, our claims to truth, to reason, to knowledge are based on a drive for control and power in the world; on a desire to privilege our chosen perspectives because they serve us well (Nietzsche 1911; Foucault 1980). It follows from this, of course, that whatever else it affords, art cannot be a source of knowledge or true understanding about an extra-linguistic world.

Even so, the highly imaginative vagaries of art – the ideas that they promulgate and instill, and the perspectives that they advocate – have come to be seen not just as instructive but as covertly powerful. In the postmodernist agenda, works of art – most especially works of popular art – are the unspoken, unsung sources of our notions of reality; more baldly, as we shall see, they are sometimes said to create that reality itself (Baudrillard 1988, 101). This view has readily translated itself into academic practice. Historians and sociologists now see fictional literature and sometimes the popular arts as an important source of historical and sociological understanding, and some historians now consult novels rather more readily than they do archives or survey results, believing as they do that there are no pre-narrative or extra-narratorial facts that can properly constrain "the play of signs". Texts have come to be seen by some as the building blocks of historical and sociological insight – not the social world "out there", and certainly not an account of what "actually" happened and was done so many centuries earlier.

One way of undermining all scepticism about the cognitive value of art is to show that art does in fact provide us with a source of knowledge, insight, and understanding about the world – a source alternative to, but as promisingly valuable as, the empirical sciences. In this chapter, I argue that art informs us richly and diversely; hence that traditional epistemology has neglected this source of knowledge and understanding to its cost.

My claim will be that an epistemology needs to be and can be developed that explains art both as a genuine source of knowledge, insight and understanding while at the same time explaining where and how it fails us epistemically. While postmodernists are correct to think that art furnishes us with sets of understandings, I explain this without embracing any of the central claims of that cluster of theories. So, for instance, it will not follow from what I have to say that there can be no rational way of deciding between two different "visions" or "perspectives" on the world, even when both are derived from art.

III THE SCOPE OF THE DISCUSSION

The history of western philosophy, the nineteenth and twentieth-century elevation of empirical science as the only proper source of knowledge about the world, and most recently the rise of postmodernism, have all attacked the idea that art can ever be a proper source of knowledge about the world. And yet, their combined force has done nothing at all to undermine the fact that claims to knowledge feature prominently in the way that people talk about art. Critics often claim to know something about the works of which they speak and write: that *Twelfth Night* is a comedy, or that a production of it is elegant, or good, or clumsy and inept. They also claim to know that certain responses to works of art are mistaken or unwarranted; that others are appropriate. It is fitting, some think, to be disdainful of Sir Walter Elliot for his vanity in *Persuasion*, and to admire his daughter Anne for her self-control in the face of Captain Wentworth's seeming love for Louisa Musgrove. More obviously, perhaps, people also claim to learn about the actual world from art, and it is widely maintained of insightful works that they engender beliefs about the world in which we live; that they can even afford knowledge of that world. Thus, for instance, it could be said of Ivan Turgenev's *A Month in the Country* that it teaches us something about the emotional vulnerability of unreflective people as they approach and move beyond middle-age.

There are, then, at least three different types of knowledge claim that are standardly made about the arts. These are distinguished by their objects. First, it seems clear enough that just by reading and interpreting a novel or a poem, or by viewing and construing a painting, a person can acquire knowledge of the art work itself. Philosophical theories about interpretation, its epistemic structure, and its justification abound, and are crucially important to hermeneutics and the philosophy of art (Beardsley 1970; Currie 1995, Ch.8; Davies 1988, 1995; Gadamer 1988; Hirsch 1968; Iseminger 1992; Livingston 1988; Margolis 1995; Novitz 1987, Ch.5; Stecker 1996, Part II). Second, it is widely argued that a work of art can be properly understood and appreciated only if we have appropriate emotional responses to it, and the problem, of course, is to know which responses – usually emotional responses – are appropriate, which inappropriate to a particular work (Carroll 1998, 249; Feagin 1996; Hjort and Laver 1997; Novitz 1987, 75-79; Walton 1990, Ch.5). These are important issues that clearly form part of what can properly concern epistemology as it pertains to art. Even so, I will not deal with them here. My reason is simple enough: it is that when people claim to acquire insight or knowledge from a work of art, they usually have in mind a third type of knowledge claim – namely, that they have acquired knowledge not about the work (or the world of the work), nor about appropriate ways of responding to the work, but about aspects of the world external to the work (Cf. Livingston 1988, 195-9). Whether any works of art ever achieve this, and if so, how, are questions that form the focus of the ensuing discussion. In what follows, then, I assume that the work of art has been properly attended to, and that it has been adequately interpreted. The question is how, when all of this has been done, can we learn about the actual world from such works?

There are, of course, many kinds of art – many different art forms – all of which employ different techniques, and many of which exploit entirely distinctive media. This is why we cannot simply suppose that all art will function in precisely the same way to convey beliefs and knowledge. Sculpture, for instance, is likely to communicate understandings of a type and in a way quite different from those communicated by poetry or the cinema or music. Not only this, but since (as I shall show) one may learn very different things from one and the same art form, it would be silly to expect that everything that we learn even from a single art form is learned in just one way; as silly as supposing that ice hockey and algebra must be learned in the same way.

Put thus, the epistemology of art promises to be richly textured and complex; rather more so, I would venture, than the epistemology of science. Science confines its enquiries to a world "out there" – to an objective world that, if dispassionately observed by an appropriately informed and equipped observer, will deliver its secrets. These limiting assumptions define the central problem of the epistemology of science, which is to explain (and rationally endorse) the sorts of methods that result in the delivery of these secrets. Art, by contrast, is informative in many different ways and delivers much more than dispassionate facts about the world, or the sort of know-how that enables us to negotiate and control it. As we will see, it frequently requires the readers' or viewers' imaginative and emotional contributions in ways that make personal change a condition of coming to know, believe, or understand – something, that, according to Bernard Harrison, makes the acquisition of understanding and knowledge from art vastly more subversive, and in this sense more dangerous, than the acquisition of scientific knowledge (Harrison 1991, 3). Put differently, art gives us access to imaginatively and emotionally charged understandings of others and of the situations they occupy, but does so in ways that, if adopted by the empirical sciences, would be thought to distort their observations and to render their findings suspect.

Again, and quite unlike a good deal of empirical science, much art captures and conveys its insights in sounds, colours and forms that frequently under-determine its meaning; that are suggestive, highly nuanced, often figurative rather than literal. Science, by contrast, has tried, since the time of John Locke, to conduct its business and express its findings in a discourse that is routinely rigorous, unadorned, dispassionate, and precise (Cf. Locke 1961, Vol.2, 105-6.).

In the case of art, however, there is no one prescribed way of conveying its various messages (Cf. Davies 1997(b); Levinson 1997; Novitz 1997(a)). Different art forms work differently to convey ideas and to enlighten or deceive us. This is why it will be helpful to focus in this chapter on one art form in particular – an art form, of course, that is also richly informative. In this context, drama, poetry, the novel, and cinema, but also figurative painting and sculpture spring to mind. Each is cognitively verdant; much more so than the abstract arts, pure music, and most dance, which, whatever else they achieve, are not as adept at informing their audiences about the actual world.

In what follows, I will attend in the first instance to fictional literature as a source of putative insight and knowledge, although much of what I say will be true of cinema, drama, and (in some respects) poetry as well. The convenience of attending just to literary fiction resides not just in its robust yet nuanced complexity, but in the fact, too, that there is considerable agreement among readers about what it is that they learn from such works. Were we to confine our attention, say, to music as a source of knowledge about the world there would be much less agreement about what, exactly, we learn from it (Cf. Bender 1993). And since any account of how we acquire knowledge from art must be able to distinguish between justified claims to knowledge and mere "spinning", it is best to begin with an art form that commands more or less uniform agreement about what can be gleaned from it about the world in which we live.

There is another good reason for attending to literary fiction as a source of knowledge. Confining ourselves to simpler art forms would, I think, disguise the richness of art as a source of knowledge, and might hide the many textured ways in which we learn from art. The disadvantage, of course, of limiting our discussion in this way is that it will have nothing (or almost nothing) to say about ways in which the visual arts influence our understanding. Even so, I will try, where appropriate, to mitigate such deficiencies.

IV EPISTEMOLOGY AND ART

To a very large extent, fictional literature is the product of an author's imaginative powers. Given the practices that surround the production of fiction and the conventions in terms of which readers respond to it, anyone who understands what fiction is will also know that it does not purport to describe the actual world. The problem, then, for those who would defend the view that fictional literature affords insights into, an understanding of, and knowledge about the actual world, is to explain how this is possible.

Plainly, an empiricist epistemology cannot easily explain how it is possible to learn about actual situations from fictional works. The worlds that they delineate do not exist, and so cannot furnish us either directly or indirectly with the experience that empiricism requires for genuine knowledge. Nor can a rationalist epistemology easily do service. It is difficult to see how the furniture of our minds, whether it be Descartes's "clear and distinct ideas" or Kantian pure categories of the understanding, or Chomsky-like innate structures, can be invoked in order to explain how we learn from fiction. Any foundationalist strategy of this sort that emphasises certainty, or a privileged class of indubitable propositions, or universal structures of the human mind as a basis for the justification of knowledge claims based on fictional literature, will find it difficult to explain how it is possible to move from mere flights of fancy to knowledge about the world. Fancy, at best, is shifting, is not bound by rules, still less by certainties; it transgresses boundaries and frameworks, defies received opinion and the psychological certainties by which people prefer to live their lives. So even if we treat the fanciful or creative imagination as a crucial part of the evolved human mind, it is still difficult to see how traditional epistemology can allow it to play any central role in the acquisition and growth of knowledge.

Still worse, since fancy frequently offers startling, sometimes shocking ways of thinking that sit uneasily with received beliefs and ways of understanding, any attempt to justify such claims to knowledge in terms of coherence must falter. Writing of Toni Morrison, Carol Shields tells us that "a seismic shift of sensation is what we feel as we reach the conclusion of one of her books. Our bones have been rearranged, and our notions of history disordered" (Shields 1998, 16)

It would seem, then, that a good deal of what we take ourselves to learn from fiction does not cohere at all well with our established beliefs. If anything, it flouts, fragments, and disturbs them – and yet we have little doubt that we have learned something useful from the fiction; that we have acquired genuine insight into a human situation. If, therefore, one holds, as Gilbert Harman does, that whether a belief "is justified depends on how well it fits together with everything else one believes" (Harman 1986, 38), a coherentist approach, at least of this universalistic variety, plainly will not help explain how beliefs derived from fiction are justified. A less universalistic coherentism – one that requires only that such a belief cohere with a selected set of beliefs – is also problematic. For how are we to determine which set of beliefs is relevant? If relevance is regarded as a function of the content of the belief in question, we run the risk of begging the question by selecting those beliefs that are most likely to cohere with it. If, on the other hand, the selected set bears no relation to the content of the belief in question, it is difficult to see why coherence with just *this* set will serve to justify the belief that we have derived from the fiction.

Nor, it would seem, can appeals to Ernest Sosa's reliabilism (Sosa 1980, 3-25) or Lorraine Code's responsibilism (Code 1987, Ch.3) straightforwardly explain how beliefs from fiction may be justified. For however one looks at it, the imaginative fabrications and fanciful speculations of an author, no matter how brilliant or how stimulating, cannot straightforwardly be regarded as a reliable source of knowledge about the world, and it is exceedingly difficult to see why we should think it a responsible way of generating claims to knowledge.

Plainly, though, any defence of the claim that people learn from fiction has to be based on a rationally convincing epistemology. Failing this, there simply can be no reason to think the claim credible. What we need to do, then, is to determine what sort of epistemology would best enable us to defend the view that reader's can acquire not just beliefs, but understanding, insight, and knowledge from fiction.

Unfortunately, and with only a few exceptions, Anglo-American aestheticians have not had much to say on the topic, and (as we have already seen) some Continental philosophers have tended to the view that since the notion of having access to the truth is itself questionable (even politically dubious), the ideal of genuine knowledge of an independently existing world is simply misplaced, as is the ideal of uncovering a rationally defensible epistemology (Foucault 1980, 108ff.).

Certainly Anglo-American philosophers have written about knowledge and its relationship to art in general and fictional literature in particular. What they have not done is examine the epistemological underpinnings of their various claims. Since contemporary epistemology has said so little about art as a source of knowledge (Bender 1993, 593-4), and since rationalist and empiricist epistemologies clearly are not equal to the task, it is helpful to look to the romantic movement which, it is well known, protested vigorously against the passive empiricist accounts of knowledge encouraged by John Locke and David Hume. Johann Fichte, Friedrich Schelling, G.W.F. Hegel and Arthur Schopenhauer in Germany, like Samuel Taylor Coleridge and William Wordsworth in England, took from Kant the idea that our concepts or our ideas help fix the nature of reality. Unlike Kant, however, they believed that the imagination is free to create its own concepts or ideas, so that (on the romantic view) there is no fixed, unchanging core of concepts, given *a priori*, in terms of which to experience the world. Since, in addition, the romantics discarded Kant's distinction between the phenomenal and noumenal world, they contended that our imaginatively produced concepts do not just determine our experiences, but help produce the world as it really is.

A romantic epistemology thus emphasizes the role of the imagination not just in our experience of the world, but in the creation of the world itself. It is thus closely allied to romantic idealism – a position that is notoriously difficult to defend. This notwithstanding, any epistemology which emphasizes the active and creative role of the imagination in the genesis and growth of knowledge seems well placed not just to explain how we come to know the world about us, but also to explain how we learn from fictional literature.

There are very good reasons for thinking that the creative imagination plays a crucial role in the genesis and growth of all knowledge. When, for instance, we run out of knowledge or established belief with which to solve a problem, we are forced to conjecture, to guess, to imagine creatively. Our guesses or imaginings are made to do service, and, if found to be successful, will eventually pass as knowledge. Again, the new-born infant is required, on pain of extinction, to make sense of its environment but can do so neither by bringing objects under empirical concepts (since it has none), nor by induction or deduction (since there is nothing to induce or deduce from). Even if we allow that the infant is to some extent "hard-wired", and even if it is helped in some measure by its parents and others who form part of its social environment, it will still need to guess (or imagine creatively and tentatively) in order eventually to make sense of the many different objects, events, situations, and states of affairs in its immediate or remote environment.

It is sometimes thought that any such reliance on the fanciful or creative imagination in an account of the genesis and growth of knowledge must commit one to the same sort of romantic idealism as is found in the work of Coleridge or Schelling. But this is an unwarranted inference. A romantic epistemology can be realist. It can be realist not just in the sense that it allows that there are independent or objective truth conditions for empirical statements, or in the sense that certain entities exist independently of us, but also in the sense that it endorses an objectivism that does not commit its proponents to cognitive relativism.

But how can this be? Well, according to a romantic epistemology, it is only when our imaginative construals ease our confusion, when they 'tame' our experience, and enable us to negotiate the world the better, that we come to regard them as adequate to the world in which we live. For as long as our imaginings continue to serve us in this way, we can have no good reason to doubt them. As a result, we are forced to assume that these construals, and the experiences bred of them, adequately reflect the way things are. A romantic epistemology can reach this conclusion via the simple device of applying its own teachings. For the imaginative construal according to which there is an independently existing external world that has most of the features we perceive it as having, is the one that best allows us to negotiate the world and make sense of our experience of it. Any other imaginative construal about the world, its properties, and its mode of existence simply flounders. This construal allows us to make sense of the fact that the world has to be discovered, that it is not simply the product of our will, that our imaginative construals do not straightforwardly shape it, that it is intractable and is not moulded by our states of mind.

Let us suppose, then, that something like a romantic epistemology is correct (Novitz 1987, Ch.3, 55-72). Can it really help us to explain how people learn from art? I think that it can. Works of art, most especially fictional literature, invite the viewer to imagine derivatively what the artist has imagined creatively. They give the reader new mental sets, new categories, concepts and possibilities in terms of which to construe their own experience of the world. In this way, they bring the reader to experience the world differently.

V WHAT DO WE LEARN FROM ART?

That we learn about our world from literary fiction seems clear enough – at least if the testimony of ordinary readers is anything to go by. In saying this, though, there is no suggestion that what we learn is always true, although, of course, it may be and often is (cf. Palmer 1992, 181). The epistemological problem is to know how it is possible to acquire false or true beliefs, and sometimes even knowledge, about the world from fictional literature, for novels and plays are not meant to be taken as journalistic or historical or scientific reports. They are not meant to be understood as being about the real world. According to Peter Lamarque and Stein Olsen, the conventions that constitute and regulate the production and the appreciation of fiction – what they call "the practice of fiction" – require *not* that we should believe the statements of fiction but that we should imagine, or make-believe, or take it as if certain people exist in particular situations (Lamarque & Olsen 1994, 32ff).

So how do we learn from fictional literature? In order to answer this question, it is vital to distinguish the many different things that readers claim to learn from novels and plays. First, people often acquire propositional beliefs about the world from the literary works that they read, some of which may be true, others false. Of course, the acquisition of a true propositional belief does not amount to the acquisition of propositional knowledge. It is only when we are justified in believing a true proposition that is somehow derived from a literary work, that we can (if we set aside certain Gettier-like scruples) be said to have acquired knowledge from the work.

The acquisition of propositional beliefs and knowledge is just a minor component of all that may be learned from fiction. A good deal of what we learn is practical rather than propositional, for people do seem to acquire a range of skills from the literary works that they read. Some of these are behavioural, and may be suggested by the action of a novel – by what a character does, how he or she moves to solve certain problems, how they deal with people, negotiate, perhaps manipulate, their social and physical environments. More or less distinct from these large-scale strategic skills, is a range of smaller, much less visible skills, that have to do with our capacity to process information about the world. Novels and plays give us new ways of organising our thoughts, or of thinking about events in the world, and in this way bring us to think and see differently. Such cognitive skills are acquired from other art forms as well – most notably painting – although how they work in the case of this art form is quite different from the way in which they work in the case of literature. Revised conceptual understandings of this sort have traditionally been seen as of particular value, affording new perspectives that forge previously unnoticed connections in the minds of many. It is here that talk of creativity and of the creative nature of art becomes especially appropriate. Even though I will not attempt to develop a theory of creativity here, no account of art and knowledge can be complete without such a theory (See Boden 1992).

At least as important, literary fiction seems often to provide an imaginative understanding of what it feels like to be in certain situations (Walsh 1969, 101-4; Novitz 1987, 132-7). Beliefs about what it feels like to be in the situation occupied by another may be called empathic beliefs, and it seems clear that such beliefs take a direct object – which is just to say that no matter how precise and vivid your descriptions are, they will never acquaint me with the feelings of desolation and despair that are occasioned by the death of a loved one or by the unjust imprisonment of one's child. This is because empathic beliefs (and the knowledge to which they are thought to give rise) deal in felt experience, with what it feels like to live through a particular experience – and whatever the descriptive content of linguistic propositions, they cannot directly convey lived experiences of this sort – the sounds of a symphony, the grief of an orphaned child, the distress of being lost in the bush.

Finally (and quite closely related to all of the foregoing), fictional literature affects our values. This is especially true of our moral values, but is true as well of all sorts of values, whether they be religious, intellectual, artistic, economic, or environmental – all of which is readily attested to by the presence of censorship boards everywhere. If it is true that one can know how one ought to behave – if it is true, that is, that moral knowledge is possible – then (as I shall show) literary fiction is fully capable of contributing to that body of knowledge.

If all of this is true, it would seem that there is a variety of things that we learn from fiction – that we acquire propositional beliefs and knowledge about the world from fiction, strategic and cognitive skills, empathic understandings of one sort of another, and a broad range of values – primarily, of course, moral values.

The standard questions that attend such claims are these: Does art – in this case, fictional literature – merely give the illusion of instruction, or does it actually afford knowledge (which includes rationally justified belief) of the actual world? And if the latter, how could this be possible? After all, the conventions that delineate the practice of fiction, require us *not* to believe its assertions but to imagine them – to make-believe, to take it as if, to pretend, to simulate – and it is difficult to understand the relevance that games of make-believe (Walton 1990, Chs. 1 & 2) or simulations (Currie 1995, Ch.5), or the "fictive stance" (Lamarque & Olsen 1990, 32) can have to the real world. As a result, it is difficult even to begin to understand why we should think that we can learn about the world from fiction, since, at most, fictional discourse or "the practice of fiction" requires us to imagine and does not seriously assert truths of the actual world.

VI ART AND THE WORLD

A first problem, then, for any would-be defence of the view that we can learn about the world from some works of art, is to explain the relationship in which the art form in question stands to the world. For if, to return to our central example, literary fiction delineates a world of an author's imagining, and if readers are themselves meant to respond imaginatively to it, a fictional work cannot be construed as describing the actual world. Why, then, do we think that it has any relevance to the world in which we live? And if it has no relevance, how can we *learn* about that world from the work?

The question is not often asked, yet needs an answer if there is to be any credible defence of the claim that we learn about the world from fictional literature. Aristotle points us in something like the right direction, for he remarks in the *Poetics* that people take a special pleasure in recognizing whatever is familiar to them in a picture or a drama – "that the man there is so and so" (Aristotle 1962, 29). Implicit in this observation is the belief that what happens in the world of the drama can and often does resemble what happens in the real world. This, if true, would enable us to explain how fiction relates to the actual world, thereby rendering it possible for us to explain how readers acquire knowledge of the world from fiction.

The trouble, though, is that the suggestion that fiction can resemble the real world is itself the source of sceptical misgivings. When a work is fictive, Max Black writes, "there can be no question of placing it side by side with its subject to check off resemblances" (Black 1972, 117-122). And Joseph Margolis puts flesh to the bones of this objection by arguing that there cannot be "a resemblance, an independently discernible resemblance, between an actual object and an intentional, fictional or imaginary object". On his view, "where there are no actual X's, there is no actual resemblance". As a result, he contends, "the best we can do with imaginary entities is to hold that they resemble actual entities because, and only in the sense that, their *descriptions* entail that we take them to resemble actual entities" (Margolis 1980, 100-101). On this view, since genuine resemblances are always asserted in virtue of shared properties, and since fictional entities do not genuinely possess properties, there can be no genuine resemblances between literary fictions and the real world (Kjørup 1977, 27). Fictional objects, the argument goes, have the properties that they do have only in virtue of the ways in which we describe, imagine, or think of them, and so do not really have these properties at all. It follows that they cannot genuinely resemble anything in virtue of them. Any hope of explaining how we learn from fiction in terms of resemblance is thus thought to vaporise.

But all is not lost. We know that in order to create a fictional character or a fictional world, an author must imaginatively impute certain properties to an imagined subject. In creating Casaubon, for instance, George Eliot imagines a person who is scholarly, insecure, yet arrogant; someone who, while possessive of Dorothea, is also a remote husband, easily threatened in his relationship with her, and, as a result, unfeelingly cruel where she is concerned. But the cruelty, arrogance and remoteness that George Eliot imputes to Casaubon are the very properties of cruelty, arrogance and remoteness that are found in the real world. Were this not so, readers simply would not understand what was being written when George Eliot delineates Casaubon in these (or similar) terms.

The fact, then, that Eliot has created an imaginary world and peopled it with imaginary characters does not entail that the properties it mentions are non-existent. When we are told that Casaubon proposes to and marries Dorothea Brooke, we can make sense of this only if, in the world of the fiction, there are entities that actually possess the properties of being human, of being male and female, that breathe air, eat food, reproduce, are guided by a legal system, form part of social institutions. Furthermore, this will have to obtain in precisely the same sense of these words, hence in the same way, as it obtains in the actual world. Failing this, it just would not be possible to decipher the novel.

Speaking from "outside" the fiction, it is plain that fictional objects do not possess their properties in the way that real-world material objects do – this simply because material objects instantiate their properties materially, and fictional objects, since they are not material, cannot instantiate their properties in this way (Lamarque

1996, 33-4). But for any reader or viewer who is immersed in the fiction, and who speaks or describes from "within" the fiction, fictional objects instantiate their properties, and have to instantiate their properties, in just the way that non-fictional objects do. From "within" the fiction, then, fictional objects really do have properties, and the properties that they have are the very same properties that are to be found in the real world – or else are constructed of them (Novitz 1987, 122-3).

VII LEARNING FROM FICTION

Although fictional worlds are normally taken to resemble the actual world in indefinitely many respects – indeed, in all respects other than those that are explicitly denied by its author – it would be wrong to think that we learn about the real world from fiction just by noting these resemblances. For in order to notice any resemblances at all, one must *already* know that certain real-world objects possess certain properties or behave in certain ways. However, since one can only learn what one does not already know or believe,· it is clear that we do not acquire new knowledge, insights or understanding about the world in which we live just by noticing certain resemblances between fiction and it. For precisely the same reason, it is wrong to suppose that we acquire knowledge about the world from art by "a kind of transference" that is based on analogical or inductive inferences (Hospers 1946, 206). For induction of the sort referred to by Hospers involves the application in new situations of what we *already* know or believe. When, for instance, I infer inductively that Alfonso, my pet dog, will be hungry this evening, I merely apply what I already know about Alfonso to the coming evening, and do not learn anything new in so doing.

a) Propositional Beliefs, Propositional Knowledge, and Conceptual Skills

On my view, resemblances between fiction and the real world furnish an occasion for that sort of imaginative response to the work that permits us to learn about the actual world from it. Let me explain. Take the case of the acquisition of propositional knowledge from fiction. Convincing resemblances between a character and real people may bring us, in exactly the way that Aristotle supposes, to entertain certain hypotheses about actual people. If, for instance, Casaubon resembles some academics in respect of his confined interests and his love of isolation, and if, as is also the case, Casaubon demonstrates a lack of security that manifests itself in a mean-spirited desire to control others, one might venture the hypothesis that the confined interests and truncated lives of some academics are indicative of a lack of confidence that, in its turn, is productive of the desire to exercise power and control over other people.

Hilary Putnam regards such hypotheses as conceptual *knowledge* (Putnam 1978, 90). But this, I think, is misleading, for although the hypothesis furnishes us with a new way of thinking about academics, it will not count as knowledge unless this way of thinking is also found to be useful. Like all putative skills, if a way of thinking is found to confuse and mislead, it will be abandoned, and will not count as knowledge (Novitz 1987, 137-9)

In general, it is misleading to think of an hypothesis derived from fiction as knowledge of any sort. For one thing, hypotheses are not, of course, beliefs. They are imaginatively derived, highly tentative suppositions based (in this case) on certain resemblances that obtain between the fiction and the real world. Should our experience falsify a given supposition, it will normally be abandoned. On the other hand, if our experience tends to confirm it, we will eventually believe it, and, in time, with repeated confirmations, may even come to treat it as knowledge.

Importantly, one would have no hesitation in claiming of such beliefs or knowledge that they were acquired or learned from the fiction. Of course, the justifications of the various propositional beliefs derived from *Middlemarch* are not, by and large, to be found in the novel itself. But this need not prevent us from owning the novel as the source of our knowledge. Take the case of newspapers or historical treatises. We certainly claim to acquire beliefs and knowledge from them – even though the beliefs in question are not justified just by perusing the treatises and newspapers in question. One would have to do a lot of additional research in order to justify them. Even so, we often locate a particular historical work or newspaper, as the source of what we claim to know. This, of course, does not explain how it is possible to justify beliefs acquired from fiction; a question that I will turn to presently.

b) Skills, Empathic Understanding and Moral Values

Propositional beliefs and knowledge, I have said, are a minor and a comparatively insignificant part of what we learn from fictional literature. Indeed, most people turn to chemistry manuals and social histories long before they turn to novels and to plays in order to learn about the chemical and social constitution of the world around them. This, I intimated earlier, is something that is gradually changing, for some historians and social scientists like to believe that there is much to be learned about the nature of our social existence or about past periods of history from novels. The claim is that such texts work over time to shape social reality, and that there is no textually neutral way of understanding it. Whether these claims are rationally defensible is a question that I will turn to later on.

In the meantime, we should notice that the resemblances between a fictional situation – say the situation of Elizabeth Bennet in *Pride and Prejudice* – and the actual world, may suggest certain strategies; certain ways of dealing with a demanding, excitable, and weak mother who thinks of the unmarried state of any of her daughters as an affliction hardly to be borne. Here, however, the reader's imaginative involvement works very differently to achieve this epistemic end; so that strategic (and conceptual) skills are acquired in ways that are substantially different from the ways in which we acquire propositional knowledge and beliefs from fiction.

Those conventions that help constitute what Lamarque and Olsen have called "the practice of fiction", require the reader to imagine the world described by the author, and to do so by considering or entertaining the authorial descriptions of this world and its events without a mind to their truth-value. Such imagining has been described by Gilbert Ryle as "derivative" rather than "originative" or creative (Ryle

1933, 32). In point, of fact, of course, one cannot imagine in this way unless there are some discernible points of resemblance between the fictional world and the real world. Without this we would not have sufficient experience to understand the novel, and it is of course impossible for us to imagine what we cannot understand.

It is in the process of imagining Elizabeth Bennet's situation – the frustrations and irritations of daily life in the Bennet family, the constraints placed on her behaviour by society at large, her feelings for Mr. Darcy, her mortification and distress at her sister Lydia's behaviour – that we come to consider her situation from her point of view rather than our own. We identify with her, where that means that we view her situation as she does, feel anxiety and joy for, and sometimes with, her.

The process of doing this will not always be straightforward. Elizabeth's responses to her mother, her father, and her sisters might strike us as strangely artificial, as foreign and untoward. And this, as Hans-Georg Gadamer argues, requires us to assess our own deep assumptions (or "prejudices"), perhaps suspend some of them and modify others. In this process of coming to understand and respond appropriately, we impose our modified assumptions on the work, which may again show that certain of our assumptions need further modification. And so, through this "to-and-fro movement", this "play" between reader and the work, we may gradually acquire a modified set of assumptions – a modified "horizon" – that enables us to see and feel the force of Elizabeth Bennet's response to her social world (Gadamer 1988, 91-9). This "to-and-fro" movement or "play" leads to what Gadamer calls a "fusion of horizons" – a temporary or permanent adjustment of the reader's deep assumptions or "prejudices" (Gadamer 1988, 239 & p.269ff.) in ways that enable him to understand or engage with what was previously foreign and incomprehensible, and so to feel for and with Elizabeth Bennet as she negotiates the brittle reality of her love for Darcy and her awkward family.

These emotional responses, in their turn, create certain expectations, hopes, and fears, in terms of which we assess and attach significance to other events in the novel, and come to apprehend, from Elizabeth's point of view, the demands placed on her. If we consider her response to her situation admirable, her manner appropriate yet novel, we may, if faced with similar demands in our own lives, adopt similar strategies. Here certain practical hypotheses are suggested to us, which, if tried in practice, may come to be adopted as useful and proven ways of dealing with demanding situations. Of course, if our empathic beliefs – beliefs about what it feels like to be in such situations – turn out to be mistaken, the strategies based on these beliefs are also likely to mislead and to fail. The problem for the epistemology of art, of course, is to know when such beliefs are justified.

What seems clear is that literary fictions can and do impart strategic skills. In achieving this, however, the novel also imparts a range of insights and understandings that derive from the reader's imaginative response to it. For, as I have already remarked, the conventions that surround the practice of fictive story telling require its readers to respond imaginatively – to derivatively imagine – Jane Austen's descriptions in *Pride and Prejudice*, so that our understanding of the novel, its characters and their actions, is not just conceptual but is richly imaginative (Cf. Kieran 1997, esp. p.338). Attentive readers who engage properly with the fiction, and who identify imaginatively, hence feelingly, with Elizabeth, will inevitably find themselves confronted by a range of demanding ethical questions; questions that are pertinent not just to the lives of Elizabeth Bennet and her sisters, but to their own lives as well (Putnam 1978, 91). Should Elizabeth defer to her silly and petulant mother? Does she owe her a binding duty of respect? Is Mr Bennet's withdrawal from the life of his family morally defensible? How should he, and how should Elizabeth, have acted with regard to the intemperate behaviour of Lydia – encouraged as it was by the ill-considered promptings of her own mother? And so on, almost without end.

Reflection on situations like these does not straightforwardly impart new moral values. What it does, however, is test and tease one's moral beliefs and understandings; beliefs, say, about filial duty, about the respect due to social custom, and to those who occupy a higher social station in life. It is not as if Jane Austen tells us what would be morally correct, or how Elizabeth ought to behave with respect to her parents, Lydia, and Mr Darcy. It is true that the reader is encouraged to be sceptical of undue deference to social convention and custom, but we are brought up short with the realisation that a failure to be properly deferential in some circumstances is also morally questionable. The problem, of course, is to know when proper deference is morally required; and this we are never told, although, as readers, we cannot help but ponder the question. And in the process we may discover that some of our moral views about everyday social intercourse are wanting or deficient, that they need to be reconsidered, broadened, or narrowed.

Unlike philosophy, fictional literature deals in the concrete and the particular, not the abstractions of ethical theory. As a result, it introduces into our moral thinking all the special and peculiar demands of circumstance, all the complexities that test or challenge the broad ethical principles that we have uncritically imbibed at some or other stage of our lives. It is in this way that literature inclines its readers to test their moral understanding by bringing them to apply it to complex imaginary situations that have been imagined in fine and nuanced detail by their authors (Nussbaum 1990, Ch.5). At times one will come to see that one's moral beliefs and theories do not cater in any adequate way for the complex situations in which they are asked to do service. As a result, the reader knows that they are not entirely satisfactory, and there may be a shift of values consequent upon this. Our moral understanding may have altered, and we may claim in light of the fiction, to know better.

VII THE PROBLEM OF JUSTIFICATION

Hypotheses, I have said, should not be confused with beliefs; nor, of course, are beliefs to be confused with knowledge. Normally, we come to believe a particular hypothesis on the basis of confirming evidence, and if such beliefs are themselves justified, we may properly claim to know what we believe. While there are well known problems with attempts to explain knowledge as justified true belief, the rational justification of the beliefs derived from art – in this case, from fictional literature – plainly is necessary before we can claim them as knowledge.

The problem, I have already suggested, is that there is widespread scepticism about the possibility of ever justifying those beliefs about the world that are derived from fiction. For, as we have seen, like most art, literary fiction is an imaginative construct that does not purport to describe the world as it really is. It describes an imaginary world, so that it is difficult to know why we should ever think it proper to derive beliefs about the real world from it. After all, the argument goes, it is not as if fiction can furnish its own internal evidence for the correctness of these beliefs; nor, I suggested earlier, need such beliefs cohere with what we already believe and know since art often shocks and surprises and leads to highly novel ways of thinking and construing the world around us. A coherentist approach to justification, therefore, seems unavailable to us. So, too, it may be thought, are reliabilist or a responsibilist approaches, for, as both Hume and Kant maintain, there is nothing particularly reliable about "idle" fantasies or an "unruly" imagination.

But it is easy to see that all of these objections are misconceived; that they commit a kind of genetic fallacy. Certainly a purely fictional novel is an imaginary construct, and so is not directly based on, and does not purport to describe, the actual world. But it simply does not follow from this that beliefs about the world that have been derived from fiction cannot subsequently be justified in terms of our experiences or in terms of how well they cohere with existing and well-confirmed beliefs. What is more, it is possible to engage in such a process of justification responsibly or irresponsibly, relying on well-tested or disreputable means of justification. Contrary to what Plato, Hume, and Kant think, the origin of our beliefs in fiction does not prevent them from being justified in precisely the way that we justify beliefs derived from experience or from chemistry textbooks. We look to our experiences – past and future – and try to see whether or not the beliefs that we have acquired from the fiction cohere with, and perhaps make sense of, other of our core beliefs. And this can be done carefully, with due attention, patience and reflection; or it can be done hurriedly and irresponsibly.

In this way, my empathic beliefs regarding what it feels like to be in Elizabeth Bennet's shoes are shown to be justified (or unjustified) by my own experiences of embarrassing families, petulant parents, loving relationships, and so on. They may also be shown to be appropriate or inappropriate in terms of the sense that they enable the reader to make of the novel as a whole. If, for instance, my beliefs about how Elizabeth feels lead to a mistaken understanding of her motives, or make nonsense of the rest of the novel, then I would need, as Gadamer points out, to modify my beliefs until I achieve the right sort of fit – one, at the very least, that makes credible sense of the novel. So there is a clear sense in which the novel itself furnishes some evidence for the adequacy or inadequacy of my beliefs. This, of course, assumes that Jane Austen is an intelligent author and that her novels make good sense; an assumption that is supported in part by the good sense that people generally discover in her writing, but also by the testimony of what Lorraine Code calls a community of knowers. It is supported, too, by our knowledge of the fact that at least some novels are works "in which the greatest powers of the mind are displayed, in which the most thorough knowledge of human nature, the happiest delineation of its varieties, the liveliest effusions of wit and humour are conveyed to the world in the best chosen language" (Austen 1962, 37).

According to Code, we depend extensively on others for the knowledge we live by, so that we form part of, and tap into, a community of knowers – an epistemic community – that shares what it knows and that guides us and others in our various endeavours. Among other things, it tells us which authors are worth reading; which are not. Until the rise of the internet and desk-top publishing, the fact of publication was one way in which a community could signal its approval of an author's work; another was (and still is) the imprint under which the author is published; yet another are the judgements of well-informed critics. In our society, Jane Austen passes these tests with flying colours, so that we do have independent grounds for supposing that *Pride and Prejudice* makes good sense. And this fact, in its turn, contributes towards the justification (or falsification) of those empathic beliefs that we form when reading the novel.

In short, then, there are various types of experience, various sources of knowledge and bodies of belief that can help to confirm or disconfirm the various beliefs that we acquire on reading a novel. And the problem of the justification of these beliefs is no more and no less complicated than the justification of beliefs derived from a volume of history or physics.

IX SOCRATIC SCRUPLES

My claim so far is that if the beliefs derived from scientific text books and encyclopedias can be shown to be justified, so too can those beliefs that are derived from fiction– for, as we have now seen, the problem of justification is no greater here than it is elsewhere . Even so, those philosophers who are sceptical about the cognitive value of art could concede this point without abandoning their scepticism.

For on their view, the problem is not primarily with the justification of beliefs derived from art. Rather, it has to do with the devious way in which drama, novels, and the cinema instill beliefs and attitudes in people. On Socrates's view in *The Republic*, for instance, drama does persuade us but it does not do so rationally. Rather, the proliferation of life-like detail in a tragedy or a comedy, the many ways in which it appeals to our emotions and our senses, together beguile us into assenting to propositions that we have not properly considered, and that we have not subjected to rational assessment. The process, in short, is one of seduction. We find, we know not how, that we have come to a particular point of view, that our beliefs have imperceptibly altered, although not in any conscious way and not through force of reason. With effort, we may trace this alteration back to our exposure a particular work of art, or to works of a certain sort. Although these may be seen as the causal origin of our shift in attitude or belief, they need not offer any rational ground for that shift.

Many of those who saw and wept over the romantic movie *Sabrina* altered their attitudes to, and beliefs about, people who read a lot. Rather than think of them as regular bookworms, they came to think of them differently and in vastly more flattering terms. How did this come about? The film is structured in a way that brings most viewers to sympathise, in the way that widowed father might with the youthful awkwardness of the growing Sabrina (Audrey Hepburn). We come to care for her as a parent might; we see her as delicately beautiful, innocent, and vulnerable, and we want well for her. Inevitably, viewers regret the fact that her father is only a lowly chauffeur in the Larrabee household. Even so, we are brought to identify with the father, so that we experience at least some of his parental discomfort as we witness Sabrina's first uncertain and ill-judged steps towards adulthood and love. Because of this strong identification, we are delighted to learn that the job of chauffeur was chosen by him not out of need but because he wanted time to read. Here the viewers' attitudes to, and beliefs about reading and readers, are unwittingly transferred to Sabrina and her father through a kind of juxtapository metaphor (Novitz 1987, 192-8) – although there is not a single good reason for doing so. The activity of reading comes to be seen as redemptive, as a way of obviating one's low occupational status and of raising oneself in the world. No reasons are ever given in support of this proposition; the audience is beguiled, seduced, into believing it – and this is a part of what Socrates finds subversive and dangerous in the dramatic arts, and that leads him to the view that poets have no proper place in the ideal republic.

In one respect, Socrates is entirely correct. Art can and does persuade by irrational means, and sometimes has to do so in order to be effective. Elsewhere I have argued that although William Blake's poetry has come to be regarded as a finely executed and brilliant example of romantic poetry, he was just too honest to be effective. His poetry, of course, was intended as a direct assault on the social and religious beliefs of his age, and in the *Songs of Innocence and Experience* he gave people clear, ascertainable reasons for looking at and understanding their social environment differently. But the result was not rational persuasion; rather, his peers responded angrily, thought him mad, and it was left to later generations to appreciate the perceptiveness of his art (Novitz 1992, Ch.10).

This is why Socrates is right to be suspicious of the cognitive value of art. What Blake hoped to achieve required not just art, but artifice as well. He needed to convince his audience that his view of society and its ills should be adopted, but the reasons that gave were powerless against the determination of his contemporaries to preserve their privilege and their way of life. In order to displace this commitment and be cognitively effective, Blake needed to lure his audience away from a perception of their own interests, and persuade them to look to the needs of others. Any direct assault on these commitments would meet with resistance, so that was needed was an element of deception. People will be persuaded by reasons and reason-giving only if they are already committed to rational procedures, and it is notorious that their religious, their economic, or their political commitments may take precedence (Novitz 1992, Ch.5; 1997(b)).

It would seem, then, that those artists who wish to affect the thinking and the commitments of others where appeals to reason and truth are bound to fail, have the job of seducing their audiences. Seduction, in this context, is best thought of as an act of enticement that plays – sometimes openly and honestly, sometimes subtly and deceptively – on people's hopes, desires, and psychological vulnerabilities; and does not appeal, in the way that reason-giving does, to their intellectual assessment of a particular situation. Hence, as I am using the term, seduction is a non-threatening form of persuasion that does not involve an appeal to reasons, but works by enticing people to certain actions or points of view.

Rational persuasion, by contrast, involves offering reasons that are designed to demonstrate the appropriateness of one's beliefs, actions, or goals and it does so by placing oneself or others in an epistemically stronger position regarding them. The process of rational persuasion is always one in which one is made *aware* of reasons for doing or believing something or other. Those who are rationally persuaded, therefore, are always in a position to know why they have reached a particular conclusion or adopted a specific belief. In this respect, seductive persuasion differs starkly from its rational counterpart, for to have been seduced, as I have already said, is quite often not to know, and is at times to wonder, how one has come to hold specific beliefs and perform particular actions.

But even if it is true that we are often seduced into beliefs and attitudes by particular works of art, this does not of itself amount to a compelling case for scepticism about the possibility of acquiring knowledge from art. For provided that one is aware of the belief or attitude that one has acquired, one can always attempt to find reasons for it. In this way, one may come to the view that one's belief is indeed rationally justified, or else that it is an irrational idea worthy only of rejection. But although this is always possible, the rejoinder misses the force of Socrates's worry. The problem, on his view, is that the cognitive and attitudinal shifts that are bred of art often pass unnoticed by an audience or readers, whose views of the world and beliefs about reality are surreptitiously altered. Much worse, there just is no knowing how powerful art can be; how many of our attitudes and beliefs are affected in this way; how greatly our views of reality are altered by the artworks we encounter.

The only solution is to ensure that the acquisition of beliefs always involves rational inquiry. But this, Socrates thinks, can only be assured if dramatists, novelists, and poets are not around to do their damage; hence the need to banish them from the ideal state. The trouble with this, as we all know, is that it is no solution at all; poets, novelists, screen writers, film directors, playwrights abound in any actual state. Nor can we escape their (or our own) fanciful constructs, which we use to impose order on large portions of our lives (Novitz 1987, Ch.2; 1992,Ch.5). There is much, for instance, to suggest that story telling is an unavoidable part of everyday life that enables us to construe and make sense of, order, and control what would otherwise be ineffable and mysterious (Turne, 1996; Hinchman & Hinchman 1997).

If Socrates is right, such narratives, like works of art generally, do not just seduce but in so doing instill beliefs and attitudes that unconsciously furnish entire perspectives in terms of which to order, structure, and construe our environment and our lives. And to this there is no solution, for even if we attempt to drive our artists from the Republic, we will be left with others who, just like ourselves, cannot help but formulate narratives, coin metaphors, and project images in terms of which we will all unwittingly construe and misconstrue reality.

X The Postmodernist Addendum

Whereas Socrates remains optimistic, and advocates a solution to the seemingly unbridled cognitive effects of art, postmodernists do not share his optimism. On this view, there is no getting beyond the perspectives imparted by art, metaphor, narrative, in order to gauge their truth – that is, in order to find some independent justification for them. All justification is internal to the perspectives within which we are reared, so that there is no knowledge in any modernist sense – no knowledge that straddles perspectives and is neutral as between them. And from this it follows, of course, that art can no more be a source of knowledge about a neutral reality than science can be.

Perhaps an example or two from the visual arts will help illustrate the problem. When, as a young postgraduate student, I first encountered Camille Pissarro's paintings on the walls of the Ashmolean Museum in Oxford, I was somewhat bemused. What struck me as odd and wholly foreign were the discrete points of sprinkled colour that Pissarro seemed wilfully to impose on everything, but most especially on the foliage of trees. What was even more surprising, though, was my discovery soon after in trees all along the Banbury Road of just those discrete colour points. And I continue to this day to see autumn foliage in this way – even though nothing could have appeared more artificial and contrived at the time (Cf. Novitz 1977, 110-139).

E.H. Gombrich tells of a similar incident. When, as a student, he was studying Rubens's schemata for the faces of young children, he simply did not believe that children could have saucer-like cheeks, and yet, he writes "I vividly remember the shock I had while I was studying these formulas..: I never thought they could exist, but all of a sudden I saw such children everywhere" (Gombrich 1968, 144). The question of whether children really have saucer-like cheeks, or whether trees really manifest discrete colour points' might seem to be futile, since, the argument goes, these ways of drawing and painting have furnished us with visual perspectives that we cannot escape. It is the perspective, the argument goes, that determines our perceptions and the truth of our perceptions – there is no getting beyond it in order to see how things really are.

Taking strength from examples like this, Jean Baudrillard contends that in "America cinema is true because it is the whole of space, the whole way of life that are cinematic. The break between the two ... does not exist: life is cinema." (Baudrillard 1988, 101). And, of course, it follows from this that there is no way of verifying the messages of the cinema. Like any other art form, the cinema, therefore, does not impart truths or knowledge – it just imparts a perspective; one, moreover, that determines our lives and what we regard as reality.

Richard Rorty is in broad sympathy with this view. As he sees it, the idiom of our "poeticized culture" (to which the cinema, fictional literature, painting, and poetry undoubtedly contribute), goes "all the way down" so that there is no truth beyond the visions that we imbibe. On this account, there is nothing that we can point to "out there" – no extra-textual, extra-artistic, reality that impinges on our consciousness, guides or constrains our descriptions and redescriptions, and makes them true (Rorty 1989, 3-22). Hence, it will be the most seductive art works – rather than reasoned arguments and truth – that will push us into new ways of looking and understanding, and will shape our language games and our world.

On this view, life-like, finely crafted, and suggestive art forms will have considerable impact on one's world because, in the end, there is nothing "out there" that can constrain the epistemic effect of these art works. When once they gain currency, they help form what Rorty calls the "vocabularies" (Rorty 1989, 5) in terms of which we think, and in terms of which we displace as outmoded our older "vocabularies" and so establish new relationships of power, new orders, new regimes.

If one is persuaded by Rorty, then one will believe that art can be every bit as powerful as Baudrillard suggests: powerful in the sense that it moulds, shapes, controls the thinking of others, and determines what we think of as reality, as reason, and as truth. But, of course, there is no possibility on this view, of our ever coming to know how things *really* are through the cinema or any other art form. If, however, one disagrees with Rorty and allows that there is an independent touchstone of truth relative to which the pronouncements of the cinema can rationally be assessed, then beliefs derived from the cinema or any other work of art can either be justified or subverted by appeals to reason and truth, so that knowledge is, after all, possible.

The problem for those who would defend the view that, despite their seductive properties, works of art are an important source of knowledge about the world, is to show that someone like Rorty is mistaken. In order to do this, one would need to show that one can give reasons for the perspectives or "vocabularies" that we derive from art; reasons, moreover, that establish such "vocabularies" as adequate, and the beliefs they engender as true or false. By showing this, we will have come some way towards solving the Socratic problem – first, by showing that it is indeed possible to become reflectively aware of the "vocabularies" or perspectives that, so to speak, house one's experience of the world; and, second, that the beliefs that derive from such perspectives or "vocabularies" can themselves be rationally warranted.

There are different ways of doing this (Novitz 1992, Ch.10). One is to resist the rigid distinction that Rorty advocates between individual sentences on the one hand, and "whole vocabularies" on the other, for on his view, while individual sentences can be justified in terms of our experience, "whole vocabularies" – the visions, perspectives, or "idioms" imparted inter alia by works of art – cannot be justified or displaced through experience of how things really are (Rorty 1989, p.5).

It is easy to show that Rorty is mistaken at this point. One thing at least seems clear about any two competing "vocabularies" or perspectives: it is that people entrenched in one of them, must nonetheless be able to understand the other in order to know that the two compete. But the only way in which I can grasp the fact that the insights imparted by Dryden are in competition with those imparted by Blake is by having a grasp of the conditions under which the assertions derived from their competing "vocabularies" would be true. Still more, we can only know that these artistic perspectives compete with one another because the assertions derived from them are incompatible. We can know this, however, only because the truth conditions of the respective assertions can be seen to differ. In other words, we must know what sorts of experiences of the world would make Blakean assertions true or false, and what sorts of experiences would vindicate Dryden's vision. Since this is so, it follows that competing "vocabularies" or competing artistic perspectives covertly appeal to precisely the same concept of truth – and, that contrary to all of Rorty's denials, the concept of truth is neutral as between the two. Unless it were, there could, as I have said, be no apprehension at all of the fact that these artistic visions compete and are exclusive of each other.

It now turns out that competing paradigms, "vocabularies" or artistically derived perspectives must have at least one concept in common; one, moreover, to whose instantiation we can now appeal as a reason for discarding certain of them. Of course, philosophers can still have different theories of truth; this is not in question. The crucial point, though, is that however we explain the concept of truth, we must all have it.

This is why it is just wrong to suppose that we cannot assess the adequacy of Rubens's or Pissarro's renditions of children or trees. There is what I have elsewhere called a two-way cognitive relation between a picture and its viewers (Novitz 1977, 109-111) – or what Gadamer might call a relationship of "play" – such that viewers may indeed come to notice aspects of the world that were previously unnoticed, where such newly acquired observational skills reciprocally allow them to notice aspects of the painting that had previously passed unremarked. But the process of adjustment depends at every turn on the viewer's ability to find in the world what the painting suggests. Had Rubens depicted children with box-like cheeks, we would not have noticed them in the world; and had Pissarro depicted trees with uniformly purple leaves, we would not find such paintings true to our experience. So it is not as if such drawings and paintings create our visual world; at best they guide our ways of looking and seeing, and bring us to see what was always there but previously unnoticed.

XI AUTONOMIST QUALMS

But even if it is the case that we can acquire true and justified beliefs from works of art, it can and has been argued that this is not the proper purpose of art. Writing of literary fiction, for instance, Lamarque and Olsen offer a "no-truth" theory of literature, and argue that it is not part of the socially prescribed response to literature to think of it as delivering truths about the world. If this is right, literature (and presumably art in general) should not be understood as a vehicle for delivering knowledge of any sort at all. This, when it happens, is something that is wholly extrinsic to the notion of literature, and certainly does not add to its value (Lamarque & Olsen 1994, 289-397).

If they are to be believed, it is simply false that literary works "have the constitutive aim of advancing truths about human concerns..." (Lamarque and Olsen 1994, 368). "Literature" is understood as an evaluative concept, so that the correct literary stance is always one of appreciation. Since literary appreciation is a practice that is constituted by a set of conventions and concepts "which both regulate and *define* the actions and products involved in the practice" (Lamarque and Olsen 1994, 256), some concepts and standards are central to it, while others, since they do not form an integral part of literary practice, are considered wholly foreign to it. "Truth" falls into the latter category. Hence any attempt by philosophers to evaluate the insights afforded by a novel in terms of its truth, amounts to a violation of literary practice; an unwarranted appropriation of literature to serve ends for which it is not suited, and for which it was not intended.

But notice how odd this claim is. According to Lamarque and Olsen, a "central defining feature" of literature is that it has something "interesting to say about human life" (Lamarque and Olsen 1994, 278). The thematic statement that emerges from Lydgate's experiences in *Middlemarch* "that noble human desires and aspirations are thwarted by forces beyond an individual's control" is said by them to be the locus of literary interest: it "is the *content* of the proposition, what it is about, not its truth as such, that confers interest..." (Lamarque and Olsen 1994, 330). But it is difficult to see how one could have an *interest* in such a proposition without having any concern for its truth. One perfectly ordinary and perhaps inevitable way of displaying one's interest in the content of a general thematic proposition (as opposed to a fictional sentence in a work) is to wonder whether it describes things as they really are. An interest in content and an interest in truth are intimately related, and while the conventions that regulate the practice of literary appreciation may discourage us at any given time from inquiring into the truth of thematic statements, there can be no doubt that there have been times when the truth of such statements has been regarded as of central interest and an obvious mark of literary value.

Lamarque and Olsen do not deny that we learn from literary fiction and from art. If anything, they treat it as obvious (Lamarque and Olsen 1994, p.5). That we acquire knowledge from fiction, they think, "can be trivially conceded"; the important question is "what role such knowledge plays in literary appreciation" (Lamarque and Olsen 1994, 13). But the acquisition of knowledge from literary fiction, we have now seen, is by no means a trivial matter. Since the epistemic richness of literature, the insights and understanding it encourages, heighten readers' appreciation both of the world and of their lives, it is something that deserves a serious philosophical explanation.

Indeed, given the impact of such knowledge on human lives, it is oddly doctrinaire to insist that the cognitive value of literature has no bearing at all on literary value. That people actually evaluate literary works in terms of their cognitive effects seems incontrovertible; yet it is precisely the *possibility* of doing so within the institutional practice of literature that Lamarque and Olsen seek to deny. Their argument for the view is straightforward: were truth central to literary appreciation, they tell us, critics would give arguments for the various propositions and insights that they derive from literature. The fact that they do not, is taken to show that, despite all appearances, considerations of truth never really enter into literary appreciation (Lamarque and Olsen 1994, 332 & 368).

To argue in this way, however, is to confuse what they call the literary stance with its philosophical counterpart. A philosophical stance would clearly require reasons and arguments for a proposition that was claimed as true. But there is no reason why a literary stance should require this. The fact that a thematic statement derived from a novel is confirmed by a critic's life's experiences, is often enough (within one version of the literary stance) to earn the critic's praise and approval of the work in question. Given that people do, as a matter of fact, appreciate literature in this way, it plainly is possible to value a text for the truths that it affords (or is supposed to afford) without treating it as a philosophical text.

Lamarque and Olsen inadvertently confuse conventions that are thought to constitute the practice of literary appreciation with those that merely regulate it. Certainly there have been times when the didactic function of art has been considered extrinsic to its value *as* art. But this, it seems to me, is no more than an arbitrary fashion of art appreciation; one that regulated the practice of art appreciation for a while last century, and then disappeared as people began to rise above fashion and to acknowledge once more the extent to which art enriches their cognitive awareness; how deeply and richly it is integrated into all of our cognitive lives.

XII CONCLUSION

In part, my aim in this article has been to explain how knowledge and understanding can be derived from art, using fictional literature as a central example. But I have tried to do more than this. For, in the process, I have tried to discredit a broad cluster of theories that are sceptical about the possibility of ever acquiring knowledge and insight from art. In this matter, I have argued, those who would confine genuine knowledge about the world to scientific practice and procedure are as mistaken as those postmodernists who defend a kind of perspectivalism, and who consequently deny that knowledge of any sort is possible. Those, on the other hand, who see art as largely autonomous, hence properly unconcerned with the business of imparting truths about the world, are also at fault – although, of course, for a range of different reasons. I do not, of course, wish to suggest that my arguments against these opponents are decisive. All need further development. At best, I have pointed the way.

David Novitz
University of Canterbury

REFERENCES

Aristotle: 1962, *On the Art of Poetry*, translated by Ingram Bywater, Clarendon Press, Oxford.

Augustine: 1956, *Letters*, in *A Select Library of Nicene and Post-Nicene Fathers*, translated by J. Cunningham, 14 Vols., Wm. B. Eerdmans, Grand Rapids.

Austen, J.: 1962 (1818), *Northanger Abbey*, Collins, London.

Bacon, F.: 1864-74, *Works*, J. Spedding, R.L. Ellis, and D. D. Heath (eds.), 14 Vols., Longmans, London.

Baudrillard, J.: 1988, *America*, translated by C. Turner, Verso, London and New York.

Beardsley, M. C.: 1970, *The Possibility of Criticism*, Wayne State University Press, Detroit.

Bender, J. W.: 1993, 'Art as a Source of Knowledge: Linking Analytic Aesthetics and Epistemology', in J. W. Bender and H. Gene, *Contemporary Philosophy of Art*, Blockerm Prentice-Hall, Englewood Cliffs.

Black, M.: 1972, 'How Do Pictures Represent?', in E. H. Gombrich, J. Hochberg and M. Black, *Art, Perception and Reality*, Johns Hopkins University Press, Baltimore, 1972.

Boden, M.: 1992, *The Creative Mind: Myths and Mechanisms*, Cardinal, Reading.

Carroll, N.: 1998, *A Philosophy of Mass Art*, Clarendon Press, Oxford.

Code, L.: 1987, *Epistemic Responsibility*, Brown University Press, Hanover and London.

Currie, G.: 1990, *The Nature of Fiction*, Cambridge University Press, Cambridge.

Currie, G.: 1995, *Image and Mind: Film, Philosophy, and Cognitive Science*, Cambridge University Press, Cambridge.

Davies, S.: 1988, 'True Interpretations', *Philosophy and Literature* **12**.

Davies, S. (ed.): 1997(a), *Art and Its Messages: Meaning, Morality, and Society*, The Pennsylvania State University Press, University Park, Pa.

Davies, S.: 1997(b), 'Introduction', in S. Davies, *Art and Its Messages*, pp. 290-7.

Derrida, J.: 1974, *Of Grammatology*, translated by G.C. Spivak, Johns Hopkins University Press, Baltimore.

Derrida, J.: 1978, *Writing and Difference*, translated by Alan Bass, Routledge & Kegan Paul, London.

Descartes, R.: 1931, *The Philosophical Works of Descartes*, translated by E.S. Haldane and G.R.T. Ross, Cambridge University Press, Cambridge.

Feagin, S.: 1996, *Reading with Feeling*, Cornell University Press, Ithaca & London.

Foucault, M.: 1980, *Power/Knowledge*, Harvester Press, Brighton.

Gadamer, H.-G.: 1988, *Truth and Method*, Crossroad, New York.

Gombrich, E.H.: 1968, *Art and Illusion: A Study in the Psychology of Pictorial Representation*, Third Edition, Phaidon Press, London.

Harman, G.: 1986, *Change in View*, MIT Press, Cambridge, Mass.

Harrison, B.: 1991, *Inconvenient Fictions: Literature and the Limits of Theory*, Yale University Press, New Haven.

Higgins, K. M.: 1991, *The Music of Our Lives*, Temple University Press, Philadelphia.

Hinchman, L. P. and S. K. Hinchman (eds.): 1997, *Memory, Identity, Community: The Idea of Narrative in the Human Sciences*, SUNY Press, New York.

Hirsch, E. D. Jr: 1967, *Validity in Interpretation*, Yale University Press, Yale and London.

Hjort, M. and S. Laver (eds.): 1997, *Emotion and the Arts*, Oxford University Press, New York & Oxford.

Hobbes, T.: 1962, *Leviathan*, Dent, London.

Hospers, J.: 1946, *Meaning and Truth in the Arts*, University of North Carolina Press, Chapel Hill.

Hume, D.: 1978, *A Treatise of Human Nature*, Second Edition, analytical index by L.A. Selby-Bigge, edited by P.H. Nidditch, Clarendon Press, Oxford.

Iseminger, G. (ed.): 1992, *Intention and Interpretation*, Temple University Press, Philadelphia.

Kieran, M.: 1996, 'Art, Imagination and the Cultivation of Morals', *The Journal of Aesthetics and Art Criticism*, **54**, 337-351.

Kjørup, S.: 1977, 'Film as a Meetingplace of Multiple Codes', in D. Perkins and B. Leanders, *The Arts and Cognition,* Johns Hopkins University Press, Baltimore.

Lamarque, P. and S. H. Olsen: 1994, *Truth, Fiction, and Literature: A Philosophical Perspective*, Clarendon Press, Oxford..

Lamarque, P.: 1996, *Fictional Points of View*, Cornell University Press, Ithaca.

Levinson, J.: 1997, 'Messages in Art', in Stephen Davies, *Art and Its Messages: Meaning, Morality, and Society.*

Livingston, P.: 1988, *Literary Knowledge*, Cornell University Press, Ithaca and London.

Locke, J.: 1890, *Of the Conduct of the Understanding*, Third Edition Clarendon Press, Oxford.

Locke, J.: 1961, *An Essay Concerning Human Understanding*, J. W. Yolton (ed.), 2 Vols., Dent, London.

Malebranche, N.: 1980, *The Search After Truth*, translated by T.M. Lemmon and P.J. Olscamp, Ohio State University Press, Columbus, Ohio.

Margolis, J.: 1980, *Art and Philosophy*, Humanities Press, Atlantic Highlands, N.J.

Margolis, J.: 1995, *Interpretation – Radical But Not Unruly*, University of California Press, Berkeley and Los Angeles.

Nietzsche, F.: 1911, 'On Truth and Falsehood in their Ultramoral Sense', in O. Levy (ed.), *The Complete Works of Friedrich Nietzsche*, translated by M. A. Mügge, 18 volumes, Allen & Unwin, London.

Novitz, D.: 1977, *Pictures and Their Use in Communication: A Philosophical Essay*, Nijhoff, The Hague.

Novitz, D.: 1986, 'The Rage for Deconstruction', *The Monist*, **69**, 40-53.

Novitz, D.: 1987, *Knowledge, Fiction and Imagination*, Temple University Press, Philadelphia.

Novitz, D.: 1992, *The Boundaries of Art*, Temple University Press, Philadelphia.

Novitz, D.: 1997(a), 'Messages 'In' and Messages 'Through' Art', in S. Davies (ed.), *Art and Its Messages*.

Novitz, D.: 1997(b), 'The Anaesthetics of Emotion', in M. Hjort and S. Laver (eds.), *Emotion and the Arts*.

Nussbaum, M. C.: 1990, *Love's Knowledge: Essays on Philosophy and Literature*, Oxford University Press, New York & Oxford.

Palmer, F.: 1992, *Literature and Moral Understanding: A Philosophical Essay on Ethics, Aesthetics, Education and Culture*, Clarendon Press, Oxford.

Paton, H.J.: 1961, *Kant's Metaphysic of Experience*, 2 volumes, Allen & Unwin, London.

Plato: 1952, *Phaedrus*, translated by R. Hackworth, Cambridge University Press, Cambridge.

Putnam, H.: 1978, 'Literature, Science and Reflection', in H. Putnam, *Meaning and the Moral Sciences*, Routledge and Kegan Paul, London

Rorty, R.: 1989, *Irony, Contingency, and Solidarity*, Cambridge University Press, Cambridge.

Ryle, G.: 1933, 'Symposium on Imaginary Objects', *Proceedings of the Aristotelian Society*, Suppl. Vol. **13**, 18-43.

Searle, J.: 1979, 'The Logical Status of Fictional Discourse', in John Searle, *Expression and Meaning: Studies in the Theory of Speech Acts*, Cambridge University Press, Cambridge.

Shields, C.: 1998, 'Paradise, by Toni Morrison', *Guardian Weekly*, 25 January.

Sosa, E.: 1980, 'The Raft and the Pyramid', in P.A. French, Y.E. Uehling Jr., and H.K. Wettestein (ed.), *Midwest Studies in Philosophy V*, The University of Minnesota Press, Minneapolis.

Stecker, R.: 1996, *Artworks*, The Pennsylvania State University Press, University Park.

Turner, M.: 1996, *The Literary Mind*, Oxford University Press, Oxford.

Walsh, D.: *Literature and Knowledge*, Wesleyan University Press, Middletown, Connecticut.

Walton, K.: 1990, *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Harvard University Press, Cambridge Mass. & London.

Young, J.-O.: 1996, 'Inquiry an the Arts and Sciences', *Philosophy* **71**, 255-273.

KATHLEEN LENNON

FEMINIST EPISTEMOLOGY

I EPISTEMOLOGY AND FEMINISM

Epistemology has become a crucial issue for feminism and feminism has become a crucial issue for epistemology. Feminist epistemologists, in common with many other contemporary thinkers, no longer regard knowledge as a neutral transparent reflection of an independently ordered reality with truth and falsity established by transcendent procedures of rational assessment. Rather most accept that all knowledge is situated knowledge, reflecting the position of the knowledge producer at a certain historical moment in a given material and cultural context. Interconnectedly feminism, along with other movements, has recognised the links between knowledge and power. The legitimisation of knowledge claims is intimately tied to networks of power relations. This recognition has moved epistemological issues into the forefront of contemporary culture. If we cannot distinguish between good and bad knowledge claims by the application of neutral and transcendent criteria how are we to address traditional epistemological concerns regarding justification and the distinction between genuine knowledge and what merely passes for it? This is not of only philosophical interest, for feminism has a commitment to social change; it wishes to establish the legitimacy of its critiques of the existing social order and devise effective strategies for change.

II KNOWLEDGES AS MASCULINE

These epistemological dilemmas have been generated by extensive feminist work exposing the masculinity of different areas of knowledge. Social and natural science attracted a great deal of attention but also literature, philosophy, history etc. It is important to pay attention to the variety of things that could be meant by calling some domain of putative knowledge masculine, for this has consequences for later attempts to devise adequate epistemological strategies. The easy and uncontroversial point is that much of what has been recognised as knowledge and passed on in academic and industrial circles has been produced by men. Consequently their experiences and concerns have served to determine its direction. History was accused of omitting herstory, art and literature of privileging male writers and artists, science of devising research directions in which women were considered only as consumers. These criticisms did not threaten the legitimacy of the research which was produced, only it's restricted range. The accusations of masculinity were more damaging when the claim was made that the theories, which had been put forward as putatively universal, did not make sense of female lives and experiences and were therefore empirically inadequate. Carol Gilligan's (1982) discussion of

I. Niiniluoto, M. Sintonen and J. Woleński (eds.), Handbook of Epistemology, 1013-1026.
© 2004 *Kluwer Academic Publishers.*

Kohlberg's theories of moral development were an example here. She argued that his conception of moral development and maturity based on experiments only with boys did not capture the patterns of moral reasoning found in girls. Similarly feminist critiques of liberal political theory (see e.g. Jagger 1983) suggested that it was not a framework that could be applicable to even all adults in the society, for the autonomous ideal was achievable only if some members of the society were playing a servicing role, raising the next generation and caring for the sick and old. The work of female primatologists on the behaviour of female apes (see Haraway 1989) showed that they played a much more major role in social organisation than had previously been suggested. Pilot studies on women with heart disease suggest the inapplicability wholesale of causal factors isolated from studies exclusively on men (Pitt 1998).

Equally damaging were criticisms which illuminated areas of knowledge as ideologically masculine; claiming that the theories constructed were working to legitimate inequalities and reinforce relations of domination. Examples were criticisms of sex differences research over several hundred years (see Longino 1990, ch.6.) suggesting that women's intellectual and physical capabilities were inferior to men's or particularly suited them to a caring role in society. Much primatology and other animal studies (Haraway 1978; Bleier 1988) served to interpret the animal world through the eyes of the existing social order and then use such an interpretation to justify that order as natural. Functionalist accounts of the nuclear family (Talcott Parsons 1951) argued that the patriarchal family was necessary for social stability. Parts of sociobiology (Lewontin, Rose and Kamin 1984) suggest that men are genetically programmed to scatter their sperm as widely as possible and women to attempt to entrap them into caring for offspring, against their better interests. (Here the structure of feminist criticisms echoed that of Marxist theorists who highlighted the ideological effect of much of what passed as knowledge, in reinforcing bourgeois interests. Parallel claims can also be made of the way in which putative knowledge supported imperialist activities and so called racial hierarchies (Harding 1993)).

Most far reaching of the attributions of masculinity has been the claim that the symbolic order by means of which knowledge claims are articulated privileges the male. There has been much broadly deconstructive work which attends to the texts within which knowledge is articulated, interrogating the structures of narrative, images and metaphors which are found there. (Fox Keller 1992) Careful unpicking and attention to the structures of language and metaphor made evident the hierarchical oppositions which underpin literary, historical, philosophical and scientific texts, and their interdependencies with gender hierarchies. (An easy example here, which doesn't take much unpicking, comes from Emily Martin (1991), who highlights the way in which conventional biological accounts of fertilisation are laden with sexist metaphor. In this conventional account sperm are described as active, battling valiantly from vagina to the oviduct and penetrating the egg, thus engendering new life. In contrast the passive egg is shed by the ovary and swept down the fallopian tubes to await its date with destiny! Given the biological reality, in which the egg's adhesive surface traps the sperm, Martin suggests a more appropriate model is to regard sperm and egg as mutually interacting in a process marked by 'feedback loops' and 'flexible adaptation'). It is important to notice here how intregal the metaphors are to the articulations of the process, structuring our conceptions of the reality, and indeed what it is possible for us to observe. As noted by Longino and Fox Keller 'metaphors guide the construction of similarities and differences- i.e. our very categories of analysis'. (Fox Keller and Longino 1996, introduction ch.7.) When Martin puts forward an alternative account of fertilisation she does not do this simply by shedding metaphor and opting for 'literal' descriptions, but rather by employing new metaphors, ('feedback loops'). It is not therefore possible to regard the gendered nature of much language and metaphor as a detachable extra, removable from the articulation of areas of knowledge, to leave an ungendered content intact. The content is tied necessarily to its mode of articulation.

Appreciating the textuality of what is offered as knowledge enables us to see that the way in which knowledge is gendered is not simply in reflecting the interests and experiences of those who produce it. The conceptual frameworks employed can themselves be gendered in a way that can continue even when women enter into the knowledge producing process.

An area where such deconstructive techniques have been particularly important for highlighting masculinity is epistemology itself. Here interrogations of conceptions of knowledge, including scientific knowledge, and conceptions of rationality, by means of which genuine knowledge is supposed to be achieved, have revealed gendered hierarchies structuring the theories which have been put forward. In 1984 Genevieve Lloyd, in a groundbreaking book, (Lloyd 1984) looked at the differing conceptions of rationality found in the history of western philosophy. What she finds is that, although there are changes in the way in which rationality is conceived, as a notion it was defined within a symbolic system which constructs notions of rationality and irrationality interdependently with constructions of masculinity and femininity, so that, by definition, as it were, women are less rational and less capable of reaching objective and thereby true knowledge than men. Such interdependence still seems to be in play. On many standard conceptions of, for example, scientific knowledge, the validity of the knowledge is seen to depend on its being subjected to a process of scientific testing which ensures that the subjectivity of the knowledge producers can be removed. In such a way, on some accounts, genuine knowledge will reflect the way the world is itself, untainted with subjectivity. It is woven into our hegemonic conceptions of (white, professional) masculinity that men are capable of detaching themselves from the objects of their study and reaching judgements untainted with emotion; while women are anchored in the emotional and the particular and have difficulty in making objective (meaning detached) judgements. Our conception of, in particular scientific knowledge is, in this way, conceptually interwoven with a conceptualisation of it as male.

One consequence of these conceptual and metaphoric interdependencies is to make it difficult to think together 'rational woman' or 'scientific woman', (a consequence not unconnected with the problem in the west of encouraging girls to take science or study philosophy). A further consequence is that these associations privilege a particular conception of knowledge, one that is tied up with just those hegemonic conceptions of masculinity. Knowledge is primarily conceived of as representations of a world, paradigmatically expressed in propositions, the truth or falsity of which can be assessed from no particular position. Other kinds of knowledge, especially embodied and practical knowledge, kinds of knowledge traditionally associated with women, are secondary to this representational and propositional paradigm and are usually given scant attention. This not only skews our account of women as knowers; it skews our account of knowledge itself. (Dalmiya and Alcoff 1992; Gonzalez Arnal 1999.)

III FEMINIST RESPONSES

Feminist articulation of the gendered nature of what counts as knowledge for us made clear the extent to which knowledge reflects both the subjectivities of the knowledge producers and inter-connectedly their position within a culture. The knowledge produced is not the outcome of a transcendent process moving towards a 'god's eye' view of the natural and social world, but a historical, social and cultural product, which reflects the contingencies of just those factors. Such a recognition was reinforced by archaeological work, careful historical excavations of the emergence of particular theories which paid attention to the concrete material practices and negotiation between different sites of power out of which what gets counted as 'facts' emerge. (Oudshoorn 1994.) The feminist conclusions here regarding the nature of knowledge production cohered well with the writings of many post positivist philosophers of science. Writers such as Thomas Kuhn and Paul Feyerabend had argued against the position that scientific theories were produced and accepted purely on the grounds of empirical adequacy. (Kuhn 1962; Feyerabend 1975.) Following the work of Quine (1953), who pointed out both the under determination of theory by data and the lack of bruteness of empirical facts, such writers argued that our classifications do not simply reflect an already categorised reality but rather mediate our observations of the world. Moreover our choice of theory is dictated not only by predictive and technological success, (taken to be a marker of empirical adequacy), but also by social, historical, cultural and aesthetic factors.

What have the responses been of feminists concerned with epistemological questions to the recognition that what passes for knowledge has been masculine in the complex numbers of ways outlined above? In 1986 Sandra Harding in a book called *The Science Question in Feminism* (1986) outlined what she considered to be the range of possible responses here. Although the classifications she outlines are ones into which few, if any, contemporary writers would neatly fit, it is worth considering them as a way of mapping possible moves.

What Harding terms feminist empiricism is the least threatening to traditional conceptions of knowledge. On this view the masculinity of knowledge has been a consequence of male knowledge collectors allowing personal bias to effect their work, and thereby not applying the standards of objective testing in a rigorous enough way. The assumption is that female investigators, alert to this possibility and reflective concerning their own positionality, would be less susceptible to bias and produce more objective results. Such a response retains a commitment to the empirical adequacy of knowledge to an independent reality, which makes the weighing and assessment of evidence a crucial part of the epistemological process. It is recognised, however, that attention to the conditions of knowledge production may count as part of the relevant evidence. Nonetheless the problem remains that the kind of objectivity aspired to here does not seem achievable, nor is there any reason why, if it were, women should be particularly good at achieving it. It is not just in the production of ideologically distorted knowledge that our account of nature is mediated by culture, but entirely generally. Donna Haraway's work on primatology illustrates the point well (Haraway 1989). When female primatologists entered the field they produced different stories of the behaviour of the apes. But these stories, bringing into view aspects of the apes' behaviour which had not been visible before, nonetheless reflected the concerns of the primatologists who were producing them and employed the linguistic and metaphorical resources of their culture.

The second response, which Harding considers, is that of feminist standpoint theory. Such a view accepts that all knowledge reflects the historical and cultural position of the knowledge producers. The feminist project was then seen as that of producing knowledge from female subject positions. This was not simply in the interests of balance or filling in of gaps, for the knowledge from female subject positions was seen as privileged. Such positions are more likely to produce true and reliable knowledge. Such a move owes a clear and acknowledged intellectual debt to Hegel's master/slave dialectic (Hegel 1977) and to a Marxist epistemology where privilege is accorded to the position of the proletariat. (Lukacs 1971.) Within Marxism the privilege which attaches to the position of the working class derives from their position in production. This makes them central to the material production of the social order while remaining marginal to the production of knowledge about it. When this framework was adopted by early feminist standpoint theorists there were a number of different considerations put forward to defend women's privileged epistemic status. (Smith 1987; Hartsock 1983) One was their role in the reproduction of everyday life, or connectedly their practical engagements with a world unmediated by abstractions. For others it was their marginal status which was crucial, a view to which I will return below. (Harding 1991)

There were a number of objections to standpoint theory in this early form. First was the lack of homogeneity within the category 'women'. Women, due to the variety of social locations which they occupy, have diverse experiences, life histories, perceptions and modes of agency. (Spelman 1988.) Arising out of these are diverse beliefs and ethical and political objectives. This line of argument opposes certain strands of essentialism running through some versions of feminist thought. Essentialist thought assumed that certain bodily forms yielded common experiences/life expectations/modes of knowledge collecting which could find expression in female ways of knowing. (Belenky et al. 1986) For some writers the absence of such commonalities, which the differences between women make apparent, undermines the coherence of feminist epistemological projects altogether.

A second problem was expressed by Donna Haraway in the following way. "Standpoints of the subjugated are not innocent positions. They are not exempt from critical re-examination, decoding, deconstruction and interpretation." (Haraway 1991, 191.) This is, in part, because of the points made clear above. We all have to make sense of our world in terms of the discourses that are available to us. Experience itself reflects and is partially constructed out of the self-understandings yielded by the imaginary and symbolic dimensions of our conceptual apparatus. Even the experience of a marginalised group is not necessarily a source of undistorted knowledge.

The third possibility which Harding outlines, she terms feminist postmodernism. Such a position abandons the justificatory concerns which characterise most traditional epistemology and which were retained in the first two responses she considers. Recognising the impossibility of producing universal standards of truth and rationality, and the complicity of those standards which were supposed to play this role in structures of domination, we give up on the project of trying to show in any entirely general way that certain accounts of the world are more valid than others. Instead our defense of preferred narratives becomes strategic and small scale and possible only amongst those who already share agreement in judgements. (Nicholson 1990.) For many this position is required by the recognition that there is no unmediated access to a world. Rather we encounter it in a way framed by the pre-judgements we bring to it. Moreover the subjects, who, in traditional epistemological projects, are attempting to stand outside of their positions to gain a 'god's eye view' of the world are themselves constituted out of the multiplicity of positions in which they stand and the multiplicity of discourses which give them their (fragmented and contradictory) self understandings.

For many feminists however such a postmodern moment remains problematic. They wish to be able to engage in critique of those knowledge collectors who do not share their assumptions, for example those employing naturalising discourses around sex differences. They also, as a result of their emancipatory projects, need narratives about the world which will facilitate effective interventions. The recognition, which attention to so called post modernist writers, forces on us, namely a recognition of the textuality and locatedness of knowledge, does not, for these, lead to the abandonment of traditional epistemological questions, but highlights their complexity. Haraway articulates the dilemma in the following way. We are forced

> "to accept two simultaneous, apparently incompatible truths. One is the historical contingency of what counts as nature for us; the throughgoing artifactuality of an object of knowledge, that makes it inescapably and radically contingent...and simultaneously ...discourses make claims...they have a sort of reality to them which is inescapable. No ...account escapes being story laden, but it is equally true that stories are not all equal here. Radical relativism just won't do." (Haraway 1991b, 2.)

All of our knowledge is laden with the cultural and social location from which it emerges. But it is nonetheless knowledge of something independent of itself, about which it can be providing more or less adequate accounts. So we cannot avoid questions of justification.

IV RETHINKING KNOWLEDGE AND OBJECTIVITY: HARDING AND LONGINO

Contemporary feminist epistemologists are difficult to assign to the categories which Harding outlined, for all are in some way touched by each of the moments of thought she identified. Most reject the possibility of a god's eye view, detached from particular locations, from which 'one true story' of the world can be produced. Most reject as both impossible and undesirable notions of objectivity which insist on the detachment of knowledge from positionality, recognising that subjects are necessarily implicated in the knowledge which they produce. Consequently our knowledge is necessarily partial, reflecting both certain perspectives onto the world and particular directions of concern. Our putative knowledges are also recognised as texts, constituted from the linguistic and metaphoric resources of our culture and, susceptible to deconstructive critique . Simultaneously, however, many theorists also recognise that our knowledge claims make demands of a reality to which they can be more or less adequate. They therefore need to concern themselves with questions of evidence and assessments of validity. Criteria for assessment, however, develop holistically against backgrounds of conceptualisations, beliefs and assumptions, in which it is not possible to disentangle factual and evaluative notions, and in relation to which considerations of the 'context of discovery' and the methods whereby the theory has gained acceptance enter into the 'context of justification'. Connectedly feminist epistemologists tend to insist on the accountability of knowledge to the communities it is designed to serve. (Code 1991) It is worth expanding on this last point. Given the acknowledgement amongst many post positivist epistemologists that theory is under-determined by data and therefore criteria of empirical adequacy don't fix unique frameworks for understanding our world, other kinds of epistemic virtues can come into play. These can include the productivity of theories in relation to practical and political objectives. An example might help here. Most commonly we understand our biology as fixing a division into two sexes, male and female. The work of feminist biologists, however, has made clear to us that these are not the only options available. It is equally possible to have a much wider range of categories, given the variety and cross over of the multiple markers, which we take to determine sex. As society has constructed an oppressive and constraining set of gendered markers onto the biological divisions of sex, it would then be perfectly legitimate for feminist theorists to chose a biological theory which did away with binary divisions into male and female.

Contemporary feminist epistemologists are also making a particular kind of intervention into the debates surrounding objectivity and rationality. They challenge the masculinist definitions to which I drew attention in the first section, not simply by resisting the definitions of femininity which were articulated there, and insisting that women can be as rational and objective as men, though they have done this; not by rejecting the value of rationality and celebrating the sensuous particularity associated with women and the distinct and valuable knowledge which it yields, though they have also done this; but by reconceptualising our notions of objectivity and rationality and consequently our conceptions of the desirable epistemic virtues. In this context it is worth paying some detailed attention to two of the best known feminist epistemologists currently writing: Sandra Harding and Helen Longino.

Longino; Local Epistemologies

Helen Longino (1990, 1993, 1997), while recognising the force of the feminist critiques, outlined above, shares with many traditional epistemologists a desire to articulate public and non arbitrary criteria for the assessment of theories, to avoid a collapse into subjectivism and relativism. In common with other post positivist philosophers of science she accepts the under determination of theory by data, and the need to supplement criteria of empirical adequacy with additional epistemological criteria. She differs from Kuhn, however, in claiming that these additional criteria will not necessarily be held in common, even within all scientific knowledge collecting communities.

Longino's first move is to insist that knowledge collection is a communal and not an individualist enterprise. The legitimacy of the knowledge collected then depends, in part, on the nature and structure of the community from which it derives, and the procedures which that community has undertaken for its assessment. Crucially this community must be diverse, so that a variety of voices and perspectives are represented. It must have structural features "to ensure the effectiveness of the critical discourse taking place within it." These include: "the provision of venues for the articulation of criticism, uptake (rather than mere toleration) of criticism, public standards to which discursive interactions are referenced, equality of intellectual authority for all (qualified) members of the community." (Longino 1997, 28-29.) Such communities, in conditions of interactive dialogue agree procedures of transformative criticism for the production of knowledge, agreeing criteria by means of which differences are to be arbitrated and theories to be assessed. These provide the public and non-arbitrary standards needed to ensure objectivity. The criteria, however, only have legitimacy for the community from which they have derived and others which share common objectives with it. This is why, for Longino, justificatory epistemology remains local. Other communities may have different objectives and evaluate theories differently. There will therefore be "a plurality of models and theories, rather than a single account that captures all facets of reality" (Longino 1997, 34). Insistence on objectivity does not, therefore, return us to one true story of the world.

In some of her papers Longino (1994 and 1997) discusses the criteria which have recommended themselves to feminist scientists, contrasting them to those put forward by Kuhn. In addition to empirical adequacy she lists novelty, ontological heterogeneity, complexity or mutuality of interaction, applicability to human needs, broadly distributed empowerment. Such a list combines political and ethical values with what appear to be more conventional epistemic ones to form a holistic framework of assessment. They are anchored in a community of feminist scientists engaged in a project of understanding the world in a way that makes gender visible and ending hierarchical power relations between men and women. The justifications in which they are used will not, however, be recognised by those who fail to share these objectives.

On Longino's account there are two bases of assessment. One is the assessment of communities from which the knowledge has derived and the procedures which have been followed whereby it has been accepted as legitimate knowledge. If the communities are not diverse or the debates have been conducted without equality of intellectual authority then the validity of the knowledge is called into question. In this way attention to the context of discovery becomes part of the context of justification. Of course none of our epistemic communities conform to her conditions of diversity and equality of intellectual authority, but in so far as they fall short then the knowledge they produce is questionable. The second basis of assessment looks more traditional. Competing knowledge claims are assessed in terms of criteria which the community has itself agreed. The difference here from traditional epistemology is that these criteria will remain 'local'. They would not necessarily be accepted by other knowledge gathering communities, although it appears that empirical adequacy in some form will appear across all communities. As a result the epistemic virtues that she identifies as being characteristic of feminist epistemic communities are distinct from those articulated by Kuhn as distinctive of most contemporary scientific inquiry.

There is some tension within Longino's theory between the two bases of assessment, for the very conditions, which promote consensus of principles of assessment, militate against diversity and vice versa. This is not necessarily a problem, because, for her, we need to accommodate both sides of this tension if we are to produce good and reliable knowledge. There is, however, a further issue, which seems more problematic. For, in the absence of agreed public standards, there is no possibility of normative engagement across difference. If a traditional scientist queried why he should be interested in the range of epistemic virtues espoused by some of the feminist scientists, the answer would seem to be that there would be no reason, unless they can identify objectives which they both share. Equally, however, the feminist scientists have no reason to attend to his list of virtues. They are all of the same standing.

There are however problems with this view. Feminists engaged in critiquing traditional disciplines did not necessarily share objectives and criteria of assessment with those they critiqued, but nonetheless they took their critiques to challenge and in some cases discredit the masculine accounts. The discrediting here was intended as discrediting per se and not just from the perspective of the feminist community. There are also many examples of encounters between, for example, women who are very differently situated, who would not be able to agree on principles of assessment but where it seems clear that some process of rational negotiation takes place. I will return to these points below.

Harding: the privilege of marginality.

A different approach to epistemological justification is found in the recent work of Sandra Harding (1991 and 1993). Harding starts from a position in which there are dominant and subjugated knowledges. She therefore accepts that there are no ideal epistemic communities of the kind which Longino recommends. She nonetheless provides us with a model of progressive epistemological engagement without relying on the kind of consensus, on which, for Longino, objectivity seems to depend.

Harding insists that our epistemological practices require 'strong reflexivity'. Our evaluation of knowledge claims requires us to reflect on the situation of knowers, and their entitlement, given their situation, to make knowledge claims. For Harding the achievement of strong objectivity was interdependent with the privileging of the epistemic position of marginal perspectives. Here attention to those who have been marginal to the production of knowledge claims makes evident the features of the positionality of those who produced them, which are crucial to assessing legitimacy. She urges us to start out theorising from marginal lives, in an attempt to address the reality of the intersection between knowledge and power. For Harding narratives from marginal lives, when counterpoised to dominant narratives, serve to expose the assumptions and exclusions in these latter, required to bring about their transformations.

Harding frequently writes as if the basis of marginal privilege lies in its making visible data which those in dominant positions did not have available to them, and which their theories could not accommodate. As such, attention to marginality would be a route to testing the empirical adequacy of theories. This in itself is, of course, not an uncontested matter, if only because the marginal lives and experiences themselves require interpretations which can be contested. But once we have given up thoughts of an unmediated access to reality this is true of all tests for empirical adequacy. There are other critical tasks for which marginality bestows a privileged position. For the margins are the places from which background assumptions, the sets of prejudgments in terms of which the dominant group structures the world can become visible. The prejudices[1], background assumptions and metaphorical associations informing particular positions and knowledges are frequently unavailable to those working within them, whose modes of characterising the world often appear as transparent reflections of the way the world is. This can provide an impression of the naturalness and inevitability of patterns of thought which are contingent and situated. It requires the intervention of differently situated viewers to unsettle the transparency and reveal the contingency of the production.

The privilege which Harding attaches to marginal perspectives does not suppose any homogeneity amongst marginal groups or the supposition that collectively they will produce a single alternative set of knowledges with which to replace those which are being critiqued. The achievement of 'strong objectivity' is a process, not a point at which theorising could rest. Marginal privilege is anchored to the critical moment of theorising. In the search for explanatory narratives which incorporate perspectives which were previously marginal new theories get produced which have to be subject to critique from their own marginalities, in a progressive project without closure or finitude.

There are a number of questions which a consideration of Harding raises. Firstly there is the issue that the division between margin and centre is not fixed (Bat-Ami Bar On, 1993), it is rather a fluid and contested one and for some writers this makes it not possible to conceive of our epistemological projects in the terms which Harding suggests. Secondly, she focuses her account on the perspectives which our epistemological projects need to address, urging men to pay attention to the lives of women and white women to start their theorising from the lives of excolonised women. This is linked to her demand for strong reflexivity, for attention to marginal positions enables us to become aware of the salient features of our own position which require critical attention; (e.g. whiteness or heterosexual practices.) It assumes, however, that perspectives of others are in some way available to us, and can encourage a picture of a subject who could somehow chose the most appropriate perspectives to adopt to advance their knowledge. This has the advantage of resisting a picture in which different perspectives are closed and self contained and only available to those who share the same life experiences. But it has problems. Naomi Scheman highlights the danger of appealing to the "experiences of people of color to provide the raw material for a more adequate theory, which it would remain the prerogative of people like me to create and authorise" (Scheman 1993). Recognising the perspectivity of knowledge also requires us to recognise the defeasible privilege of those occupying the situations to which it is tied, in its articulation. Consequently the progressive epistemological project, which Harding outlines, has to be conceived of not as individual subjects testing the validity of their theories against a marginality which they adopt at will, but as a project of epistemic communities, consequent on the opportunities for marginal voices to be heard. The need for the process of strong objectivity to be reflected in the constitution of our epistemic communities links Harding's account here with Longino's insistence that epistemic communities be constituted of diverse voices.

Here however there is a key difference between Harding and Longino. Faced with diverse voices Longino insists that consensual principles for negotiating them are required if rational justification of knowledge claims is to occur. Harding, however, makes no such assumption. The justification of certain claims, (critical ones), for Harding, seems to consist simply in pointing to their origin in marginal voices. The justification consists in pointing to their origins. But this does not seem satisfactory. It does not seem sufficient to accept marginal critiques just because they are marginal. For one reason these critiques may be in opposition to one another, given the fact that the margins is not a shared space. Secondly, as we noted above, subjugated accounts are not innocent. They themselves require interpretation and interrogation.

V RATIONAL NEGOTIATION ACROSS DIFFERENCE

Thus far the abandonment of transcendent criteria of epistemological assessment and the recognition of the situatedness of knowledge seems to have left us with a number of options. One is to abandon the project of justification and simply accept an unnegotiable pluralism in our knowledge claims. A second is to restrict normative assessment to contexts of consensual agreement. The third is to allocate epistemic privilege on the basis of material and social position. Are there any others?

The attention of many feminist writers in the past decade has been devoted to the issue of understanding and negotiating across difference (Strickland 1993 and 1994; Seller 1994; Whitford 1996; Narayan 1997). If the situatedness of knowledge is not to just lead to a standoff, then we must be able to think together experiences which are discrepant, and recognise that they bear some kind of relation to one another, so that attention to one perspective impacts on and modifies another. If attention to this process is to remain within the domain of epistemology then this process of engagement across difference must be more than a brutely causal one. It must not simply be the case that engagement with the perspectives of other modifies causally my world view, or that diversifying the epistemic community modifies causally the kind of knowledge claims which it makes. Such brutely causal transformations may occur but on their own they do not contribute to the justificatory task. What must happen instead is that the encounter with difference must enable recognition that certain kinds of modifications of view are rationally required. The difficulty comes in articulating this process of rational transformation in a way that does not presuppose transcendent or consensual principles of assessment. What has become evident from the work of several writers is that the process of understanding the perspectives of others is simultaneously a process of rational assessment. It is not possible to engage without also making judgements. Such judgements emerge out of the encounter and are not restricted to adopting any one of the perspectives within it. I have not understood other perspectives if I simply attempt to understand the world of others from within my own. But neither is it possible to simply switch perspectives, like jumping into another box and see the world from there.

To understand the perspective of another requires engaging with the way our shared world appears from another perspective within it, and the appearances of the world here includes the salience and significance it carries. The way the world appears is tied up with whole ways of life and sets of practices. To access it requires entering sufficiently into a way of life to recognise the appropriateness of a way of characterising the world. It is not easily available from elsewhere and there are limits to all of our capacities for such understanding. It is always partial. We cannot leave our own perspective behind, but bring that of another into play beside it. We recognise that the different viewpoint has an impact on our own. Engagement makes us assess the possibility and plausibility of seeing the world that way. The resulting judgements are the rational outcome of the encounter (Lennon 1997; Gadamer 1975; Strickland 1993).

To defend the possibility of rational assessment in the absence of criteria of assessment enables us to rearticulate the role of marginal perspectives to which Harding draws our attention. It is not simply that such perspectives deliver epistemically privileged critiques just because they are marginal. It is rather that engagement with such marginality brings into view considerations which impact on the dominant views and require rationally their modification. The force of such requirements, however, cannot be seen without engagement with these marginal perspectives. It is not something that can be shown by lining up sets of propositions whose relations of mutual entailment are visible from anywhere.

VI CONCLUSIONS

Feminist epistemology is not a single and unified theory. It has developed from the early work done by feminists in a number of disciplines highlighting the masculinity of the knowledge which had been produced. From this work came a recognition of the situatedness and partiality of our knowledge claims, a recognition which emerged from many sources and not just feminist ones. Subsequent work has been anxious to explore the epistemological consequences of such a recognition. Much of this has been devoted to rethinking conceptions of objectivity and rationality in ways that is respectful of a world to which our knowledge is answerable and accountable to a diverse feminist community.

Kathleen Lennon
University of Hull

NOTES

[1] 'prejudices' is here used non-pejoratively, see Gadamer 1976, p. 9.

REFERENCES

Alcoff, L. and E. Potter (ed.): 1993, *Feminist Epistemologies*, Routledge, London and N Y

Bar On, Bat–Ami: 1993, 'Marginality and Epistemic Privilege', in Alcoff and Potter 1993.

Belenky, M., B. M. Clinchy, N. R. Goldberger and J. M. Tarule: 1986, *Womens Ways of Knowing; The Development of Self, Voice and Mind*, Basic Books, New York.

Bleier, R.:1983, *Science and Gender*, Pergamon, Elmsford, New York.

Bleier, R. (ed.): 1988, *Feminist Approaches to Science*, Pergamon, New York.

Code, L.: 1991, *What Can She Know: Feminist Theory and the Construction of Knowledge*, Cornell University Press, Ithaca.

Dalmiya, V. and L. Alcoff: 1993, 'Are 'Old Wives Tales' Justified?', in Alcoff and Potter (eds.).

Feyerabend, P.: 1975, *Against Method*, Verso, London.

Fox-Keller, E.: 1985, *Reflections on Gender and Science*, Yale Univeristy Press, New Haven CT.

Fox-Keller, E.: 1992, *Secrets of Life, Secrets of Death: Essays on Language, Gender and Science*, Routledge, New York and London.

Fox-Keller, E . and H. Longino (eds.): 1996, *Feminism and Science*, Oxford University Press, Oxford, New York.

Gadamer, H.G.: 1975, *Truth and Method*, Sheed and Ward, London.

Gadamer, H.G.: 1976, *Philosophical Hermeneutics*, University of California Press, Berkeley.

Gilligan, C.: 1982, *In a Different Voice: Psychological Theory and Women's Development*, Harvard University Press, Cambridge Mass.

Gonzalez, A. S.: 2000, *Practical Knowledge*, Ph.D. Thesis, University of Hull.

Haraway, D.: 1978, 'Animal Sociology and a Natural Economy of the Body Politic,' *Signs: Journal of Women in Culture and Society* 4/1, reprinted in E. Fox Keller and H. Longino (ed.) 1996, *Feminism and Science*, Oxford University Press, Oxford.

Haraway, D.: 1989, *Primate Visions: Gender, Race and Nature in the World of Modern Science*, Routledge, New York.

Haraway, D.: 1991a, *Simians, Cyborgs and Women: The Reinvention of Nature*, Free Association, London.

Haraway, D.: 1991b, 'Cyborgs at Large; Interview with Donna Haraway', in C. Penley and A. Ross (eds.), *Technoculture*, University of Minnesota Press, Minneapolis.

Harding, S.: 1986, *The Science Question in Feminism*, Cornell University Press, Ithaca and London.

Harding, S.: 1991, *Whose Science? Whose Knowledge?*, Open Univerisity Press, Milton Keynes.

Harding, S. (ed.): 1993, *The "Racial" Economy of Science: Toward a Democratic Future*, Indiana University Press, Bloomington and Indianapolis.

Hartsock, N.: 1983, 'The Feminist Standpoint: Developing the Ground for a Specifically Feminist Historical Materialism', in S. Harding and M. Hintikka (eds.), *Discovering Reality*, Reidel, Dordrecht.

Hegel, G.W.F.: 1977, *Phenomenology of Spirit*, Oxford University Press, Oxford and New York.

Jagger, A.: 1983, *Feminist Politics and Human Nature*, Harvester, Sussex.

Kuhn, T.: 1962, *The Structure of Scientific Revolutions*, University of Chicago Press, Chicago.

Lennon, K.: 1995, 'Gender and Knowledge', *Journal of Gender Studies*, Vol **4** No2, 33-145.

Lennon, K.: 1997 'Reply to Helen Longino', *PASS* **LXXI**, 37-55.

Lennon, K. and M. Whitford (eds.): 1994, *Knowing the Difference, Feminist Perspectives in Epistemology*, Routledge, London and New York.

Lukacs, G.: 1971, *History and Class Consciousness*, MIT Press, Cambridge Mass.

Lewontin, R., S. Rose and L. Kamin: 1984, *Not in Our Genes: Biology, Ideology and Human Nature*, Penguin, London and New York.

Lloyd, G.: 1984, *The Man of Reason: 'Male' and 'Female' in Western Philosophy*, Methven, London.

Longino, H.: 1990, *Science as Social Knowledge*, Princeton University Press, New Jersey.

Longino, H.: 1993, 'Subjects, Power and Knowledge: Description and Prescription in Feminist Philosophies of Science', in Alcoff and Potter 1993.

Longino, H.: 1994, 'In Search of Feminist Epistemology', *The Monist* **77**, no 4, 472-485.

Longino, H.: 1997, 'Feminist Epistemology as Local Epistemology', *PASS*, **LXXI**, 19-37.

Martin, E.: 1991, 'The Egg and Sperm', *Signs; Journal of Women in Culture and Society* **16/3**, reprinted in E. Fox Keller and H. Longino (eds.), 1996, *Feminism and Science*, Oxford University Press, Oxford.

Narayan, U.: 1997, *Dislocating Culture*, Routledge, London and New York.

Nelson, L. H.: 1990, *Who Knows: From Quine to a Feminist Empiricism*, Temple University Press, Philadelphia.

Nelson, L. H.: 1993, 'A Question of Evidence', *Hypatia*, vol **8**, no 2.

Nicholson, L. J.: 1990, *Feminism / Postmodernism*, Routledge, New York and London.

Oudshoorn, N.: 1994, *Beyond the Natural Body*, Routledge, London and New York.

Parsons, T.: 1951, *The Social System*, The Free Press, Glencoe Ill.

Pitt, D.: 1988, 'Women and Heart Disease', paper to 1998 Women's Studies Network Conference, Hull, UK.

Quine, W.: 1953, 'Two Dogma's of Empiricism', in *From a Logical Point of View*, Harvard University Press, Cambridge Mass.

Scheman, N.: 1993, *Engenderings*, Routledge, London and New York.

Seller, A.: 1994, 'Should the Feminist Philosopher Stay at Home', in Lennon and Whitford 1994.

Smith, D.: 1987, *The Everyday World as Problematic*, Open University Press, Milton Keynes.

Spelman, E.: 1988, *Inessential Woman: Problems of Exclusion in Feminist Thought*, Beacon Press, Boston.

Strickland, S.: 1993, *Objectivity Perspectivity and Difference*, Ph.D. University of Hull 1994.

Stickland, S.: 1994, 'Feminism Postmodernism and Difference', in Lennon and Whitford 1994.

Whitford, M.: 1996, 'Doing Feminist Research – Making Links', *Womens Philosophy Review* Issue no.**16**, 33-41.

INDEX OF NAMES

SUBJECT INDEX