Philosophy and the Cognitive Sciences

Proceedings
of the
16th International Wittgenstein Symposium

15-22 August 1993
Kirchberg am Wechsel (Austria)

EDITORS

Roberto Casati, Barry Smith and Graham White

Vienna 1994
Verlag Hölder-Pichler-Tempsky

Wir danken
dem Bundesministerium für Wissenschaft und Forschung in Wien
und dem Kulturamt der Landesregierung Niederösterreich
für die Förderung dieses Werkes

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Philosophy and the cognitive sciences : proceedings of the 16th International Wittgenstein Symposium, 15 - 22 August 1993, Kirchberg am Wechsel (Austria) / ed.: Roberto Casati ... - Wien : Hölder-Pichler-Tempsky, 1994
(Schriftenreihe der Wittgenstein-Gesellschaft ; Bd. 21)
ISBN 3-209-01747-6
NE: Casati, Roberto [Hrsg.]; Internationales Wittgenstein-Symposium <16, 1993, Kirchberg, Wechsel>; Österreichische Ludwig-Wittgenstein-Gesellschaft: Schriftenreihe der Wittgenstein-Gesellschaft

ISBN 3-209-01747-6

Umschlaggestaltung: Herwig Kempinger
Gesamtherstellung: G.G. Buchbinderei, Hollabrunn
TEX typesetting and LATEX style: Graham White
TEX is a trademark of the American Mathematical Society

Table of Contents

Philosophy and the Foundations of Cognitive Science

Wittgenstein and Cognitive Science

Content and Object

Logic and Foundations

Language and Linguistics

Ontology and Mereology

Introduction

The Sixteenth International Wittgenstein Symposium, which was held in Kirchberg am Wechsel, Lower Austria, from 15 to 21 August 1993, was devoted to the topic of Philosophy and the Cognitive Sciences. Special sections were devoted to current developments in such fields as artificial intelligence research, cognitive linguistics and Wittgenstein's contribution to philosophical psychology, as well as to the historical roots of the cognitive sciences in the work of Ernst Mach and Franz Brentano. A volume of Preprints containing the texts of 110 papers which were read at the meeting has been published already as Volume I of the newly established series "Contributions of the Austrian Ludwig Wittgenstein Society"; the table of contents of this volume is reproduced as an appendix below. Both collections demonstrate the enormous diversity of systematic and historical work which is to be observed on both sides of the Atlantic in this still nascent field.

The three editors of these Proceedings have for the last three years been engaged on a project of the Swiss National Foundation on the topic of Formal-Ontological Foundations of Contemporary Artificial Intelligence Research. The preparation of this volume and the organisation of the Kirchberg conference would not have been possible without the support of the SNF. The editors would like to thank also the Austrian Federal Ministry of Science and Research, the Peter Kaiser Foundation (Vaduz), and Prince Hans Adam II of Liechtenstein for their generosity in supporting the conference, as well as the Cultural Department of the Government of Lower Austria for its unfailing support of the Ludwig Wittgenstein Society. Finally, for invaluable organisational help, we should like to thank Klaus Puhl of the University of Graz.

The Editors

Triesen, July 1994

The Problem of Consciousness

John R. Searle

The most important scientific discovery of the present era will come when someone – or some group – discovers the answer to the following question: How exactly do neurobiological processes in the brain cause consciousness? This is the most important question facing us in the biological sciences, yet it is frequently evaded, and frequently misunderstood when not evaded. In order to clear the way for an understanding of this problem. I am going to begin to answer four questions: 1. What is consciousness? 2. What is the relation of consciousness to the brain? 3. What are some of the features that an empirical theory of consciousness should try to explain? 4. What are some common mistakes to avoid?

1 What is Consciousness?

Like most words, 'consciousness' does not admit of a definition in terms of genus and differentia or necessary and sufficient conditions. Nonetheless, it is important to say exactly what we are talking about because the phenomenon of consciousness that we are interested in needs to be distinguished from certain other phenomena such as attention, knowledge, and self-consciousness. By 'consciousness' I simply mean those subjective states of sentience or awareness that begin when one awakes in the morning from a dreamless sleep and continue throughout the day until one goes to sleep at night or falls into a coma, or dies, or otherwise becomes, as one would say, 'unconscious'.

Above all, consciousness is a biological phenomenon. We should think of consciousness as part of our ordinary biological history, along with digestion, growth, mitosis and meiosis. However, though consciousness is a biological phenomenon, it has some important features that other biological phenomena do not have. The most important of these is what I have called its 'subjectivity'. There is a sense in which each person's consciousness is private to that person, a sense in which he is related to his pains, tickles, itches, thoughts and feelings in a way that is quite unlike the way that others are related to those pains, tickles, itches, thoughts and feelings. This phenomenon can be described in various ways. It is sometimes described as that

An earlier version of this article has appeared in the publications of the CIBA Foundation. The theses advanced in this paper are presented in more detail and with more supporting argument in Searle, J.R. *The Rediscovery of the Mind*, Cambridge MA: MIT Press 1992.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 1-10

feature of consciousness by way of which there is something that it's like or something that it feels like to be in a certain conscious state. If somebody asks me what it feels like to give a lecture in front of a large audience I can answer that question. But if somebody asks what it feels like to be a shingle or a stone, there is no answer to that question because shingles and stones are not conscious. The point is also put by saying that conscious states have a certain qualitative character; the states in question are sometimes described as 'qualia'.

In spite of its etymology, consciousness should not be confused with knowledge, it should not be confused with attention, and it should not be confused with self-consciousness. I will consider each of these confusions in turn.

Many states of consciousness have little or nothing to do with knowledge. Conscious states of undirected anxiety or nervousness, for example, have no essential connection with knowledge.

Consciousness should not be confused with attention. Within one's field of consciousness there are certain elements that are at the focus of one's attention and certain others that are at the periphery of consciousness. It is important to emphasize this distinction because 'to be conscious of' is sometimes used to mean 'to pay attention to'. But the sense of consciousness that we are discussing here allows for the possibility that there are many things on the periphery of one's consciousness – for example, a slight headache I now feel or the feeling of the shirt collar against my neck – which are not at the centre of one's attention. I will have more to say about the distinction between the center and the periphery of consciousness in Section 3.

Finally, consciousness should not be confused with self-consciousness. There are indeed certain types of animals, such as humans, that are capable of extremely complicated forms of self-referential consciousness which would normally be described as self-consciousness. For example, I think conscious feelings of shame require that the agent be conscious of himself or herself. But seeing an object or hearing a sound, for example, does not require self-consciousness. And it is not generally the case that all conscious states are also self-conscious.

2 What are the Relations between Consciousness and the Brain?

This question is the famous 'mind-body problem'. Though it has a long and sordid history in both philosophy and science, I think, in broad outline at least, it has a rather simple solution. Here it is: Conscious states are caused by lower level neurobiological processes in the brain and are themselves higher level features of the brain. The key notions here are those of *cause* and *feature*. As far as we know anything about how the world works, variable rates of neuron firings in different neuronal architectures cause all

the enormous variety of our conscious life. All the stimuli we receive from the external world are converted by the nervous system into one medium, namely, variable rates of neuron firings at synapses. And equally remarkably, these variable rates of neuron firings cause all of the colour and variety of our conscious life. The smell of the flower, the sound of the symphony, the thoughts of theorems in Euclidian geometry – all are caused by lower level biological processes in the brain; and as far as we know, the crucial functional elements are neurons and synapses.

Of course, like any causal hypothesis this one is tentative. It might turn out that we have overestimated the importance of the neuron and the synapse. Perhaps the functional unit is a column or a whole array of neurons, but the crucial point I am trying to make now is that we are looking for causal relationships. The first step in the solution of the mind-body problem is: brain processes *cause* conscious processes.

This leaves us with the question, what is the ontology, what is the form of existence, of these conscious processes? More pointedly, does the claim that there is a causal relation between brain and consciousness commit us to a dualism of 'physical' things and 'mental' things? The answer is a definite no. Brain processes cause consciousness but the consciousness they cause is not some extra substance or entity. It is just a higher level feature of the whole system. The two crucial relationships between consciousness and the brain, then, can be summarized as follows: lower level neuronal processes in the brain cause consciousness and consciousness is simply a higher level feature of the system that is made up of the lower level neuronal elements.

There are many examples in nature where a higher level feature of a system is caused by lower level elements of that system, even though the feature is a feature of the system made up of those elements. Think of the liquidity of water or the transparency of glass or the solidity of a table, for example. Of course, like all analogies these analogies are imperfect and inadequate in various ways. But the important thing that I am trying to get across is this: there is no metaphysical obstacle, no logical obstacle, to claiming that the relationship between brain and consciousness is one of causation and at the same time claiming that consciousness is just a feature of the brain. Lower level elements of a system can cause higher level features of that system, even though those features are features of a system made up of the lower level elements. Notice, for example, that just as one cannot reach into a glass of water and pick out a molecule and say 'This one is wet', so, one cannot point to a single synapse or neuron in the brain and say 'This one is thinking about my grandmother'. As far as we know anything about it, thoughts about grandmothers occur at a much higher level than that of the single neuron or synapse, just as liquidity occurs at a much higher level than that of single molecules.

Of all the theses that I am advancing in this article, this one arouses the most opposition. I am puzzled as to why there should be so much opposition, so I want to clarify a bit further what the issues are: First, I want

to argue that we simply know as a matter of fact that brain processes cause conscious states. We don't know the details about how it works and it may well be a long time before we understand the details involved. Furthermore, it seems to me an understanding of how exactly brain processes cause conscious states may require a revolution in neurobiology. Given our present explanatory apparatus, it is not at all obvious how, within that apparatus, we can account for the causal character of the relation between neuron firings and conscious states. But, at present, from the fact that we do not know *how* it occurs, it does not follow that we do not know *that* it occurs. Many people who object to my solution (or dissolution) of the mind-body problem, object on the grounds that we have no idea how neurobiological processes could cause conscious phenomena. But that does not seem to me a conceptual or logical problem. That is an empirical/theoretical issue for the biological sciences. The problem is to figure out exactly how the system works to produce consciousness, and since we know that in fact it does produce consciousness, we have good reason to suppose that are specific neurobiological mechanisms by way of which it works.

There are certain philosophical moods we sometimes get into when it seems absolutely astounding that consciousness could be produced by electro-biochemical processes, and it seems almost impossible that we would ever be able to explain it in neurobiological terms. Whenever we get in such moods, however, it is important to remind ourselves that similar mysteries have occurred before in science. A century ago it seemed extremely mysterious, puzzling, and to some people metaphysically impossible that life should be accounted for in terms of mechanical, biological, chemical processes. But now we know that we can give such an account, and the problem of how life arises from biochemistry has been solved to the point that we find it difficult to recover, difficult to understand why it seemed such an impossibility at one time. Earlier still, electromagnetism seemed mysterious. On a Newtonian conception of the universe there seemed to be no place for the phenomenon of electromagnetism. But with the development of the theory of electromagnetism, the metaphysical worry dissolved. I believe that we are having a similar problem about consciousness now. But once we recognize the fact that conscious states are caused by neurobiological processes, we automatically convert the issue into one for theoretical scientific investigation. We have removed it from the realm of philosophical or metaphysical impossibility.

3 Some Features of Consciousness

The next step in our discussion is to list some (not all) of the essential features of consciousness which an empirical theory of the brain should be able to explain.

Subjectivity

As I mentioned earlier, this is the most important feature. A theory of consciousness needs to explain how a set of neurobiological processes can cause a system to be in a subjective state of sentience or awareness. This phenomenon is unlike anything else in biology, and in a sense it is one of the most amazing features of nature. We resist accepting subjectivity as a ground floor, irreducible phenomenon of nature because, since the seventeenth century, we have come to believe that science must be objective. But this involves a pun on the notion of objectivity. We are confusing the *epistemic* objectivity of scientific investigation with the *ontological* objectivity of the typical subject matter in science in disciplines such as physics and chemistry. Since science aims at objectivity in the epistemic sense that we seek truths that are not dependent on the particular point of view of this or that investigator, it has been tempting to conclude that the reality investigated by science must be objective in the sense of existing independently of the experiences in the human individual. But this last feature, ontological objectivity, is not an essential trait of science. If science is supposed to give an account of how the world works and if subjective states of consciousness are part of the world, then we should seek an (epistemically) objective account of an (ontologically) subjective reality, the reality of subjective states of consciousness. What I am arguing here is that we can have an epistemically objective science of a domain that is ontologically subjective.

Unity

It is important to recognize that in non-pathological forms of consciousness we never just have, for example, a pain in the elbow, a feeling of warmth, or an experience of seeing something red, but we have them all occurring simultaneously as part of one unified conscious experience. Kant called this feature 'the transcendental unity of apperception'. Recently, in neurobiology it has been called 'the binding problem'. There are at least two aspects to this unity that require special mention. First, at any given instant all of our experiences are unified into a single conscious field. Second, the organization of our consciousness extends over more than simple instants. So, for example, if I begin speaking a sentence, I have to maintain in some sense at least an iconic memory of the beginning of the sentence so that I know what I am saying by the time I get to the end of the sentence.

Intentionality

'Intentionality' is the name that philosophers and psychologists give to that feature of many of our mental states by which they are directed at, or about states of affairs in the world. If I have a belief or a desire or a fear, there must always be some content to my belief, desire or fear. It must be about something even if the something it is about does not exist or is a hallucination. Even in cases when I am radically mistaken, there must be some mental content which purports to make reference to the world. Not all conscious

states have intentionality in this sense. For example, there are states of anxiety or depression where one is not anxious or depressed about anything in particular but just is in a bad mood. That is not an intentional state. But if one is depressed about a forthcoming event, that is an intentional state because it is directed at something beyond itself.

There is a conceptual connection between consciousness and intentionality in the following respect. Though many, indeed most, of our intentional states at any given point are unconscious, nonetheless, in order for an unconscious intentional state to be genuinely an intentional state it must be accessible in principle to consciousness. It must be the sort of thing that could be conscious even if it, in fact, is blocked by repression, brain lesion, or sheer forgetfulness.

The Distinction Between the Center and the Periphery of Consciousness

At any given moment of non-pathological consciousness I have what might be called a field of consciousness. Within that field I normally pay attention to some things and not to others. So, for example, right now I am paying attention to the problem of describing consciousness but very little attention to the feeling of the shirt on my back or the tightness of my shoes. It is sometimes said that I am unconscious of these. But that is a mistake. The proof that they are a part of my conscious field is that I can at any moment shift my attention to them. But in order for me to shift my attention to them, there must be something there which I was previously not paying attention to which I am now paying attention to.

The Gestalt Structure of Conscious Experience

Within the field of consciousness our experiences are characteristically structured in a way that goes beyond the structure of the actual stimulus. This was one of the most profound discoveries of the Gestalt psychologists. It is most obvious in the case of vision, but the phenomenon is quite general and extends beyond vision. For example, the sketchy lines drawn in Figure 1 do not physically resemble a human face.

If we actually saw someone on the street that looked like that, we would be inclined to call an ambulance. The disposition of the brain to structure degenerate stimuli into certain structured forms is so powerful that we will naturally tend to see this as a human face. Furthermore, not only do we have our conscious experiences in certain structures, but we tend also to have them as figures against backgrounds. Again, this is most obvious in the case of vision. Thus, when I look at the figure I see it against the background of the page. I see the page against the background of the table. I see the table against the background of the floor, and I see the floor against the background of the room, until we eventually reach the horizon of my visual consciousness.

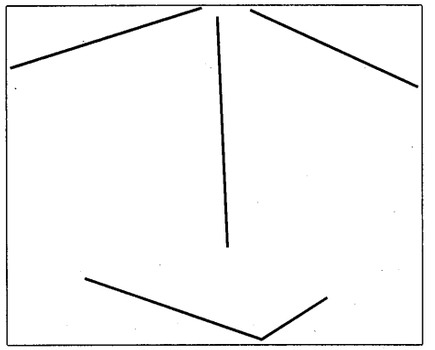


Figure 1

The Aspect of Familiarity

It is a characteristic feature of non-pathological states of consciousness that they come to us with what I will call the 'aspect of familiarity'. In order for me to see the objects in front of me as, for example, houses, chairs, people, tables, I have to have a prior possession of the categories of houses, chairs, people, tables. But that means that I will assimilate my experiences into a set of categories which are more or less familiar to me. When I am in an extremely strange environment, in a jungle village, for example, and the houses, people and foliage look very exotic to me, I still perceive that as a house, that as a person, that as clothing, that as a tree or a bush. The aspect of familiarity is thus a scalar phenomenon. There can be greater or lesser degrees of familiarity. But it is important to see that non-pathological forms of consciousness come to us under the aspect of familiarity. Again, one way to consider this is to look at the pathological cases. In Capgras's syndrome, the patients are unable to acknowledge familiar people in their environment as the people they actually are. They think the spouse is not really their spouse but is an imposter, etc. This is a case of a breakdown in one aspect of familiarity. In non-pathological cases it is extremely difficult to break with the aspect of familiarity. Surrealist painters try to do it. But even in the surrealist painting, the three-headed woman is still a woman, and the drooping watch is still a watch.

Mood

Part of every normal conscious experience is the mood that pervades the experience. It need not be a mood that has a particular name to it, like depression or elation; but there is always what one might call a flavour or tone to any normal set of conscious states. So, for example, at present I am not especially depressed and I am not especially ecstatic, nor indeed, am I what one would call simply 'blah'. Nonetheless, there is a certain mood

to my present experiences. Mood is probably more easily explainable in biochemical terms than several of the features I have mentioned. We may be able to control, for example, pathological forms of depression by mood-altering drugs.

Boundary conditions

All of my non-pathological states of consciousness come to me with a certain sense of what one might call their 'situatedness'. Though I am not thinking about it, and though it is not part of the field of my consciousness, I nonetheless know what year it is, what place I am in, what time of day it is, the season of the year it is, and usually even what month it is. All of these are the boundary conditions or the situatedness of nonpathological conscious states. Again, one can become aware of the pervasiveness of this phenomenon when it is absent. So, for example, as one gets older there is a certain feeling of vertigo that comes over one when one loses a sense of what time of year it is or what month it is. The point I am making now is that conscious states are situated and they are experienced as situated even though the details of the situation need not be part of the content of the conscious states.

4 Some Common Mistakes about Consciousness

I would like to think that everything I have said so far is just a form of common sense. However, I have to report, from the battlefronts as it were, that the approach I am advocating to the study of consciousness is by no means universally accepted in cognitive science nor even neurobiology. Indeed, until quite recently many workers in cognitive science and neurobiology regarded the study of consciousness as somehow out of bounds for their disciplines. They thought that it was beyond the reach of science to explain why warm things feel warm to us or why red things look red to us. I think, on the contrary, that it is precisely the task of neurobiology to explain these and other questions about consciousness. Why would anyone think otherwise? Well, there are complex historical reasons, going back at least to the seventeenth century, why people thought that consciousness was not part of the material world. A kind of residual dualism prevented people from treating consciousness as a biological phenomenon like any other. However, I am not now going to attempt to trace this history. Instead I am going to point out some common mistakes that occur when people refuse to address consciousness on its own terms.

The characteristic mistake in the study of consciousness is to ignore its essential subjectivity and to try to treat it as if it were an objective third person phenomenon. Instead of recognizing that consciousness is essentially a subjective, qualitative phenomenon, many people mistakenly suppose that its essence is that of a control mechanism or a certain kind of set of dispositions to behavior or a computer program. The two most common mistakes about consciousness are to suppose that it can be analysed behavioristically

or computationally. The Turing test disposes us to make precisely these two mistakes, the mistake of behaviorism and the mistake of computationalism. It leads us to suppose that for a system to be conscious, it is both necessary and sufficient that it has the right computer program or set of programs with the right inputs and outputs. I think you have only to state this position clearly to enable you to see that it must be mistaken. A traditional objection to behaviorism was that behaviorism could not be right because a system could behave as if it were conscious without actually being conscious. There is no logical connection, no necessary connection between inner, subjective, qualitative mental states and external, publicly observable behavior. Of course, in actual fact, conscious states characteristically cause behavior. But the behavior that they cause has to be distinguished from the states themselves. The same mistake is repeated by computational accounts of consciousness. Just as behavior by itself is not sufficient for consciousness, so computational models of consciousness are not sufficient by themselves for consciousness. The computational model of consciousness stands to consciousness in the same way the computational model of anything stands to the domain being modelled. Nobody supposes that the computational model of rainstorms in London will leave us all wet. But they make the mistake of supposing that the computational model of consciousness is somehow conscious. It is the same mistake in both cases.

There is a simple demonstration that the computational model of consciousness is not sufficient for consciousness. I have given it many times before so I will not dwell on it here. Its point is simply this: *Computation is defined syntactically.* It is defined in terms of the manipulation of symbols. But the syntax by itself can never be sufficient for the sort of contents that characteristically go with conscious thoughts. Just having zeros and ones by themselves is insufficient to guarantee mental content, conscious or unconscious. This argument is sometimes called 'the Chinese room argument' because I originally illustrated the point with the example of the person who goes through the computational steps for answering questions in Chinese but does not thereby acquire any understanding of Chinese.[1] The point of the parable is clear but it is usually neglected. *Syntax by itself is not sufficient for semantic content.* In all of the attacks on the Chinese room argument, I have never seen anyone come out baldly and say they think that syntax is sufficient for semantic content.

However, I now have to say that I was conceding too much in my earlier statements of this argument. I was conceding that the computational theory of the mind was at least false. But it now seems to me that it does not reach the level of falsity because it does not have a clear sense. Here is why.

The natural sciences describe features of reality that are intrinsic to the world as it exists independently of any observers. Thus, gravitational attrac-

[1] Searle, J.R., "Minds, Brains, and Programs", *Behavioral and Brain Sciences*, 3 (1980), 417-457.

tion, photosynthesis, and electromagnetism are all subjects of the natural sciences because they describe intrinsic features of reality. But such features such as being a bathtub, being a nice day for a picnic, being a five dollar bill or being a chair, are not subjects of the natural sciences because they are not intrinsic features of reality. All the phenomena I named – bathtubs, etc. – are physical objects and as physical objects have features that are intrinsic to reality. But the feature of being a bathtub or a five dollar bill exists only relative to observers and users.

Absolutely essential, then, to understanding the nature of the natural sciences is the distinction between those features of reality that are intrinsic and those that are observer-relative. Gravitational attraction is intrinsic. Being a five dollar bill is observer-relative. Now, the really deep objection to computational theories of the mind can be stated quite clearly. Computation does not name an intrinsic feature of reality but is observer-relative and this is because computation is defined in terms of symbol manipulation, but the notion of a 'symbol' is not a notion of physics or chemistry. Something is a symbol only if it is used, treated or regarded as a symbol. The Chinese room argument showed that semantics is not intrinsic to syntax. But what this argument shows is that syntax is not intrinsic to physics. There are no purely physical properties that zeros and ones or symbols in general have that determine that they are symbols. Something is a symbol only relative to some observer, user or agent who assigns a symbolic interpretation to it. So the question, 'Is consciousness a computer program?', lacks a clear sense. If it asks, 'Can you assign a computational interpretation to those brain processes which are characteristic of consciousness?' the answer is: you can assign a computational interpretation to anything. But if the question asks, 'Is consciousness intrinsically computational?' the answer is: nothing is intrinsically computational. Computation exists only relative to some agent or observer who imposes a computational interpretation on some phenomenon. This is an obvious point. I should have seen it ten years ago but I did not.

Naturalizing Computationality

Joseph Margolis

I

In the chronicles of analytic epistemology, W.V. Quine's enormously influential paper, "Epistemology Naturalized",[1] has given impetus to a fashionable and very spare form of theorizing that is doubly anti-Fregean: first, because, in construing epistemology as a special inquiry within empirical psychology, it distances itself from Frege's well-known diatribe against psychologism; second, in construing epistemology as empirical, Quine precludes all theories that pretend to *a priori* resources or necessary and indubitable truths.

I welcome both themes, but a caveat or two may be in order. Empirical psychology may not be quite what Quine supposes it is; he himself has almost nothing to say about the relationship between the biological and cultural sources of individual psychology; and the very relevance of psychology for epistemology may take a number of diverse and disputed forms that the validity of Quine's own account must surely confront (but nowhere does). Similarly, although with very few exceptions, contemporary analytic philosophies oppose any reliance on self-evidence, privileged access to the structure of the real world, and the like, it is not always clear what Quine intends in his account of the normative, justificatory, legitimative, even transcendental questions of classical epistemology.

In fact, it is not really clear what should now be meant by the epithet "naturalized".[2] Some profess to see in Quine's account an attempt to eliminate the normative questions of canonical epistemology, without untoward results; others think Quine has given mixed signals on this score. Some think the normative can be successfully retired; others think that retiring the normative is abandoning epistemology altogether; still others think the normative can be naturalized. Some think a causal account of the conditions of knowledge obviates the need for separate normative provisions; others think that is a mistake. Some think the question can be divided; others think it cannot.[3] My own view is that these options do not quite grip on to the central issue. In all candor, the confirmation of none of these claims is supplied in Quine's essay (or elsewhere in Quine's writings) or anywhere else (as far as I can see) in the naturalistic epistemologies of our day. (We

[1] Quine 1969.

[2] For a reliable overview, which nevertheless confirms the informality I draw attention to, see Kitcher 1992.

[3] See Kim 1988, Stroud 1981, Bloor 1974, Goldman 1991.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 11-23

need not confine ourselves to Quine's view.) It may therefore not be too naive to press the question: What, precisely, makes a theory "naturalistic"?

I shall come to that in a moment; but, before I do, I want to draw attention to the fact that, on the naturalist's own view, the theory of knowledge and the theory of mind inevitably overlap, since knowledge is thought to be a psychological state of some sort. Hence, wherever the cognitive sciences are thought to be concerned to model the human mind, one may fairly suppose that, if epistemology can be naturalized, in all likelihood the computational modeling of the mind cannot be far behind. I think this is part of the reason for the recent fascination with Donald Davidson's reclamation of the supervenience theory. I shall come to that shortly as well; it is central to my argument. But my present point is that the argument on supervenience, narrowly addressed to epistemology (*pro* or *con*), may be reasonably thought to apply across the philosophical board and may be shown to bear on more than epistemology. The vision is a bold one and very grandly conceived on all sides.

The truth is that the various strategies of "naturalizing" are the converging beneficiaries of the very large failed systematic programs of early twentieth-century analytic philosophy: in particular, positivism, the unity of science program, and logical atomism. Themes centrally favored in these programs continue to attract adherents, of course. That, I take it, is the point of the "naturalizing" venture. But, as in the recovery of positivist themes (in rather different ways) in the recent work of Bas van Fraassen and Nancy Cartwright,[4] the global vision has had to give way to, for instance, small-scale studies of the logical and epistemological features of explanatory laws.

Quine's paper indicates something of what may be retired of the grander visions, but the deeper continuities remain. The most important is an uncomplicated realism that treats the world as determinately structured independently of human inquiries – hence, essentially unaltered by cognition as such – the structures of which, inquiry is directed to investigate. The effort, by and large empirical, tends to be methodologically solipsistic.

Within these very informal constraints, the marks of the "naturalizing" orientation are these:

(1) empirical discoveries tend to favor extensionalist or physicalist formulations or formulations relatively free of intensional (and intentional) complexity;

(2) philosophy tends to favor the marginalizing, near-elimination, or elimination, of second-order legitimative questions;

(3) very little in the way of rationalizing maneuvers in accord with (2) is required, or offered, on the basis of what is discovered in accord with (1); and

[4]See, for instance, the piecemeal (and tactful) reclamation of "positivist" themes in Cartwright 1983 and Van Fraassen 1989.

(4) interpretive schemata mediating between inquiring agents and investigated world are thought to function instrumentally or heuristically only or to reflect the level of admitted ignorance at particular stages of inquiry – they tend not to play a constituting role with regard to the intelligible structure of reality and in principle may be retired.

Naturalizing, then, is the piecemeal pursuit of questions across the philosophical board, loosely tethered in accord with (1)–(4).

I admit I am one of those benighted enough to believe that the naturalizing of epistemology, of mind and language, of the computational modeling of the mind, and of moral philosophy are all failed projects. I put this as baldly as I can, so that I shall not be misunderstood.

Still, I must make two concessions about all this that may mollify some. First, I admit the idea of naturalizing epistemology is not in the least incoherent or self-contradictory or paradoxical; but then I am bound to say, neither is its rejection. Second, I believe the way to defeat the naturalist is, broadly speaking, on his own grounds, so that, for instance, if he insists on a naturalized account of truth or reference or meaning, the "non-naturalist" is bound to show that these matters resist or cannot convincingly accommodate a naturalized strategy.

Philosophy proceeds here by placing dialectical bets – not usually by a *reductio* or happy paradox, though such may arise to surprise us from time to time. At the present time, much stronger bets (I say) can be mustered against the naturalizing stance than can be mustered in its favor; the best objections against naturalizing cannot be seriously disarmed; and *none* of the resources or strategies of the naturalists (except the pretense to have defeated the non-naturalists by such means) need be rejected by their opponents – but the reverse is not true. The resources of the non-naturalists, I say, are greater than those of the naturalists. There's the challenge at least.

II

I address here only the competing strategies. The arguments should be reasonably clear, once these are in place; and in any case they would require an extended labor that I cannot now afford. Furthermore, the point of my reflection is not merely to oppose naturalized epistemologies (though I am happy to do that). It is rather that I believe naturalizing epistemology is the vanguard of a very ambitious (but, in my view, completely flawed) philosophical program prepared to believe that success in epistemology augurs well for the success of the naturalizing strategy applied to narrower questions like those regarding the analysis of meaning, truth, reference, predication, facts, values, validity, and legitimation, as well as applied to such large domains as those of the philosophy of mind, the cognitive sciences, and moral philosophy. My thought is that if the inherent (metonymic) weakness of naturalized epistemology or supervenience were made clear, then pretensions of the sorts just indicated would fall fairly quickly into place and we should have the benefit of a clear picture of the full contest.

I have no difficulty, however, in conceding the promise of a program like that described in K.M. Colby's well-known "Computer Simulation of a Neurotic Process" or his later effort to model paranoia.[5] Nor do I have any difficulty with the usefulness of the Chinese Room puzzle, even admitting that Searle's objections are both worth considering in general and considerably flawed in particular.[6] Nor do I have any difficulty in admitting the canonical importance of Turing's famous "intelligence test".[7] I simply don't view any of these factors as *decisive*, one way or the other, for the resolution of the quarrel about the naturalizing strategy applied to the analysis of mind in general or of any of the cognate questions usually linked with this regarding intelligence, cognition, understanding, or anything of the sort. I grant, then, that naturalizing signifies at least:

(a) the full pertinence of empirical psychology (descriptive, causal, and explanatory) in characterizing the work of epistemology, and

(b) the total rejection of all forms of apriorism and cognitive privilege.

That, of course, is still too bland.

I must point out here that I am deliberately mingling questions regarding the naturalizing of epistemology, of mind, and of the cognitive modeling of mind. I do so to confirm that the arguments across the whole of philosophy are essentially the same in this regard, that there is a natural declension from epistemology to the theory of mind – and even to the methodology of the human sciences and the puzzles of moral philosophy. For example, Searle, who, in his Chinese puzzle case, hopes to show that the computer simulation of the human understanding of a language fails, nevertheless himself holds that human understanding may be satisfactorily explained in terms of the powers of the physical brain and the causal relations in which the brain enters.

In my opinion, Searle opposes one version of the naturalizing strategy but favors another. That is, against Allen Newell and Herbert Simon's well-known (naturalizing) claim, Searle affirms (reasonably enough) that "mental processes are [not] computational processes over formally defined elements", since the mind is intentional, has semantically pertinent "content", and purely formal or syntactically defined programs do not: "they have by themselves", Searle says, "no causal powers except the power, when instantiated, to produce the next stage of the formalism when the machine is running".[8] On the other hand, Searle claims but does not demonstrate: "*only* a machine could think, and indeed only very special kinds of machines, namely brains and machines that had the *same causal powers* as brains... Whatever else intentionality is, it is a biological phenomenon".[9]

[5]Colby 1963 and 1975.
[6]Searle 1980.
[7]Turing 1950.
[8]Searle 1980, 1981: 298–299. Cf. Newell and Simon 1976.
[9]Searle 1980, 1981: 305.

I cannot stay to quarrel with Searle about the details of his thesis. My point is rather that one strong naturalizing strategy (Newell and Simon's) might supplement the defining marks of naturalizing undertakings thus: by adding (c′), the adequacy of characterizing the mental in terms of "computational processes over formally defined elements"; whereas another (Searle's) might reject (c′) and substitute instead (c″), the adequacy of characterizing the mental in terms of the causal powers of the biological brain. I have no doubt that (c′) and (c″) – and similar formulas – do capture more satisfactorily what the naturalizing stance is all about, but I would be less than candid if I did not say at once that I think both (and similar additions) must fail.

I think Searle is right about Newell and Simon; but I suggest that Searle cannot make his own strategy convincing without explaining what the sense is in which intentional and linguistic properties and the like *can* be directly ascribed to the brain – and, therefore, adequately accounted for in specifically causal terms (nomologically, for instance). That question I regard as affecting a provisional stalemate for certain well-known naturalizing strategies, of which Searle's is merely one well-known specimen.[10] It raises what I call the question of conceptual and ontic "adequation", the matching of the supposed nature of the brain and the properties that may be coherently ascribed to it and formulated in standard causal terms.[11] Searle nowhere addresses the matter. It obviously affects what we should mean by "physical" and the possible limits of naturalizing.

Return, then, to Quine. Quine begins his paper with a perfectly conventional remark: "epistemology", he says, "is concerned with the foundations of science".[12] He then reviews the reductive translational epistemology proposed by Carnap (and suggested by Hume) and finds it philosophically hopeless. The truths of natural science, Quine says, cannot be defined in terms of "immediate experience" any more than the truths of mathematics can be defined in terms of "elementary logic".[13] The reason is essentially the same, viz.: "a statement about the world does not always or usually have a separable fund of empirical consequences that it can call its own".[14]

This single remark is intended by Quine to signify: his pragmatist indebtedness to Peirce, in terms of linking the meanings of sentences to observational effects; his indebtedness to Duhem, construing Peirce's thesis in terms of whole theories (as complexes of sentences) rather than of discrete sentences that may be taken singly; his indebtedness to Frege, in terms of taking sentences rather than terms as the minimal bearers of meaning; his having exposed the analytic/synthetic dogma that the empiricisms of Carnap and Hume share; his advocacy of the thesis of the "indeterminacy

[10]See Searle 1984 and Searle 1992, Chapter 10.
[11]Margolis 1989a.
[12]Quine 1969: 69.
[13]Quine 1969: 74.
[14]Quine 1969: 82.

of translation", which follows directly from the sentence cited; and, finally, his own epistemology. But *what*, you may well ask, is the latter? and *where* is it defended?

Here we are getting closer to the essential quarrel. First of all, none of the implications I have just drawn out of the single sentence of Quine's that I cite – or indeed the truth of that sentence – could possibly be confirmed by invoking nothing but the resources of items (a) and (b) of the tally I have begun to draw up; or, for that matter, from (c)-resources alone. Secondly, Quine concedes – deliberately employing an old-fashioned idiom – that epistemology *is* concerned with the "*foundations* of science". He actually invokes the Kantian formula: he speaks of discerning "the ground of mathematical knowledge" – of determining "how mathematical certainty is possible" – as being the epistemologist's task;[15] and he himself offers an alternative to Carnap's translational program – call Carnap's strategy (c‴) – that we may now suppose is meant to compete with Newell and Simon's and Searle's options as well as Carnap's. Quine treats Carnap's proposal and his own, rather lightly, as "rational reconstruction[s]" of epistemology or as "legitimation[s]" of such reconstructions;[16] but he himself is obviously more interested in the actual strategy (call it (c⁗)) than with the dialectical argument by which, in a fully legitimative way, it may be installed. The full force of this maneuver is not always clear – and has plainly misled some naturalists, even where they have preferred other options.[17]

III

Before proceeding further, therefore, I seize the occasion to add to the naturalizing marks already tallied:

(d) the second-order legitimation of first-order naturalizing policies.

I think we are now very close to understanding what "naturalizing" means– hence, also, close to understanding its inadequacy. But item (d) is more complicated than may appear. It is clear that (c′)-(c‴) and similar formulas are simply alternative *first*-order strategies for installing some version of the naturalizing stance, whereas (d), as Quine himself appreciates, is the second-order argument by which to"justify" (Quine's own term[18]) one or another of those alternative (c)-strategies.

Quine, of course, has his own favorite theory, his own empirical proposal in accord with (a), (b), and (d), namely, whatever best answers the following constraint: "what we want of observation sentences [on which, as Quine claims, the whole of science depends] is that they be the ones in closest causal proximity to the sensory receptors".[19] Call that, as I say,

[15]Quine 1969: 70.
[16]Quine 1969: 78.
[17]I think Kim scants it, for instance, despite the fact that he is one of the firmest among the naturalists on the normative nature of epistemology. See Kim 1988.
[18]Quine 1969: 84.
[19]Quine 1969: 85.

(c⁗). You can see that (c⁗) is not entirely separable from arguments of the (d)-sort, arguments that say, in effect, that strategy (c⁗) *is* a viable and non-question-begging form of our (c)-like strategies for explaining how knowledge obtains.

I don't see how *any* (c)-strategy could possibly retire the (d)-question Quine raises, though I am entirely prepared to admit that (c)-strategies *are* eligible – under the auspices of the second-order query – to support naturalizing. I don't believe Quine's option works, and I don't believe he ever shows us that (or how) it can. For instance, I don't think he shows us how – or could ever show us how – to naturalize the referential use of proper names or the propositional attitudes. Failing to do that, I say, signifies failing jointly to supply a winning option of the (c)-series and to supply a (d)-argument for justifying invoking that particular (c)-strategy. (I take the contest metonymically.) You see now what I meant in speaking of dialectical bets.

I cannot stop to assess the comparative strengths of these and similar alternatives. I am trying to isolate what it means to speak of the naturalizing stance, why it fails in any and all of its forms, and what it would mean if it succeeded. The line of attack is not entirely obvious. But the first step is probably clear by now. In a fairly straightforward sense, the second-order argument (the "legitimative" argument, in Quine's own idiom: the argument explaining the "possibility" of epistemology) is not on its face an argument that is naturalized. It is an argument that would confirm (if valid) that epistemology *can* be naturalized – by way of one alternative or another drawn from the (c)-series, which would itself be naturalized. I won't say straight out that such an argument – that is, a version of (d) – is inimical to naturalizing.

Of course, if some strategy from the (c)-series were to succeed (Quine's, preeminently – or Kim's or Goldman's, say), then it would appear that arguments of the (d)-sort, which plainly address the point of competing (c)-strategies, *could* be subsumed or rendered harmless, somewhere (at a higher level of dispute), as applications of the same sort of strategies selected from the (c)-series. Furthermore, if you were to take the questions the (c)-strategies address to be first-order questions, and those (d)-strategies address to be second-order questions, it would soon be apparent that first- and second-order questions could not be neatly segregated; although, as with questions of truth and the legitimation of truth-claims, they are hardly identical.

I suggest that, regarding truth and knowledge, first- and second-order questions are indissolubly linked at every point of relevant inquiry and that the distinction between the two is itself a second-order distinction that may be variably drawn. For reasons of this sort, I am not prepared to insist *tout court* that Quine's admission of the (d)-question shows in a knockdown way that naturalizing strategies cannot possibly work. It does put naturalizing on the defensive, however, and that is not usually admitted. But I still say

Quine's strategy cannot succeed, on any reasonable assessment of what it sets out to do, and *should* set out to do.[20] Furthermore, *that* failure provides an exemplar of how to defeat the naturalizing strategy in general.

There is another side to the issue. Success (of a certain sort) of one or another of the (c)-strategies might confirm the benign standing of the second-order foundational issue – the (d)-question – as to what to regard as a legitimative account of knowledge or mind or the computational modeling of mind, or, more narrowly, of truth or meaning or reference or the like. Hence, if, by some suitable second-order argument, it *could* be shown that particular options drawn from the (c)-series failed, then, to that extent, the second-order objections would (by parity of reasoning) then count as non-naturalizable or as not sufficiently congenial to naturalizing strategies. The fate of first- and second-order (c)- and (d)-questions go hand in hand, you see. This explains why the foundational questions Quine raises suddenly seem invisible in his argument.[21] But they do not really disappear.

Once you admit that naturalizing strategies are committed to constraints (a) and (b), you cannot really dismiss the (d)-issue and, on the argument just sketched, you cannot really separate the (c)-question from the (d)-question.[22] This explains Quine's very canny pronouncement – with which I entirely agree, though not (I must say) in Quine's own favor – namely

> The old epistemology aspired to contain, in a sense, natural science; it would construct it somehow from sense data. Epistemology in its new setting, conversely, is contained in natural science, as a chapter of psychology. But the old containment remains valid too, in its way.[23]

Yes, of course. Except that *if* the naturalizing stance fails, then this pronouncement can still be read much as Quine wrote it but with the reverse sense, that is, as signifying that science and psychology and perception and epistemology are *not* convincingly naturalized or naturalizable! That is the genius of inquiries of the (d)-sort. The mutual "containment" Quine speaks of is, as such, entirely neutral to the fortunes of the naturalizing stance; in other words, it cuts both ways. It cuts both ways for all naturalizing options.

This yields two further very strong constraints. One is this. It is reasonably clear that (d)-questions are normative questions, whereas (c)-questions are not. But from what I have already said, you must see that I do not hold that naturalizing fails simply because legitimation is normative. That would be incompatible with what I have just conceded in favor of Quine's argument. The adverse judgment in fact is one that Putnam seems to favor – mistakenly, I suggest:[24] it is an argument that naturalizing analysts have been quick to neutralize by showing that values and norms

[20] See Margolis 1989b.
[21] See Stroud 1981.
[22] Quine 1969: 84–85.
[23] Quine 1969: 83.
[24] Putnam 1983.

may, without generating incoherence or contradiction, also be naturalized.[25] Of course, on the argument I have been mounting, if there were normative questions that could not be naturalized (second-order legitimation), then, by parity of reasoning, the naturalists would lose (or risk losing) here as well. So the argument on either side is entirely conditional.

The second constraint is this. There may be very strong, possibly even knockdown, first-order arguments that plainly show that and why the naturalizing stance fails in particular cases – that may be reasonably generalized to the whole of philosophy. That is indeed my conviction.

To be perfectly blunt, I hold that there are no satisfactory naturalizing strategies for reference, predication, intentionality, meaning, truth, confirmation, legitimation, knowledge, moral norms, and the like; and if there are none, then the naturalizing option fails on its own grounds and non-naturalism is vindicated. I cannot see how this line of attack can be deemed impertinent on the naturalist's own view, and I cannot see how, if that is conceded, the (d)-question can be absorbed within any strategy of the (c)-sort. And yet, the counterattack will not have invoked any considerations other than those the naturalists themselves admit.

I am in fact just beginning to broach here a set of considerations belonging to a new series – I shall call it the (e)-series: in effect, the (non-naturalizing) counterparts of the (c)-series – that either stalemate the sanguine expectations of the other or pose potentially decisive difficulties for naturalizing policies across the board. Let me give you a particular example (a single decisive example) by way of a well-known reflection not initially intended to serve this purpose.

In his treatment of the mind-body problem in the paper, "The Material Mind," Davidson affirms the following:

> Although, as I am urging, psychological characteristics cannot be reduced to [physical or biological or physiological ones], nevertheless they may be (and I think are) strongly dependent on them. Indeed, there is a sense in which the physical characteristics of an event (or object or state) *determine* the psychological characteristics; in G.E. Moore's view, psychological concepts are *supervenient* on physical concepts. Moore's way of explaining this relation (which he maintained held between evaluative and descriptive characteristics) is this: it is impossible for two events (objects, states) to agree in all their physical characteristics (or in Moore's case, their descriptive characteristics) and to differ in their psychological characteristics (evaluative).[26]

Davidson surely means that neither an eliminativist view (notoriously: the disquotational theory in the case of truth, the naturalistic fallacy in the case of moral properties) nor a reductive identity would be convincing. The parallels in the case of the mind-body problem – something like Churchland's

[25] See Kim 1988.
[26] Davidson 1980b: 253. See also Davidson 1980c: 214 and McGinn 1991.

eliminativism and Smart's reductive physicalism – are regarded by Davidson as not serious contenders of what I am calling the (c)-series, for carrying the naturalizing flag.[27]

But if you ask what could possibly vindicate the supervenience thesis – a specimen of the (c)-series – it stares you in the face that either the thesis is completely arbitrary (as some have supposed[28]) or else it signifies that the mental or psychological is *necessarily* determinable in every instance by a rule or principle or algorithm that, in effect, models the mental computationally.

I have no objection to the supervenience proposal, if all we are asked to consider is its bare coherence. But that is clearly insufficient for legitimating the (c)-option it offers; and it is certainly fair to say that Davidson nowhere provides the least argument in favor of the (d)-claim, the *modal* claim – nor does McGinn, nor, as far as I am aware, does anyone else (Hare, for instance, with respect to values; Kim, with respect to epistemology).

Let me put it to you this way: if there were good reasons for believing that the mental cannot be computationally modeled in this way, then we should have a good clue to arguments of the (e)-sort, that would do in the pertinent (c)-pretensions and would therefore color the (d)-question in a non-naturalizing way. Both because of its recent salience and surprising ubiquity and because, *au fond*, it now appears as the most general strategy for smoothing the entry of a computational model of mind and knowledge into the resolution of the relevant legitimative question, I shall take supervenience as the current favorite of the naturalizing options. In that way I gain a convenient focus and economy: for the essential weakness of the supervenience thesis collects the weaknesses of all the naturalizing options and, at the same time, points ineluctably to its older heritage (unity of science, extensionalism, physicalism) and to its more fashionable new objective (the global adequacy of computational models or models congruent with same, the entrenchment of direct realism, the near-abandonment of legitimative inquiry). What I offer, of course, is the bare sketch of a complex argument, not the full argument itself. I mean to show that:

(i) supervenience can be denied without self-contradiction,

(ii) that it can and should be denied in a way that fruitfully links (c)- and (d)-considerations, and

(iii) that showing *that* shows us how to defeat the naturalizing strategy across the board (even in the absence of strict supervenience).

IV

Let me cut through all the subtleties. I suggest that supervenience, which I take to be a theory of some sort about the way the world is – not a

[27] Churchland 1989: 2–6; Hare 1952: 50–81, 145, 153.
[28] Blackburn 1985.

theory about the mere use of terms – must be distinguished from at least two doctrines, with which it is easily confused, that are themselves about one or another aspect of linguistic usage. I am certain that that's what G.E. Moore had in mind originally;[29] and I am certain that that's what Davidson had in mind and what naturalizing epistemologists have in mind.

The first notion that supervenience ought not to be confused with is the notion of consistency of usage – what Hare famously called the "principle of universalizability": the prescriptive commitment (hence, the "nontrivial" commitment) to the logically trivial "descriptive meaning-rule", viz. "that we cannot without inconsistency apply a descriptive term to one thing, and refuse to apply it to another similar thing (either exactly similar or similar in the relevant respects)". Moral judgments, Hare says, are "*in the same sense*, universalizable".[30] Hare seems to have been partly aware, and partly not, that consistency of usage, applied to interpreted terms, is indifferent to the particular interpretation of the terms in question. He seems in fact to have thought that universalizability (thus construed) *was* a moral principle – which of course it cannot be (even though deliberately violating consistency may be a morally serious matter).

The other notion that supervenience must not be confused with is the notion of entailment – which, I say, neither Davidson (with regard to the mind-body problem) nor analytic epistemologists (Goldman, for instance) ever quite shows us how to avoid. For, speaking informally, I mean, by entailment, that A entails B if, given the truth of A, the truth of B "follows" as a consequence of the sense of A. Now, then, my argument – put in the baldest terms – is that no one has ever given a compelling reason for supposing that the mental supervenes on the physical or the cognitive on the noncognitive or the moral on the nonmoral or the nonnatural on the natural in any way that could be said to hold "necessarily" in some *de re* sense. That's all. I insist that *any* affirmative or negative argument addressing the claim is a (d)-argument that, ineluctably, yields a (c)- or an (e)-argument.

The supervenience argument cannot be confined to constraints of a purely linguistic sort: consistency or entailment, for instance. It cannot be a causal argument or one involving nomologicality in any familiar sense. It specifically eschews identity arguments. It is meant, in some sense, to be a *de re* analysis of the way the world is – which, of course, was precisely what Moore had in mind. But it clearly falls foul of constraint (b), which, as I say, is characteristic of recent naturalizing strategies. One could relent here, but we need the (d)-argument that would justify doing so. It is nowhere to be found.

Many interesting analyses in epistemology and the philosophy of mind fail to supply the affirmative (d)-argument needed. For example, the various forms of what has come to be called "reliabilism", which Alvin Goldman has

[29] Moore 1942: 588. See Kim 12984.
[30] Hare 1963: 12–13. See further Margolis, 1971, Chapter 4.

done so much to strengthen, fails to explain, as far as I can see, what the *real* connection is between causal matters that purposedly improve the epistemic reliability of our beliefs and why any such linkages should actually count as such:[31] any strong admission of holism along Quine's or Duhem's lines would seriously challenge the causal account; there is no clear sense in which a causal account of theoretical beliefs is sufficiently straightforward; intentional complications cannot be ignored or dismissed; and no causal linkages that are of the right gauge may actually be convincingly characterized as nomological. Also, the various forms of what is generally called "eliminativism" in the philosophy of mind – notably, Churchland's thesis – fail in precisely the same way to answer the pertinent (d)-question.

I think both maneuvers are, or are linked to the fortunes of, supervenience-arguments: Goldman's, implicitly to the effect that supervenience obtains between the causal and epistemological; Churchland's, explicitly to the effect that nothing like supervenience need obtain between the mental and the physical, because the phenomena to be explained are a philosophical delusion. If they are not supervenience theories, then either they are completely arbitrary (which is barely possible) or they rely on some alternative (d)-strategy that is at least as strong and that works in a way very similar to the supervenience doctrine. I don't think it matters what to make of such arguments, so long as the answer to the (d)-question remains invisible.

Here, finally, I am prepared to relax the logical rigor of the arguments required. I think supervenience requires a very strong modal reading of "necessity". But, for the sake of philosophical dialogue, I should be entirely willing to accept contingent arguments in favor of a plausible linkage between (c)-arguments and (d)-arguments stronger than what can be constructed linking counterpart (e)-arguments and their (d)-arguments. I know of none.

In this adjusted sense, I say that naturalizing utterly fails. If it does, then of course it follows trivially that the mental cannot be convincingly modeled computationally. The relevant (e)-arguments will inevitably concern themselves, in realist terms, with the analysis of cultural emergence, intentionality, the role of *Lebensformen*, the inherent informality and ineliminability of context, historicity, constructed or artifactual aspects of the entire intelligible world and our own natures as persons, and the inseparability of first- and second-order questions. All these matters are known to be peculiarly strenuous for the naturalizing strategy and are largely ignored.

References

Simon Blackburn 1985 "Supervenience Revisited", in Ian Hacking (ed.), *Exercises in Analysis: Essays by Students of Casimir Lewy*, Cambridge: Cambridge University Press.

David Bloor 1974 *Knowledge and Social Imagery*, London: Routledge.

Nancy Cartwright 1983 *How the Laws of Physics Lie*, Oxford: Clarendon.

Paul M. Churchland 1989 *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, MA: MIT Press.

[31] See Goldman 1991b.

K.M. Colby 1963 "Computer Simulation of a Neurotic Process", in S.S. Tomkins and Samuel Messick (eds.), *Computer Simulation of Personality, Frontiers of Psychological Research*, New York: John Wiley.

K.M. Colby 1975 *Artificial Paranoia*, New York: Pergamon.

Donald Davidson 1980a *Essays on Actions and Events*, Oxford: Clarendon.

Donald Davidson 1980b "The Material Mind", in *Essays on Actions and Events.*

Donald Davidson 1980c "Mental Events", in *Essays on Actions and Events.*

Bas C. Van Fraassen 1989 *Laws and Symmetry*, Oxford: Clarendon.

Alvin I. Goldman 1991a *Liaisons: Philosophy Meets the Cognitive and Social Sciences*, Cambridge: MIT Press.

Alvin I. Goldman 1991b "Strong and Weak Justification", *Liaisons.*

Alvin I. Goldman 1991c "What Is Justified Belief", *Liaisons.*

R.M. Hare 1963 *Freedom and Reason*, Oxford: Clarendon.

R.M. Hare 1952 *The Language of Morals*, Oxford: Clarendon.

John Haugeland (ed.) 1981 *Mind Design*, Montgomery, VT.: Bradford Books.

Jaegwon Kim 1984 "Concepts of Supervenience", *Philosophy and Phenomenological Research* LXV.

Jaegwon Kim 1988 "What Is 'Naturalized Epistemology'?" *Philosophical Perspectives* II.

Philip Kitcher 1992 "The Naturalists Return", *Philosophical Review* CI.

Colin McGinn 1991 *The Problem of Consciousness*, Oxford: Basil Blackwell.

Joseph Margolis 1971 *Values and Conduct*, Oxford: Clarendon.

Joseph Margolis 1989a "Constraints on the Metaphysics of Culture", in Margolis, *Texts without Reference.*

Joseph Margolis 1989b "The Grammar and Ontology of Reference", in Margolis, *Texts without Reference.*

Joseph Margolis 1989c *Texts without Reference: Reconciling Science and Narrative*, Oxford: Basil Blackwell.

G.E. Moore 1942 "A Reply to My Critics", in Paul Arthur Schlipp (ed.), *The Philosophy of G.E. Moore*, Evanston: Northwestern University Press.

Allen Newell and Herbert A. Simon 1976 "Computer Science as Empirical Inquiry: Symbols and Search", *Communications of the Association for Computing Machinery* XIX. Reprinted in Haugeland 1981.

Hilary Putnam 1983 "Why Reason Can't Be Naturalized", *Philosophical Papers* 3, Cambridge: Cambridge University Press.

W.V. Quine 1969 "Epistemology Naturalized", in Quine, *Ontological Relativity and Other Essays*, New York: Columbia University Press.

John R. Searle 1980 "Minds, Brains, and Programs", *Behavioral and Brain Sciences* III. Reprinted in Haugeland 1981.

John R. Searle 1984 *Minds, Brains and Science*, Cambridge: Harvard University Press.

John R. Searle 1992 *The Rediscovery of the Mind*, Cambridge: MIT Press.

Barry Stroud 1981 "The Significance of Naturalized Epistemology", *Midwest Studies in Philosophy* VI, Minneapolis: University of Minnesota Press.

A.M. Turing 1950 "Computing Machinery and Intelligence", *Mind* LIX.

Creativity and Representational Redescription

Margaret A. Boden

1 Introduction

Creativity involves the mapping, exploration, and transformation of conceptual spaces: METCS, for short (Boden 1990). Conceptual spaces are styles of thinking in people's minds. They include styles in sculpture and painting, literary genres, patterns generated by composers, choreographers, and couturiers, and scientific and mathematical theories.

"Maps" of conceptual spaces are internal representations, or descriptions, of the creator's thinking-skills. They may be many-levelled, and are not necessarily conscious – though some are. They articulate the structure of the spaces concerned. For example, they indicate pathways, boundaries, and, sometimes, potential "tunnels" into closely-related spaces. Some heuristics for exploring and transforming conceptual spaces are especially relevant to certain sorts of conceptual "landscape".

Recent research in developmental psychology has shown that children's skills are utterly inflexible when the child first achieves fluency. Only later does imaginative flexibility become possible. This results from spontaneous redescription, at successively higher levels, of a (fluent) lower-level skill (Karmiloff-Smith 1990, 1993). This "representational redescription" (RR) is closely linked to creativity, since it provides many-levelled maps of the mind which can be used by the subject to do things which they *could not* do before.

The importance of that 'could not', with respect to a theory of creativity, is explained in Section 2. In Section 3 (drawn from Boden 1990, Chapter 4), I show how RR supports this account of creativity. Finally, in Section 4, I explore the relevance of RR to creativity in adult life.

2 What is Creativity?

Creativity is sometimes said to be the ability to use intuition (or insight) to produce new ideas. But this will not do. Creative ideas are not only new, but deeply surprising – though some also appear surprisingly *obvious*. As for intuition, this term marks the fact that artists and scientists typically have their creative ideas unexpectedly, with little if any conscious awareness of how they arose. But to mark a fact is not to explain it. To include "intuition" in the definition of creativity is to multiply mysteries.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 25-35

The philosophical challenge, then, is to distinguish creativity from (mere) novelty – preferably, in a way that helps us to understand both the surprise-value of creative ideas and their sometimes startling obviousness. The psychological challenge is to explain creativity in scientifically intelligible terms.

People of a scientific cast of mind often try to define creativity in terms of "novel combinations of old ideas". In that case, the surprise caused by a creative idea must be due to the improbability of the combination, and purely statistical tests – as used by psychometricians (Eysenck 1994) – could identify creativity.

Combination-theorists typically leave something important unsaid. The "novel combinations" have to be not only new, but interesting: to call an idea creative is, in part, to say that it is valuable. Combination-theorists, however, usually omit value from their definition of creativity. It is not surprising, then, that they seldom discuss the strong *social* influences determining which ideas are judged to be "creative" (Schaffer 1994).

This cavil aside, what is wrong with the combination-theory? Many ideas we regard as creative are indeed based, at least in part, on unusual combinations. Much of Coleridge's enchanting imagery in "The Ancient Mariner", for example, draws together diverse ideas scattered in his eclectic reading. So combination-theory is not utterly irrelevant. However, it cannot explain, or even adequately describe, the most intriguing cases. "Creativity" is not a natural kind. Quite apart from the many, and untidily various, criteria used in evaluating candidate-ideas, creative novelties are of significantly different types – the most interesting of which lie beyond combination-theory.

Many creative ideas are surprising not because they involve unusual juxtapositions of familiar ideas, but in a deeper way. They concern novel ideas which not only *did* not happen before, but which – in a sense that must be made clear (and which combination-theory cannot express) – could not have happened before.

Before considering just what this seemingly paradoxical 'could not' means, we must distinguish two senses of this (deep, or impossibilist) type of *creativity*. One is psychological (P-creativity), the other historical (H-creativity). An idea is P-creative if the individual person in whose mind it arises could not (in the relevant sense) have had it before; it does not matter how many times other people have already had the same idea. By contrast, an idea is H-creative if it is P-creative *and* no-one else has ever had it before.

There can be no systematic explanation of H-creativity. Whether an idea survives, whether it is lost for a while and resurfaces later, and whether historians (and/or peers) at a given point in time happen to know about it, and happen to value it, depend on a wide variety of unrelated factors – including fashion, flood, and fire. It follows that there can be no psychological explanation of H-creativity as such. But all H-creative ideas, by definition, are P-creative too. So an explanation of P-creativity would apply to H-creative ideas as well.

What does it mean to say that an idea "could not" have arisen before? Unless we know that, we cannot make sense of P-creativity (or H-creativity either), for we cannot distinguish radical originality from mere "first-time" newness.

Noam Chomsky, discussing what he called the "creativity" of natural language, reminded us that language is an unending source of novel (even H-novel) sentences. But these sentences are novelties which clearly *could* have happened before, being generated by the same rules that can generate other sentences in the language. Any native speaker (and many computers, too) could produce novel sentences using the relevant grammar. In general, to come up with a new sentence is not to do something P-creative.

The 'coulds' in the previous paragraph are computational 'coulds'. That is, they concern the set of structures described and/or produced by one and the same set of generative rules. Sometimes, we want to know whether a particular structure is describable by a specific schema, or set of abstract rules. – Is this a sonnet, and is that a sonata? – To ask *whether an idea is creative or not* (as opposed to how it came about) is to ask this sort of question. But whenever a particular structure is produced in practice, we can also ask what generative processes actually went on in the computational system concerned. – Was the sonata composed by following a textbook on sonata-form? – To ask how an idea (creative or otherwise) actually arose, is to ask this type of question.

We can now distinguish first-time novelty from radical originality. A merely novel idea is one which can be described and/or produced by the same (specified) set of generative rules as are other, familiar, ideas. A genuinely original, or creative, idea is one which cannot. It follows that *constraints*, far from being opposed to creativity, make creativity possible. To throw away all constraints would be to destroy the capacity for creative thinking.

To justify calling an idea creative, then, one must specify the particular set of generative principles (the conceptual space) with respect to which it is impossible. So literary critics, musicologists, and historians of art and science have much to teach the psychologist. But their knowledge of the relevant conceptual spaces must be made as explicit as possible, to clarify just which structures *can*, and which *cannot*, be generated within them.

With respect to the familiar structures in the relevant domain (chemistry, poetry, music ...), a "deeply" creative idea is not just improbable, but impossible. How can it arise, then, if not by magic? A generative system defines a certain range of possibilities: molecules, for example, or jazz-melodies. These structures are located in a conceptual space whose limits, contours, and pathways can be mapped, explored, and transformed in various ways.

The "mapping" of a conceptual space involves the representation, whether at conscious or unconscious levels, of its structural features (Boden 1990, Chapter 4). The more such features are represented in the mind of the person concerned, the more power (or freedom) they have to navigate and

negotiate these spaces. A crucial difference – probably the crucial difference – between Mozart and the rest of us is that his cognitive maps of musical space were very much richer, deeper, and more detailed, than ours. In addition, he presumably had available many more domain-specific processes for negotiating them.

What counts, in this context, as exploration? One example is the development of post-Renaissance Western music, which is based on the generative system known as tonal harmony. Each piece of tonal music has a "home key", from which it starts, from which (at first) it did not stray, and in which it must finish. Travelling along the path of the home key alone soon became insufficiently challenging. Modulations between keys then appeared, within the body of the composition. But all possible modulations did not appear at once. At first, only a small number within one composition were tolerated, and these early modulations took place only between keys very closely related in harmonic space. With time, the modulations became more daring, and more frequent. By the late nineteenth century there might be many modulations within a single bar, not one of which would have appeared in early tonal music.

Eventually, the very notion of the home key was undermined. With so many, and so daring, modulations within the piece, a "home key" could be identified not from the body of the piece but only from its beginning and end. Inevitably, someone (it happened to be Schoenberg) suggested that the convention of the home key be dropped altogether, since it no longer made sense in terms of constraining the composition as a whole.

Exploring a conceptual space is one thing. Transforming it is another. In general, novel ideas gained by exploring an unknown niche in a pre-existing conceptual space are regarded as less creative than ideas formed by transforming that space in radical ways.

One example of transformation has just been mentioned: Schoenberg's dropping the home-key constraint to create the space of atonal music. *Dropping a constraint* is a general heuristic for transforming conceptual spaces. Non-Euclidean geometry, for instance, resulted from dropping Euclid's fifth axiom, about parallel lines meeting at infinity. Another very general way of transforming conceptual spaces is to *consider the negative*: that is, to negate a constraint.

One well-known instance concerns Kekulé's discovery of the benzene-ring. He described having an image of atoms "in long rows, sometimes more closely fitted together; all twining and twisting in snakelike motion. But look! What was that? One of the snakes had seized hold of its own tail, and the form whirled mockingly before my eyes. As if by a flash of lightning I awoke." This vision was the origin of his hunch that the benzene-molecule might be a ring. Previously, Kekulé had assumed that all organic molecules are based on strings of carbon atoms.

We can understand how it was possible for him to pass from strings to rings, as plausible chemical structures, if we assume three things (for each

of which there is independent evidence): that snakes and molecules were already associated in his thinking; that the topological distinction between open and closed curves was present in his mind; and third, that the "consider the negative" heuristic was present also. A string-molecule is an open curve. If one considers the negative of an open curve, one gets a closed curve. Moreover, a snake biting its tail is *a closed curve which one had expected to be an open one.* For that reason, it is surprising, even arresting ("But look! What was that?"). Finally, the change from open curves to closed ones is a topological change (a change in neighbour-relations), and Kekulé knew that a change in atomic neighbour-relations is likely to have some chemical significance. So it is understandable that he had a hunch that this tail-biting snake-molecule might contain the answer to his problem.

(Hunches are common in human thinking, and an adequate theory of creativity must be able to explain them. It must show how it is possible for someone to feel – often, correctly – that a new idea is promising *even before* they can say just what its promise is. The example of Kekulé suggests that a hunch is grounded in appreciation of the structure of the space concerned, and some notion of how the new idea might fit into it.)

A third common way of transforming a conceptual space is to *vary the variable.* So chemists after Kekulé, who knew very well that carbon is just one of about 90 elements, asked whether ring-molecules might exist with (for instance) nitrogen or phosphorus atoms in the ring. And many examples could be given of people's substituting one numeral for another, where the space is partly described in numerical terms. Thus Kekulé's successors asked whether there might be less than six atoms in a ring-molecule, and Hindus asked whether Kali might have six arms, not two. (Even Humpty-Dumpty varied the variable: when Alice remarked "One can't help growing older", he grimly replied "One can't, perhaps – but two can. With proper assistance, you might have left off at seven [years old]".)

In these ways, and others, our maps of conceptual space can be explored, and even transformed. The notion of mental maps helps us to see how "appreciation of the structure of the space" may be grounded. Much as a real map helps a traveller to find – and to modify – his route, so mental maps enable us to inhabit our conceptual spaces in imaginative ways.

3 Developing Maps

Some intriguing recent work in developmental psychology illuminates how simple mental maps arise, and how they can be used. Annette Karmiloff-Smith's (1990; 1993) theory of representational redescription (RR) explains the growing fluency of children's skills in terms of spontaneously arising representations of increasing descriptive power. RR is closely linked to creativity, since it provides many-levelled maps of the mind which can be used by the subject to do things which they *could not* do before (Boden 1990, Chapter 4).

The fundamental idea of RR-theory is that when children (and adults) practise new skills, they spontaneously develop explicit mental representations of knowledge they already possess in an implicit form. As Karmiloff-Smith puts it: "knowledge embedded in procedures gradually becomes available, after redescription, as part of the system's data-structures".

These representations arise on several successive levels, each time enabling the person to exploit the prior knowledge in ways that were not possible before. The person progresses from a skill that is fluent but "automatic" (being varied only with much effort, and limited success), to one that can be altered in many ways. Eventually, the skill can be deliberately altered, and its results carefully evaluated, by means of self-reflective consciousness (for which RR is a necessary precursor).

Karmiloff-Smith's work shows, for instance, that young children need explicit (though not necessarily conscious) representations of their lower-level drawing-skills in order to draw imaginatively: to P-create a picture of a one-armed man, for instance, or a seven-legged dog. Before developing such RRs, the child simply *cannot* draw a one-armed man (and finds a two-headed man extremely difficult even to copy).

In her experiments, a child (aged between four and eleven) would be asked to draw "a house". Then the first drawing would be removed, and the child would be asked for "a house that does not exist" (or "a funny house", "a pretend house", "a house you invent", and so on). Similarly, the children were asked to draw a man and then a funny man, or an animal followed by a pretend animal. She recorded not only the resulting drawings, but the way in which each individual child went about drawing each picture.

The children's drawings are "imaginative", or "interesting", in a number of different ways. There may be a change in the shape, or the size, of component elements; so a door is spiky, and a head is tiny or square. There are cases where the shape of the whole thing is changed; so we have houses like tripods or ice-cream cones. Sometimes, elements are deleted, giving doorless houses or one-legged men. Other times, there are changes in the position or orientation of elements, and/or of the whole thing; we see doors opening into mid-air, an arm and a leg switched, and a house upside-down. In some cases, extra elements are inserted into the structure (as opposed to being added after the thing has been drawn as a normal whole), resulting – for example – in many-headed monsters of various kinds. And sometimes, the extra elements come from a different category of thing; so a man is given an animal's body, and a house is given wings.

These imaginative transformations do not happen at random. The flexibility of the drawing-skill – the creative range – depends on the age of the artist. All the children could draw (real) houses, men, and animals fluently: their sketches were done quickly and effortlessly. But drawing funny houses, or men that do not exist, required them to alter their usual drawing-method. The younger children found this difficult, being unable to vary their drawings in all the ways possible for a 10-year-old. Children of all ages varied

size or shape, and deleted elements. But the 8- to 10-year-olds were much more likely to insert elements (whether same-category or cross-category), or to change position or orientation, than the 4- to 6-year-olds.

It seems that 4-year-olds have rather uninformative mental maps of their own drawing-skills, for they can explore these conceptual spaces only in very superficial ways. A path here or there can be made wider or more crooked, or sometimes (under special conditions, described below) deleted altogether. But this path cannot be inserted into that one. Orientation and position are fixed: it is as though a river flowing from North to South could be made wider, and more meandering, but could not be made to run from East to West – nor transferred from the Himalayas to the Alps. And there are no pathways made up of alternating stretches of river and road: such a mixture appears to be inconceivable. The 10-year-olds, by contrast, can explore their mental territory in all these ways. Their mental maps seem to make the necessary sorts of distinction, for they can ask what would happen if the river became part-road – and they can make it happen.

These changes in imaginative power come about, according to RR-theory, because children develop explicit representations of knowledge they already possess implicitly. In other words, the skill is *redescribed* at a higher level. (The earlier representation is not destroyed, and is still available for routine use if required.) Whereas implicit knowledge can be used but not explored, explicit descriptions allow an activity to be transformed in specific ways.

The 4-year-olds are constrained by a fixed, "automatic", sequence of bodily actions, which they can vary only in very limited ways. They can draw a man with ease (quickly, and without hesitation or mistake), but not a two-headed man. Because their knowledge of their own drawing-skill is almost entirely implicit, they can generate hardly any variations of it. They are like someone who knows how to reach a place by following a familiar route, but cannot vary the route because they have no map showing how the various parts relate to one another. The oldest children represent their skill in a much more explicit manner, and as a result can produce drawings which the younger children cannot.

At the first level of redescription, a drawing-skill (already mastered as a rigid sequence of bodily actions) is represented in the mind as a strictly-ordered sequence of parts – for instance, head-drawing or limb-drawing parts. This sequence must be run from beginning to end, although it can occasionally be stopped short. Variable properties of the parts (like size and shape) are explicitly marked, allowing for certain sorts of imaginative distortion: heads can be made square, or arms very small. But the relation between the parts is represented only implicitly, depending on their order in the drawing-procedure as a whole.

Consequently, body-parts are dropped by 4-year-olds only rarely, and then only if they are at the end of the procedure. So an arm or a leg may be dropped (by a child who normally draws it last of all), but the head –

because it is usually drawn first – is almost never left out.

This first-level description of the basic bodily skill can generate funny men, but their inner structure does not vary. It does not allow for repetition of a part "at the same place" within the sequence. Nor can it generate any re-ordering of the parts, or the insertion of a part into the sequence. To be sure, 5-year-olds were able (if asked) to draw "a house with wings". But they did this by *adding* the wings to the completed house, not by interrupting their drawing of the wall-lines so as to *insert* the wings smoothly into the picture. Similarly, of the very few who mixed categories without being asked to do so, all added the foreign part after the main item had been drawn normally.

In short, very young children *cannot* insert extra elements into their drawings. When Karmiloff-Smith asked 5-year-olds to draw "a man with two heads", she found (as she predicted) that most could not do so. Typically, they would draw two heads and then attach a body-with-arms-and-legs to each head. If they were dissatisfied with the result, they would start again – but they succeeded only after very slow and elaborate efforts. They even found it difficult to *copy* drawings of two-headed men. They seemed to have an inflexible ("compiled") man-drawing procedure, which had to be run straight through. The first line of the head triggered the rest of the procedure, and it was impossible to go back and correct what had been done.

At the next level of description, the structure of the skill is mapped as a list of distinct parts which can be individually repeated and rearranged in various ways. The ordering-constraint is relaxed (though not dropped): a single part can now be deleted from the middle of the process, without disrupting the rest of the drawing. As a result, we see much more flexible behaviour. Funny men with extra arms are drawn, and houses are spontaneously given wings. But the flexibility is limited: there are no two-headed men, for instance, and the wings are still added to the house rather than inserted in it.

As the representation develops further, many of the structural relations between the (second-level) parts come to be explicitly mapped, and can then be flexibly manipulated. Subroutines – even some drawn from different categories – can be perfectly inserted into a drawing-procedure, the relevant adjustments (such as interruption of lines) being made without fuss. For the first time, we see funny men with two heads and three legs, fluently drawn with no rubbings-out. Also, we see parts of one representation being integrated into another, so that man and animal, for instance, are smoothly combined.

Evidently, 10-year-olds can explore their own man-drawing skill in a number of systematic ways. They can create funny men by using general strategies such as distorting, repeating, omitting, or mixing parts chosen from one or more categories. In effect, their conceptual space has more dimensions than the conceptual space of the 4-year-olds, so they can generate a wider – and more interesting – range of creations. In short, their RR-maps

enable them to act in ways which were simply not possible before.

4 RR and Adult Creativity

We have seen that creativity involves METCS: mapping the structures in one's mind, and using those maps to negotiate and transform the conceptual spaces concerned. Representational redescription is an example of such mapping.

Very likely, these sorts of spontaneous redescription go on in adult minds too, constructing conceptual spaces on many different levels. (Karmiloff-Smith cites some evidence, drawn from piano-playing and adult literacy, that this is so.) As the successive representations multiply, the skill becomes more complex and subtle – and more open to insightful, self-disciplined, control.

The question then arises whether the process of mapping, or re-describing, is much the same in the adult as in the child – and if so, why.

For example, consider the order in which descriptions of distinct types appear. We saw in Section III that there are some dramatic age-related differences between the sorts of transformations made by young children. The size and shape of parts can be altered before the size and shape of the whole. Size or shape (of a whole drawing or of a drawing-part) can be varied long before position or orientation. Deletion precedes insertion (and seems to work backwards: it is easier to cut short a triggered procedure than to start it half-way through). Both deletion and insertion precede changes in position or orientation. And cross-category insertions come last of all.

This evidence suggests that size and shape (of whole or parts) are described in ways different from those used to describe position and orientation. Moreover, the former representations seem to be easier for the system to generate than the latter. Why is this?

Once something (a whole or a part) has been labelled as a unit, its size and shape can be varied without needing to refer to other units – and without having to make compensatory changes in them. In this sense, size-shape changes are simple, requiring only simple descriptors to control them. Position and orientation, by contrast, relate a unit to other units or to a wider field. Accordingly, more complex descriptors are required for these transformations.

The early appearance of final-deletion suggests that, once the procedure has been described as a sequence of parts, it can be cut short by merely dropping one (or more) of the final parts at the relevant "starting-point". But the deletion of "inner" parts requires that the end-point of the deletion be noted, too – and, sometimes, that "extra" actions be improvised in order to link the two procedural units on either side of the gap. What counts as an acceptable link, in turn, will depend on the descriptions of the newly-neighbouring units (and, perhaps, on their relation to other parts of the whole as well).

The late appearance of cross-category insertions suggests that descriptors are somehow grouped, and that it is easier to select descriptors from within one and the same group than it is to "mix and match" from different groups. Just why this should be so, however, is not clear.

Until we know much more about the specific nature of the descriptions generated, we shall not understand clearly why the order of development is as it is. Some suggestions have been made as to how a connectionist system might construct "symbolic" models of the activity of another such system, but these ideas are still sketchy (Clark & Karmiloff-Smith 1993). C. Thornton (forthcoming) has argued that a certain type of constructive induction (exploring compositions on a set of base-functions) might be a plausible computational model for a sort of representational redescription. However, he points out that this implies that even sea slugs should be capable of reflexive thinking, so constructive induction is not enough.

In computer models of creativity, of course, we can decide on the descriptions – if any – ourselves. And we can decide what sorts of transformations – again, if any – the system can carry out on those descriptions.

AI-models of creativity that do not contain reflexive descriptions marking the program's own procedures, or which lack ways of varying them, are necessarily limited to exploring their conceptual spaces, rather than transforming them. Harold Cohen's program that produces line-drawings of acrobats, for example, is incapable of drawing "funny" acrobats (McCorduck, 1991). It can draw an acrobat with only one arm visible (because the other is occluded by another acrobat's body), but it cannot draw a one-armed acrobat.

If its knowledge of human anatomy included the declarative statement that people have 2 arms, and if it also had a vary-the-variable heuristic, it might change this 2 to a 1 – or, for that matter, to 6. Even so, the resulting drawing would be in the same graphic style as before, and would depict clearly humanoid figures (like Kali).

In general, varying parameters (such as size) will effect a less surprising, less fundamental, transformation than changing functions (compare substituting a wing for an arm). This can be clearly seen in two graphics-programs which employ "evolutionary" methods to vary their own drawing-rules.

Karl Sims' (1991) system uses self-modifying "genetic algorithms" (modelled on biological mutations) to generate new images, or patterns, from pre-existing ones. At each "generation", the programmer selects the most aesthetically pleasing examples, and these are used to "breed" the next generation. A similar method is used by the sculptor William Latham, to generate 3D-forms likely to satisfy specific aesthetic constraints (Todd & Latham, 1992).

So as not to jeopardize those constraints, Latham allows self-transformations only at relatively superficial levels in the program (changes in parameters, not functions). In consequence, the varying forms produced by Latham's system are much less diverse than those produced by Sims. In

Sims' system, the mutations can affect the image-generating functions themselves: one function – chosen at random – may be nested within another, or concatenated with it, or substituted for it. Sometimes, Sims' system goes in for relatively unadventurous exploration, giving us what is obviously "the same" image with its colours changed, or its lines blurred. But sometimes one cannot say, merely by looking at the parent-child image-pair, how they are related – or if they are related at all. The one appears to be a radical transformation of the other, or even something entirely different. (It does not follow that Sims' program is "better" than Latham's: if the artist-programmer is trying to achieve a particular type of aesthetic effect, the system's freedom of transformation must be limited.)

In sum: the ways in which one is able to explore and transform a conceptual space depend crucially on the maps one has available. Representational redescription, in providing a variety of explicit representations of previously-implicit skills, lays the groundwork for imaginative and self-disciplined change. This has been convincingly demonstrated in young children, and some supportive evidence exists (for a limited number of skills) in adults also. My suggestion is that to understand (for example) Mozart's creativity we would need to know what were his (many-levelled) mental maps of musical space, and what processes he employed to explore and alter them.

References

Boden, M. A. 1990 *The Creative Mind: Myths and Mechanisms*, London: Weidenfeld & Nicolson.

Clark, A., & A. Karmiloff-Smith 1993 "The Cognizer's Innards", *Mind and Language* 8, 487–519.

Eysenck, H. J. 1994 "The Measurement of Creativity". In M. A. Boden (ed.), *Dimensions of Creativity*. Cambridge MA: MIT Press.

Karmiloff-Smith, A. 1990 "Constraints on Representational Change: Evidence from Children's Drawing." *Cognition* 34, 57–83.

Karmiloff-Smith, A. 1993 *Beyond Modularity: A Developmental Perspective on Cognitive Science*, Cambridge MA: MIT Press.

McCorduck, P. 1991 *Aaron's Code*, San Francisco: W. H. Freeman.

Schaffer, S. 1994 "Making Up Discovery", in M. A. Boden (ed.), *Dimensions of Creativity*, Cambridge MA: MIT Press, pp. 13–52.

Sims, K. 1991 "Artificial Evolution for Computer Graphics", *Computer Graphics* 25 (no.4), July 1991, 319–328.

Thornton, C. forthcoming "Representational Redescription for Sea Slugs". Paper given to the AAAI "AI & Creativity" symposium, March 23-25 1993 (Stanford, California).

Todd, S., & W. Latham. 1992 *Evolutionary Art & Computers*, London: Academic Press.

Artificial Agents and Mental Properties

Francesco Orilia

1 Introduction

Two quite different theses can be distinguished at the core of what Searle[1] has called *strong AI*:

(1) It will be technologically feasible – essentially by the means and strategies of current AI – to build an inorganic artificial agent that exhibits human-like intelligent behavior.

(2) Such a hypothetical artificial agent would have (*genuine*) *mental properties*, i.e., it would have beliefs, desires, fears, hopes, pleasures and pains, in a sense that implies having subjective mental states, experiences, *qualia*, awareness.

Thesis (1) would be verified, should there be a candidate artificial agent that is able to pass a version of the Turing test. Presumably, this version should be more sophisticated than the original one and resemble what Harnad[2] calls the Total Turing Test (TTT). I am inclined to believe that it is in principle possible that some future artificial agent will pass TTT,[3] but I shall argue that (2) can nonetheless be questioned. My argument differs from other attacks on strong AI, by relying heavily on a defense of the view on the mind-body problem known as *property dualism*. I shall assume that our pre-theoretical mutual understanding of what it is like to have subjective mental states, essentially Nagel's sense,[4] is sufficiently clear for a critical scrutiny of (2), contrary to what, e.g., Sloman or Dennett[5] might appear to imply.

In the following, I shall for convenience use "Roby" to refer to the hypothetical artificial agent that has passed TTT, thereby verifying (1).

2 The Antisolipsistic Hypothesis and the Physical Stance

According to old-fashioned behaviorism, (2) semantically or conceptually follows from (1). As against a rash acceptance of an implicational line from

[1] Searle, "Minds, Brains, and Programs".
[2] Harnad, "Other Bodies, Other Minds".
[3] I have given my reasons for this in Orilia, "Intelligenza artificiale e proprietá mentali".
[4] T. Nagel, *The View from Nowhere* and "What is it Like to be a Bat?"
[5] Sloman, "The Emperor's Real Mind", Dennett, *Brainstorms* ch. 11, "Why You Can't Make a Computer that Feels Pain".

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 37-50

(1) to (2), it should be observed that intelligent behavior and possession of mental properties are conceptually distinct and in fact the *solipsistic hypothesis* is perfectly intelligible, although utterly unbelievable. According to this hypothesis, *I* am the only one really endowed with mental properties, whereas any other intelligently behaving entity possesses them only apparently. Contrariwise, the *anti-solipsistic hypothesis*, uncritically accepted by common sense, claims that not just *I*, but all other human beings (and other animals) have mental properties. Undoubtedly, this is motivated by the intelligent behavior of other human beings and animals. Nevertheless, the anti-solipsistic hypothesis does not appear to be quite as obvious, if extended to Roby. This is mainly because this hypothesis underlies folk-psychological causal explanations of other people's (intelligent) actions. I am able to account for most of my actions by attributing a causal power to my mental properties, of which I am directly aware. The most natural way to explain similar actions of other human beings is to attribute to them analogous mental properties with analogous causal powers. The situation is different as regards Roby's similar actions. Since Roby is a human artifact, we can apply to it what we could call *Vico's principle*, according to which *we can be certain of what we have made.*[6] Vico's principle guarantees that Roby's behavior can always be explained in terms of its physical properties, i.e., we can always assume Dennett's *physical stance* with respect to it.[7]

Vico's principle thus *suggests* that Roby has no mental properties: we did not really "put them in", we just made Roby behave *as if* it had them, by endowing it with certain physical properties, which are sufficient to explain its behavior. This can be made more cogent by appealing to what Kim has called "the principle of explanatory exclusion", according to which "there can be at most one *complete* and *independent* explanation of a single explanandum".[8] This principle is related to the traditional Ockham's razor.

According to Sloman,[9] many opponents of strong AI argue against a straw man in that they assume an excessively naive version of strong AI. According to this naive version, "there is an algorithm, the UAI,[10] any instantiation of which is sufficient for the existence of mental processes".[11] Thesis (1), as here understood, is not meant to involve this assumption. Sloman is inclined to accept a version of strong AI based on the functionalist-oriented thesis that "there is a collection of computational processes such that if they run on some distributed collection of processors they will produce

[6]The Italian philosopher G.B. Vico (1668-1744) centered his epistemological views around this idea. Vico's principle has been seen as a ground for the relevance of AI to the scientific study of mind (Genesereth & Nilsson, *Logical Foundations of Artificial Intelligence* pp. 1–2).

[7]Dennett, *Brainstorms*.

[8]Kim, "Explanatory Exclusion", p. 41.

[9]Sloman, "The Emperor's Real Mind".

[10]The "Undiscovered Algorithm for Intelligence"; Sloman, "The Emperor's Real Mind", p. 360.

[11]Sloman, "The Emperor's Real Mind", p. 367.

mental states".[12] Some such hypothesis – indeed advocated by many AI theoreticians – could be built into (1), if desired. In general, the standpoint that I am going to outline is defensible as long as the physical stance, as granted by Vico's principle, can in principle be adopted with respect to Roby.

3 Computational Functionalism

Apart from behaviorism, functionalism is the current theory of mind that most closely supports strong AI and thesis (2) in particular. According to functionalism, mental properties are to be viewed as the occupants of causal roles specifiable in terms of some principles of psychology. Since the principles of psychology are meant to be hardware independent, functionalism has been taken to imply the rejection of the traditional *type identity theory*, according to which types of mental states are to be contingently identified with types of physical states.[13] This supports the strong AI program, and in particular the inference from (1) to (2), provided of course that a computational version of functionalism is viable. This presupposes that the principles of (human) psychology be algorithmic in nature.

Functionalism has undergone various attacks. Typically, these aim at showing that the principles of (human) psychology, as understood by functionalism, can be satisfied by organisms with different mental properties (*inverted qualia objection*), or even without any genuine mental properties (*absent qualia objection*), and thus do not capture the essence of the mental.[14] The possible defense of functionalism is well synthesized by Churchland:[15] as against the absent qualia objection, functionalism should reply that qualia are the "inevitable concomitant" of the fact that an organism satisfies the principles of psychology, no matter what its hardware is; as against the inverted qualia objection, functionalism should argue that qualia are not captured by functional definitions, because they depend on the specific hardware on which the psychological principles are implemented. In other words, according to functionalism, if the principles of psychology are implemented in a piece of hardware, then this piece of hardware has qualia. But the nature of the qualia depends on the nature of the hardware. Let us call this claim *the inevitable concomitance thesis.*

[12]Sloman, "The Emperor's Real Mind", pp. 375, 378. Sloman (p. 378) parenthetically remarks that perhaps some non-computational mechanisms are required. I shall neglect this, as it undermines the essence of strong AI.

[13]The typical functionalist is however committed to the token identity theory. I shall not discuss here the plausibility of this theory.

[14]See Bealer, "Mind and Anti-Mind", for a global attack along these lines and for references.

[15]Churchland, *Matter and Consciousness* pp. 40–42.

4 The Identity Theory and Property Dualism

The inevitable concomitance thesis can be supported by associating functionalism to a relativized version of the type identity theory, according to which restricted types of mental properties are to be contingently identified with restricted types of physical properties. Thus, for example, human pain would be an "organic physical property", whereas inorganic (artificial) agent pain would be an "inorganic physical property". A reappraisal of the type identity theory in this spirit is presented by Kim, who speaks of it as *the multiple realisability thesis.*[16]

Prima facie, empirical introspective data strongly support *property dualism*, i.e., the position that mental and physical properties are distinct, albeit causally correlated. Hence, in a dispute between an identity theorist and a property dualist the burden of proof lies with the former. Typically, the traditional identity theorist invites us to endorse her position by suggesting an analogy between the identification of mental properties with physical (brain) properties – which future scientific progress should insure – and other identifications already asserted by science. Typical examples are the identification of the temperature of a gas with its mean kinetic energy, or the identification of water with H_2O. At the present state of our knowledge and capacity for understanding, the best that the identity theorist can hope for in light of the current state of neurophysiology is an ever richer series of nomologically valid biconditionals correlating mental and physical events, to be viewed also as "bridge laws" correlating neurophysiology and psychology. Let us call them *mental-physical bridge laws.* I would say that the identity theory invites us to see these bridge laws as brute facts that cannot further be explained.[17]

Such a view lends plausibility to (2): if the mental-physical bridge laws cannot further be explained, we seem to have no reason to suppose that they hold only for biological hardware. If this is true, one might argue that these brute facts are brought about as soon as the right network of functionalist principles of psychology is implemented in whatever kind of hardware. But then Vico's principle can no longer be invoked to reject (2): even though in constructing Roby we did not "directly put in" mental properties, they would be, as a matter of brute fact, the inevitable concomitant of the physical properties we endowed Roby with.

Many philosophers have pointed out that even on the assumption that all the presently conceivable mental-physical bridge laws had been determined, there would still remain an "explanatory gap" between the mental and the physical which is in the way of considering such laws as a reduction of the mental to the physical.[18] McGinn has even argued that we are intrinsically

[16] Kim, "Multiple Realization".
[17] Kim, "Supervenience as a Physical Concept", p. 26.
[18] Levine, "Materialism and Qualia"; Nagel, "What is it Like to Be a Bat?" and *The View from Nowhere*; Kim, "Concepts of Supervenience" p. 173.

unable to fill this gap.[19] I think that the explanatory gap thesis is strongly supported by the property dualist's typical rejoinder to the identity theorist, which consists in showing that the latter's analogies fail to shed light on the plausibility, or even the intelligibility, of the type identity theory.

The strategy of rebutting the identity theorist's analogical argument is followed by Kripke,[20] but his argument is heavily dependent on his debatable theory of rigid designators. A more theory neutral application of this strategy is provided by Thomson, who considers a long list of potential candidates for the analogy, including the ones already mentioned. I find Thomson's line of attack in general convincing, albeit weak regarding the temperature case, for she fails to appreciate that, e.g., feeling pain can be seen as a property of things just as temperature can.[21] I shall thus concentrate on the temperature analogy by providing an argument of a nature quite different from Thomson's.

The identification of temperature with mean kinetic energy can roughly be described as follows.[22] In thermodynamics the concept of temperature is operationally understood approximately along these lines: *the x whose presence in higher or lower degrees causes corresponding variations in a thermometer.* Thermodynamics postulates that there is such an x in order to explain a number of phenomena (including the thermometer's variations). On the basis of such phenomena it formulates general laws which are empirically supported. Statistical mechanics, on the other hand, has different theoretical and observational tools that could in principle allow for a definition of the concept of gas mean kinetic energy in a way which is quite independent from the thermodynamic concept of temperature. With some inessential simplification such a definition could go as follows: *the y whose presence in higher or lower degrees causes corresponding variations in the "measurement tools" of statistical thermodynamics.* It is then systematically observed that a thermodynamically measured variation in a gas temperature is proportional to the variation in mean kinetic energy of the same gas, as measured with the tools of statistical mechanics. This justifies, together with other complex considerations of a methodological and formal nature, the contingent identification of gas temperature with gas mean kinetic energy. This means that, although the two concepts in question are logically distinct, the x which causes the thermometer's variations and the y which causes parallel observations based on the methods of statistical mechanics are identical. Let us record the following fact: *temperature, as understood by thermodynamics, is a theoretical entity which is not directly observed. The thermometer's variation is the directly observed phenomenon that grounds the postulation of temperature.*

Let us now consider a hypothetical attempt to identify the mental

[19] McGinn, "Can We Solve the Mind-Body Problem?"
[20] Kripke, "Identity and Necessity"
[21] Thomson, "The Identity Thesis" p. 221.
[22] See Nagel, *The Structure of Science* for more details.

property of feeling pain with, say, the physical property of having P-fibers firing. Let us imagine that neurophysiology can avail itself of an instrument for the measurement of P-fiber firing. Let us further suppose that a neuroscientist applies such an instrument to her own brain and finds out that the intensity of pain which she is experiencing is systematically proportional to the amount of P-fiber firing in her brain. If the analogy with the temperature case could go through, the scientist should be able to conclude that feeling pain is contingently identical with having P-fibers firing. Nevertheless, this conclusion is not legitimate. The analogy with the temperature case breaks down because feeling pain, contrary to temperature, is not a theoretical entity, but an observable phenomenon. On its basis, we could define a concept, let us call it *sh-pain*, roughly as follows: *the x that causes pain.* In view of her experiment the neuroscientist could perhaps conclude that sh-pain is contingently identical with P-fibers firing. Sh-pain would thus be reduced, on analogy with temperature, to a concept of a more mature science, such as a future neurophysiology. In other words, feeling pain plays the role of the thermometer's variations, whereas sh-pain plays the role of temperature, i.e., the theoretical entity that could be "reduced".

5 Parallellism and Epiphenomenalism

A pervasive system of mental-physical bridge laws holding for humans is compatible not only with the identity thesis but also with parallellism, epiphenomenalism, a non-reductionist form of supervenience of the mental on the physical, or even Cartesian dualism. However, for the reasons outlined above, only to the extent that these views lend support to a conception of the mental-physical bridge laws as brute facts do they in turn lend support to (2).

In any case, as regards epiphenomenalism and parallellism, these views can be questioned on the basis of the aforementioned principle of explanatory exclusion, and its keen Ockam's razor. For example, given the antisolipsistic hypothesis, subjective mental states are to be considered causally efficacious in explaining human actions. Hence, it would go against Ockham's razor to postulate parallel physical properties with the same causal role. Ockam's razor is violated even more clearly if mental properties are viewed as epiphenomenal. Conversely, in view of the physical stance, granted by Vico's principle, Ockham's razor suggests that we should not postulate mental properties that have, so to speak, an existence running "in parallel" with – or epiphenomenally "over and above" – the physical properties that are by themselves sufficient to explain Roby's actions.

6 The Supervenience of the Mental

In a series of papers,[23] Jaegwon Kim has formulated various notions of supervenience and has examined the thesis that a materialist non-reductionist view of the mind-body relationship can be framed in terms of them. The basic idea here is that mental properties are not identical to physical properties, although the former are supervenient or dependent upon the latter, i.e., their *subvenient base.* The mental-physical bridge laws that the neurophysiologist aims at are to be viewed as supporting such a dependence relation rather than the identity thesis. We should now consider whether this relation can also help support the strong AI thesis (2).

The thesis that the mental supervenes on the physical must face the allegation that it is just another route for epiphenomenalism. For example, Heckmann seems to view this conclusion as inescapable.[24] His argument depends on a view of intentional ascription based on broad content. I shall not follow Heckmann on this line, as the supervenience thesis has to do primarily with narrow content ascription.[25] On different grounds, Kim has expressed doubts on the plausibility of what he describes as *emergentism*, i.e., essentially, a non-reductionist, non-epiphenomenalist, anti-Cartesian view of the supervenience of the mental; he presents emergentism as follows:[26]

> The fact that mentality has emerged, on the emergentist view, must make a *genuinely new causal difference to the world.* So the following summarizes the heart of the emergentist doctrine on mental causation: *mentality must contribute genuinely new causal powers to the world, that is, it must have causal powers not had by any physical biological properties, not even by those from which it has emerged.*

According to Kim,[27] this implies the emergentist's commitment to *downward causation*, i.e., to a causal connection whereby the fact that certain emergent or supervenient properties are instantiated is responsible for the fact that certain other properties of the subvenient base are instantiated.

According to Kim, this position, involving a combination of "upward determination" and downward causation results in a sort of paradox which threatens its consistency: "mentality emerges out of... the physical, and... in spite of this ontological dependence, it begins to lead a life of its own".[28] He goes on to "challenge... the [non-epiphenomenalist] non-reductivists... to state an alternative principle on just how the causal powers of a realized property are connected with those of its realization base".[29]

[23] An interesting critical examination of Kimian supervenience (as defined in Kim, "Concepts of Supervenience") can be found in ch. 2 of Castañeda, *Intentionality, Modality and Supervenience.*

[24] Heckmann, "How Not to Make Mind Matter".

[25] See Kim, "Psychophysical Supervenience", pp. 60 ff.

[26] Kim, "'Downward Causation' in Emergentism", p. 135.

[27] Kim, "'Downward Causation' in Emergentism", p. 136.

[28] Kim, "'Downward Causation' in Emergentism".

[29] Kim, "The Non-Reductivist Troubles with Mental Causation", p. 209.

To give an account of downward causation ultimately means that the mental-physical bridge laws are not viewed as brute facts. This amounts to telling a "story" regarding the causal path from the mental to the physical. And presumably a similar "story" could then be told regarding the causal path from the physical to the mental.[30] By Vico's principle, we should not be able to tell the same story for Roby. But if Kim's challenge cannot be taken up, any attack against the identity thesis, or more generally against the view of the mental-physical bridge laws as brute facts loses some of its strength, and conversely the plausibility of a functionalist inevitable concomitance thesis, supporting (2), is increased.

I shall take up Kim's challenge by outlining an account of the psycho-physical relation based on the idea of causation as energy and information transfer.

7 Causation as Energy and Information Transfer

The mind-body problem depends crucially on the notion of causality. Yet it is typically discussed without a specific account of causality in the background. There is an ordinary notion of causality, which philosophers have tried to reconstruct essentially with the tools of conceptual analysis. But it can also be inquired whether the ordinary notion of causality can be contingently identified with some notion provided by scientific investigation. Such an identification might have a feedback on the mind-body problem. I know of two such attempts to identify the causal relation: Castañeda and Fair.[31] Both consider (and Fair more explicitly proposes) the identification of causation with energy transfer, whereas energy is understood in terms of modern physics.

Castañeda focuses on framing the conceptual background behind the identification of causation with energy transfer. In particular, this involves specifying the notion of *causity*, i.e., something in principle measurable that can flow from cause to effect. Amounts of causity should be assignable to sets of properties, whereas the amount assigned to epiphenomenal properties, if any, would be zero.[32] He suggests that causity *could* be identified with "what scientists nowadays call *energy*" and then warns us that the identity of energy and causity would be "a *contingent identity*... [and] a topic for scientific investigation".[33]

Fair focuses instead on the empirical evidence that currently supports

[30] If this story were available, it could also be used for a version of epiphenomenalism in which the physical-mental bridge laws are not brute facts.

[31] Castañeda, "Causes, Energy and Constant Conjunctions"; Fair, "Causation and the Flow of Energy".

[32] Castañeda, "Causes, Energy and Constant Conjunctions", p. 99

[33] Castañeda, "Causes, Energy and Constant Conjunctions", p. 95. The late professor Castañeda gave me in 1984 a manuscript, entitled "Causes, Causity, and Energy", that differs from "Causes, Energy and Constant Conjunctions" in some interesting details. In particular, the manuscript seems to suggest with somewhat more confidence the contingent identity of causity and energy.

the contingent identity of causation and energy transfer.[34] Fair speaks of energy as something that can be transferred from an object to another. This is not incompatible with assigning energy primarily to sets of properties rather than to objects, in that energy can be seen as accruing to objects insofar as they have certain properties. The two papers can thus be seen as complementing each other. It will be convenient in the following to stick to Castañeda's terminology involving causity (= energy) as something that can be associated to (sets of) properties. This in fact is in tune with the above description of downward causation.

Both Castañeda and Fair briefly touch on the relevance of their views for the psycho-physical relation. According to Fair, the theory of causation as energy flow requires that in general A causes B only if there is a "physical redescription" of A and B. This means that in the case of mental causation the theory "awaits a very detailed neurophysiology". Given this, "if physical theory is descriptively and explanatorily adequate, the connection [e.g. between a mental event and an ensuing action] will be one of energy flow."[35] But then Fair goes on to suggest that "the theory that causation is a matter of energy flow might be compatible with certain forms of dualist interactionism; the reduction of the mental to neurophysiology is not essential.[36] Given the problems surrounding the reductionist position based on the identity thesis, we should take this possibility seriously, although Fair does not unfold his suggestion.

Fair's suggestion is however expanded a bit by Castañeda in the following hypothesis:

> Perhaps what science nowadays calls energy is all there is to causity. But perhaps there are still unknown forms of energy, i.e., of causity. And perhaps there are further forms of causity that should not be called energy because of their anomalous or very different properties.[37] Perhaps psychophysical interaction involves a form of causity which is not physical energy of the well-known types.[38]

Consider a prototypical case where physical to mental causation seems to be involved: a bee, stinging you. We might suppose that, as consequence of the sting some energy will flow through your afferent nerves to the brain up to a point where the causity = energy of a network of physical properties of the brain will increase. This is the proximate cause[39] of your pain *qua* genuine mental property. The pain is brought about, let us suppose, because some causity is transferred from the set of brain physical properties in question to a set of mental properties of your brain. In this process, the causity

[34] Fair, "Causation and the Flow of Energy". To be more precise, Fair speaks more generally of causation as energy or momentum transfer.

[35] Fair, "Causation and the Flow of Energy" p. 236.

[36] Fair, "Causation and the Flow of Energy" p. 237.

[37] This idea might possibly be connected to Davidson's anomalous monism.

[38] Castañeda, "Causes, Energy and Constant Conjunctions", p. 95.

[39] In Ducasse's terminology; see Ducasse, *Nature, Mind and Death.*

has presumably been transformed.[40] Perhaps it is no longer describable as energy as understood by physics today, but whatever it is now, *qua* causity, it is the same amount. This is an empirical hypothesis. Suppose we can measure the amount of causity = physical energy in the brain. We might imagine that the amount is n immediately after the sting and just before the pain. We might then find out that it is $n - \Delta$ at the moment in which the pain ensues. We might therefore assume that mental energy has increased by Δ (even if we are not directly able to measure this).

Consider now mental to physical causation. The pain immediately causes you to retract. We might suppose that some quantity $\Delta' \leq \Delta$ of causity flows from the mental to the physical properties of the brain, thereby being transformed into energy as known by physics today. Again, this could in principle be empirically verified.

Mental states convey information. This is particularly obvious if we consider a mental property such as that of confronting a visual field. Hence, psycho-physical and physico-psychical energy transfer should be viewed also as information transfer. Psychical energy could perhaps be conceived as a peculiar "structurable substrate" for the representation of information.[41] Jackendoff hypothesizes that the elements present to awareness are "projected" from information structures of the computational mind physically implemented in the brain.[42] On the basis of a sort of phenomenological analysis of the elements of awareness, he investigates how information must be physically and computationally structured for it to give rise to the whole that the elements of awareness compose.[43] On the basis of an epiphenomenalist stand, he assumes that such a whole or any element of it does not have any causal efficacy.[44] Causal efficacy should instead be present at the physical-computational level from which awareness is projected. This might lead one to suppose that there must be a causally efficacious physical structured whole or unitary level which maps onto the epiphenomenal structured whole of awareness.[45] But if epiphenomenalism is ruled out, we have no need to suppose that there is some area of the brain containing structured information that is isomorphic with the information contained in consciousness. Perhaps the information supported by psychic energy is a synthesis of physically supported scattered pieces of information. Such a synthesis would thus give rise to a causally efficacious psychic whole which does not correspond to any specific structured physical area of the brain (if there were such

[40] Castañeda speaks of mental to physical causation as "transforming mental or psychic energy into physical energy"; *Intentionality, Modality and Supervenience*, p. 78. In physical to mental causation the transformation could presumably go the other way around.

[41] Arguably, this "substrate" would have only a subjective reality, and the Berkelian motto *esse est percipi* would hold for it. I cannot elaborate on this here.

[42] Jackendoff, *Consciousness and the Computational Mind* pp. 24, 276.

[43] Jackendoff, *Consciousness and the Computational Mind* part IV.

[44] Jackendoff, *Consciousness and the Computational Mind* p. 267.

[45] Although this might not be Jackendoff's position; see however Jackendoff, *Consciousness and the Computational Mind* pp. 296–301.

a physical area, we would have one aspect of the "redundancy", chastised by Ockham's razor, which afflicts epiphenomenalism). We could speculate that the human brain planning module is somehow able to "read" and take advantage of such a synthesized information which would not be available without psychic energy supporting or realizing it.

This picture is clearly compatible with a Cartesian dualism of mental and physical substances. But I think it is also compatible with a moderate materialistic conception according to which the mental supervenes and depends on the physical.

8 The Physical Closure Principle

Perhaps the fundamental challenge for both substance dualism and emergentist downward causation is that they may not be compatible with the alleged causal closure of the physical world (cf. Kim 1993, p. 209), to be called, for brevity's sake, the *physical closure principle*.[46] Strictly related to it is the alleged law of conservation of energy. We must face up to this challenge with an open mind, given the problems afflicting the alternative positions.

It is thus worth noting that there are good reasons to dismiss a dogmatic attitude toward these principles. In defending a dualist position, Ducasse points out that the law of conservation of energy has the status of a postulate rather than that of an established fact:

> the conservation of energy is a ... *defining postulate* of the notion of "an isolated physical system"... That is, conservation of energy is something one *has to have, if* (as the materialistic ontology of physics and more generally of naturalism demands) one is *to be able to conceive the physical world as wholly self-contained, independent, "isolated"*... Conservation of energy, thus, would be an obstacle to the possibility of psychophysical (and of physicopsychical) causation only if it were known to be a *universal* fact;... But it is not known – only postulated, and postulated only to save the universality of the conservation principle.[47]

Furthermore, in his dialogue with Eccles, Popper points out that the law of conservation of energy is not a problem for the interactionist, given a sufficiently pluralistic view of physics:

> ... [The] openness of the mechanical world to the world of electricity was the main challenge which led to a new reconstruction of physics in which electricity became basic and mechanics derivative with respect to electricity... [But] there is no monistic world of electricity.

[46] Kim, "The Non-Reductivist Troubles with Mental Causation", p. 209, points out that the violation of the physical closure principle is more problematic in the emergentist position than in Cartesian dualism.

[47] Ducasse, *Nature, Mind and Death* p. 241.

There are forces other than the electrical forces; forces such as nuclear forces and weak interactional forces in addition to gravitational forces. Accordingly, we can say that each of the two physical worlds, the mechanical world and the electrical world is, on our current understanding, "open" to at least one other physical world which somehow or other interacts with the mechanical and the electrical world. In other words, modern physics is pluralistic (and the law of conservation of energy had constantly to be generalized whenever the physical world was enlarged). Thus we should not be too worried about a *prima facie* violation of this law: somehow we may be able to smooth it all out.[48]

Is thus the existence of mental causity, as outlined here, compatible with the principle of physical closure? Given the Popperian pluralistic view of physics, the answer may be positive, although this must be reconciled with the fact that *physical* is typically taken to imply *objective*, whereas mental properties have clearly subjective features. If this could be worked out, we could perhaps have a physicalistic world view in which the mental depends on the physical in the supervenientist sense, although the former influences the latter through downward causation. On the other hand, given a less pluralistic notion of *physical*, we are left either with a non-physicalist monism or with some form of substance dualism.

Reasons of time and space prevent me from further investigating this here. For the same reasons I have not considered mind-body theories such as Davidson's anomalous monism, or the eliminativist position championed by the Churchlands. This would have been useful as a more thorough support for my attack on (2), but I must reserve this for some future occasion.

9 Conclusion

I have outlined a plausible picture of the causal role played by mental properties in humans. This picture takes into account the methodological importance of the antisolipsistic thesis and of Ockam's razor and the related principle of explanatory exclusion. It is also compatible with the possibility of a pervasive network of mental-physical bridge laws such as the ones invoked by the identity theorist. Nevertheless, according to the present picture, there would be an explanatory gap between the opposite ends of such "bridges". The above account outlines a way to fill this gap. But this involves the idea of a peculiar causity transfer, which by Vico's principle we know not to be at work in Roby. For example, in all likelihood, Roby would have, in order to pass TTT, a planning module which has access to "synthesized information" analogous to that available to humans through consciousness. But by Vico's principle, we would know that such "synthesized information" would simply be realized or supported by purely physical energy, i.e., by physical properties. In view of this, the prospects for Roby's having mental properties are meager.

[48] Popper and Eccles, *The Self and Its Brain* p. 542

References

Bealer, G., "Mind and Anti-Mind: Why Thinking Has No Functional Definition", *Midwest Studies in Philosophy* 9 (1984), 283–328.

Castañeda, H.-N., "Causes, Energy and Constant Conjunctions", in P. van Inwagen, ed., *Time and Cause*, Dordrecht: Reidel 1980, pp. 81–108.

Castañeda, H.-N., *Intentionality, Modality and Supervenience*, M. J. van den Hoven and G. J. C. Lokhorst, eds., *Rotterdamse Filosofische Studies* XII, Faculteit der Wijsbegeerte, Erasmus Universiteit Rotterdam 1990.

Churchland, P. M., *Matter and Consciousness*, revised edition, Cambridge MA: MIT Press 1988.

Dennett, D., *Brainstorms*, Cambridge MA: MIT Press 1978.

Ducasse, C. J., *Nature, Mind, And Death*, La Salle: Open Court 1951.

Fair, D. "Causation and the Flow of Energy", *Erkenntnis* 14 (1979), 219–250.

Genesereth, M. R. and Nilsson, N. J., *Logical Foundations of Artificial Intelligence*, Los Angeles CA: Morgan Kaufmann 1987.

Harnad, S., "Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem", *Minds and Machines* 1 (1991), 43-54.

Heckmann, H.-D., "How not to Make Mind Matter More or Why Fodor's Cure for Epiphobia Doesn't Work", *Grazer Philosophische Studien* 43 (1992), 101–124.

Jackendoff, R., *Consciousness and The Computational Mind*, Cambridge MA: MIT Press 1987.

Kim, J., "Psychophysical Supervenience", *Philosophical Studies* 41 (1982), 51–70.

Kim, J., "Concepts of Supervenience", *Philosophy and Phenomenological Research* 45 (1984), 153–176.

Kim, J., "Supervenience as a Philosophical Concept", *Metaphilosophy* 21 (1990), 1–27.

Kim, J. 1990a "Explanatory Exclusion and the Problem of Mental Causation", in E. Villanueva, ed., *Information, Semantics, and Epistemology*, Oxford: Blackwell 1990, pp. 36-56.

Kim, J., "Downward Causation in Emergentism and Nonreductive Physicalism", in A. Beckermann, H. Flohr, and J. Kim, eds., *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*, Berlin: de Gruyter 1992, pp. 119–138.

Kim, J., "Multiple realization and the Metaphysics of Reduction", *Philosophy and Phenomenological Research* 52 (1992), pp. 1–26.

Kim, J., "The Non-Reductivist Troubles With Mental Causation", in J. Heil and A. Mele, eds., *Mental Causation*, Oxford: Clarendon Press 1993, pp. 189–210.

Kripke, S., "Identity and Necessity", in M. Munitz, ed., *Identity and Individuation*, New York: New York University Press 1971.

Levine, J., "Materialism and Qualia: The Explanatory Gap", *Pacific Philosophical Quarterly* 64 (1983), pp. 354–361.

McGinn, C., "Can We Solve the Mind-Body Problem?", *Mind* 98 (1989), 349–366.

Nagel, E., *The Structure of Science*, Indianapolis: Hackett 1979.

Nagel, T., "What Is It Like to Be a Bat", *Philosophical Review* 83 (1974), 435–50.

Nagel, T., *The View from Nowhere*, Oxford: Oxford University Press 1986.

Orilia, F. "Intelligenza artificiale e proprietà mentali", *Nuova Civiltà delle Macchine* X n. 2 (38) (1982), 44–63.

Popper, K. R. and Eccles, J. C., *The Self And Its Brain*, Berlin: Springer-Verlag 1981.

Sloman, A., "The Emperor's Real Mind: Review of Roger Penrose's *The Emperor's New Mind: Concerning Computers, Minds and the Laws of Physics*", *Artificial Intelligence* 56 (1992), 355–396.

Searle, J. R. 1980 "Minds, Brains. and Programs", *The Behavioral and Brain Sciences* 3 (1980), 417–457.

Thomson, J. J. 1969 "The Identity Thesis", in S. Morgenbesser, P. Suppes and M. White, eds., *Philosophy, Science, and Method – Essays in Honor of Ernest Nagel*, New York: St. Martin's Press 1969, pp. 219-234.

Wittgenstein's Conception of Philosophy as Grammar

Newton Garver

In his "Notes on Logic", Wittgenstein characterized philosophy as consisting of "logic and metaphysics, the former its basis". In his later work, beginning about 1930, he took philosophy to be a part of grammar, in a way that parallels key ideas of Kant and Aristotle. The later view is obviously different from the earlier one, but it is better seen as a modification of the early view than a total rejection of it. In a somewhat similar way the later view needs to be seen as an extension of ordinary grammar that requires some shift in our conception of that familiar subject, rather than as the invention of a totally new philosophical concept.

In Wittgenstein's work through 1921 what he called "logic" had two faces. It was both the arbiter of meaning and inference and also the subject matter of the *Principia Mathematica.* By 1929 he realized that formal logic could not serve as the arbiter of meaning and inference, at least in part because 'A is orange' entails 'A is not blue', even though the entailment is not formally valid. After a brief flirtation with phenomenology, he settled on grammar as the arbiter of meaning and inference, and hence as providing the all-important critical criterion for his version of critical philosophy.

The grammar in question is not easy to comprehend. It is descriptive rather than normative. It is self-referential, as a criterion of significance in critical philosophy must be. While philosophers have often noted the strong Kantian element in Wittgenstein's later philosophy, including not only the self-referentiality but also the use of "grammatical propositions" to replace both analytic and synthetic *a priori* judgments, Wittgenstein's grammar is naturalistic in a way that resembles Aristotle much more than Kant. Particularly noteworthy is that Wittgenstein begins by describing "language-games" that categorize speech-episodes, in that it is impossible to understand an utterance without understanding which language-game is being played. In this way language-games replace categories, and can be seen as a generalization of Aristotle's account of them. Wittgenstein's remark that "Essence is expressed by grammar" is parallel to the use that Aristotle makes of categories in his metaphysics.

Wittgensteinian grammar consists overwhelmingly of descriptions of language-games, whereas grammarians (or linguists) ordinarily take language-games for granted and confine their descriptions to phonology, morphology, and syntax. Wittgenstein said to Moore that he was using 'grammar'

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 51-61

in its ordinary sense but making it apply to things it did not usually apply to. This is a clever and plausible comment; but it remains in tension with his remark that "the meaning of a word is its use in the language", and requires an explication of the meaning of 'grammar' that goes beyond what linguists are generally comfortable with.

Wittgenstein's conception of philosophy as a kind of grammar therefore incorporates important elements of Aristotelian naturalism, Kantian idealism, and linguistic methodology; but it forces us to rethink each of these three traditions in order to see the use that Wittgenstein makes of them.

This brief introduction to the issues makes it clear that Wittgenstein's conception of philosophy as a kind or branch of grammar has multifarious implications, each of which has subtlety and depth.[1]

Rather than skim over each of them, I propose in this paper to pay close attention to the naturalism inherent in his conception of philosophy as grammar, and especially to the parallels with Aristotle that are involved in it.

Aristotle's *Categories* is an anomalous work. Anyone who wishes to systematize or categorize our thinking about thinking must puzzle about where to put the straightforward and seemingly unobjectionable remarks Aristotle makes, and what to make of his leaving them so obviously incomplete. His students and editors (and perhaps he himself) put them at the beginning of his *Organon*. That is of course the right place, except that putting them there answers none of the questions. In particular it leaves us just as perplexed about what *kind* of activity it is to distinguish categories, whether it is an activity that belongs to metaphysics or to linguistics, and in either case how it relates to the more familiar linguistic domains of grammar, logic, and rhetoric. Perhaps we shall never be able to answer the latter question without revising our conception of those traditional domains, but we certainly need to ask the question. In what follows I shall spell out reasons for thinking that Wittgenstein follows more closely in Aristotle's footsteps than either Kant or Ryle does, and that Wittgenstein's language-games can profitably be seen as a generalization of what Aristotle was doing in the Categories.

If Aristotle's *Categories* provide a classification of things and not of sayings, as is traditionally insisted, the things classified are at any rate 'things that are said' (1^a16). The *Categories* may therefore be regarded as presenting in rudimentary form results that might possibly be appropriately and more completely formulated in terms of current methods of linguistic or grammatical analysis, applied to a level of language or discourse that linguists and grammarians usually ignore. While Aristotle's methods for making his distinctions should not seem strange to a contemporary linguist, linguists do not in fact bother with the distinctions he was making. In that respect – and as traditional metaphysicians would certainly insist – he, like

[1] I have discussed them all at length in various essays in Garver 1994.

Wittgenstein, was applying the concept and method of linguistic description beyond its normal range, thus "making things belong to grammar which are not commonly supposed to belong to it". (M276)

Both the name 'categories', which signifies predications or sayings, and the position of the work at the beginning of the *Organon*, which deals with matters of logic and language, reinforce a temptation to interpret the *Categories* linguistically. Although neither the title nor the position of the work in the corpus is directly due to Aristotle, they do show that the inclination to treat the *Categories* as at least partially linguistic goes back to the very earliest tradition of Aristotelian scholarship. This observation need not trouble philosophers. The determination that the categories can be given a linguistic interpretation – even the conclusion that they are linguistic, Ackrill (1963: 71) and Benveniste (1971, chapter 6) notwithstanding – would not suffice to show that they are not also (in some sense) metaphysical, nor that they are not universal.

The most useful linguistic method to employ in this inquiry is distinctive feature analysis,[2] which has been used in several kinds of linguistic analysis. Passages in the *Categories* can be interpreted as employing a related method, if not an early version of the method itself.

This method is based on a complex presupposition: that nothing is linguistically significant (or real) unless it contrasts with something else, that what it contrasts with is an alternative possibility within a systematic array of possibilities, and that the possible alternatives are determined by binary (sometimes ternary, positive/negative/neutral; or at any rate finitary) alternation along a finite number of dimensions, called features. It is unlikely that all types of phenomena admit of a fruitful distinctive feature analysis. The method does not, for example, seem fruitfully applicable either to mechanics or to formal logic. Admitting of a distinctive feature analysis may be a distinctive feature of some types of linguistic phenomena.

In phonology there are, theoretically, a finite number of articulatory and acoustic dimensions along which spoken sound can vary. In the phonemic analysis of a given language, each phonological dimension is either relevant or irrelevant for the identification of given phonemes, and the relevant dimensions, or features, are either positive or negative. Phonemes can then be regarded as bundles (that is, simultaneous collocations) of distinctive features. The English phoneme /p/, for example, can be described as the simultaneous presence of one set of phonetic features (the positive ones) and absence of another set (the negative ones), with the remaining phonetic features (e.g. aspiration) being nondistinctive or irrelevant. In semantic theory lexical meanings can analogously, though somewhat more precariously, be regarded as bundles of abstract semantic markers. J. J. Katz (1966) has

[2] This method of analysis is due to Roman Jakobson more than anyone else. For an account of the method and its uses, see Jakobson, Fant and Halle 1952; Chomsky and Halle 1968; or Householder 1971. Most recent linguistics textbooks have a discussion of features.

been especially impressive in developing semantics from this perspective.

Aristotle does not define the categories, but he is careful to say what is distinctive about each. Some features, such as contrariety and whether something in the category can be said to be more or less so, are specified either positively or negatively for each category. There seems no difficulty in regarding each of the categories as a bundle of features (some positive and some negative, as in the case of phonemes). What sort of thing the categories are would then depend principally on what sort of features occur in the bundles.

Aristotle's categories are derived from predication; they are the kinds or species of the values of the variables in the form 'X is predicated of some a'. This is not to say that every member of each category can be predicated of something, but only that it must be distinctively involved in such predication and that it is what it is because of this distinct sort of involvement. A "this", for example, cannot be predicated of anything, but it may be the subject of a predication, either as a substance or as something inhering in a substance.

Katz (1966: 224-239) has suggested that Aristotle's categories can be interpreted as abstract semantic markers which (a) are entailed by other semantic markers and (b) do not themselves entail other semantic markers. Even leaving aside epistemological questions that arise about the entailments, Katz's suggestion is implausible. His account does not fit what Aristotle listed as categories, for it seems that *being-in-a-position* entails other markers, and neither *where* nor *when* seems to fit into entailment patterns at all; it gives no place to the features that Aristotle singled out as distinctive, such as whether a predicate admits of a more and a less, or of contrariety; and it presupposes a full-blown logical apparatus instead of providing a basis for it. Since it is difficult to imagine a more sophisticated and detailed elaboration of this proposal than Katz has provided, it is reasonable to conclude that Aristotle's categories are not semantic categories.

Predication, or making truth-claims, is a genus of speech-acts (language-games). Aristotle assumes it can be distinguished from other sorts, such as inferring, praying, commanding, imploring, promising, reciting poetry, and so on. Viewed linguistically, therefore, Aristotle's *Categories* form (in Searle's terms) a small subsection in the general theory of speech-acts or (in Wittgenstein's terms) a partial description of a limited range of language-games.

It is certain that predication is more basic than some other sorts of language acts (such as inferring, which clearly presupposes predication), and there are considerations both from generative grammars and from common sense which suggest (falsely[3]) that it may be the most basic sort of speech-

[3]The background supporting this negative judgment includes Malinowski's insistence on the primacy of phatic communion, in the appendix to Ogden and Richards; Husserl's emphasis on the primacy of prepredicative judgments in *Formal and Transcendental Logic*; and Wittgenstein's introduction of language-games that involve no predication, in the early sections of the *Philosophical Investigations*.

act. What should be granted is that predication is basic to science and other truth-seeking linguistic activities, which no doubt accounts for the prominence Aristotle gave this use of language over the others he recognized. But from a linguistic point of view the very idea of some use of language being basic or foundational is suspect, since its significance would depend on its having been recognized or identified initially as one kind of speech-act among many. Speech acts are distinguished from one another by two sorts of criteria, the circumstances in which they are appropriate and the sort of questions and comments that can be made in response to them.[4] I will call them both 'discourse features', the first having to do with discourse conditions and the second with discourse possibilities. The discourse features that Aristotle cites to distinguish the categories belong mainly to the second group.

Ackrill points out (1963: 79) that "one way in which he [Aristotle] reached categorical classification was by observing that different types of answers are appropriate to different questions". This is true, and useful for seeing the overall design of the *Categories*. But the distinctive features that Aristotle cites are often based on the reverse insight, that different questions are appropriate to different sorts of predication. It may be useful to look at some examples:

(a) "Substance, it seems, does not admit of a more and a less". (3^b33) Suppose X is predicated of some a (someone says 'a is X'); it goes hand-in-hand with X being in the category of substance that no question can be raised whether a is more X than b or less X than a was yesterday. If the question could be raised, the predicate would belong to some other category, where this feature is positive or neutral rather than negative. If someone says "a is more a man than b", the presence of the word 'more' shows the predication to be qualitative rather than substantial, even though 'man' normally signifies a substance.

(b) A substantial predication involves not only *predicating* X of a but also *saying* X of a. The latter (but not the former) carries with it a commitment to predicate the definition of X of a; that is, both the genus of X and the differentia of X are also implicitly predicated of a, when X is *said* of a. This obviously shapes the subsequent discourse possibilities. For example, I can attack a substantial predication by contending that the definition of the predicate does not apply to the subject; but I could not attack a qualitative or quantitative predication in this manner. And conversely, the discourse possibilities determine the category; for if I can impugn a predication by charging that the definition of the predicate does not apply to the subject, then the predication must have been substantial rather than merely qualitative.

[4]This characterization holds for the work of Wittgenstein (PI 1-25, 304) as well as that of Austin 1962 and Searle 1969.

The distinction between predicatingand saying is subtle, but Aristotle seems to have gotten it right.

Each of Aristotle's discourse features governs a specific range of possible subsequent discourse. When a feature is positive, a certain set of responses (questions, challenges, comments, etc.) is open or permitted to predications in that category. When a feature is negative, another set of responses is open or permitted. From this point of view, therefore, categories involve or entail, in addition to distinct discourse conditions (which Aristotle largely ignores), distinct clusters of discourse possibilities.

One advantage of such a linguistic reading is that it brings the discussion of categories into a field of active scholarly research. It thereby makes possible a rational and potentially useful criticism of Aristotle's work. Within his category of substance, for example, discourse features can certainly be found to distinguish substances in the modern sense (gold, coal, mud, water, etc.) both from individuals and from natural kinds (species and genera) – perhaps (as John Corcoran has suggested) making use of the distinction between mass nouns and count nouns. There are of course reservations to be kept in mind. Although predication is a universal speech-act, it is not clear that the discourse features which distinguish the categories are universal; nor is it clear what the import would be of their not being universal. Another ground of caution is that discourse features belong to the domain of rhetoric, whereas the categories have always seemed a matter of logic. A third caution is that the theory of speech-acts, which perhaps has the potential for revitalizing rhetoric in the way that the theory of quantification revitalized logic, is itself in a youthful state, and its precise relation to other branches of linguistics remains as uncertain as does that of Aristotle's categories.

Behind these specific issues there lurks the problem that this reading may turn Aristotle from a philosopher into a linguist or some sort of empirical scientist. It seems a betrayal of Aristotle to interpret him as engaged in an empirical rather than an *a priori* inquiry. Ryle (1971, II.180) puts the point cautiously:

> The danger is, of course, that we shall be taken and shall unwittingly take ourselves to be talking grammar, as if it was all part of one topic to say 'Plural nouns cannot have singular verbs' and 'The dotted line in "... is false" can be completed with "What you are now saying..." and cannot be completed with "What I am now saying..."'.

Kant rises to the defense of philosophy with a more strident comment:

> It has not arisen rhapsodically, as the result of a haphazard search after pure concepts, the complete enumeration of which, as based on induction only, could never be guaranteed. Nor could we, if this was our procedure, discover why just these concepts, and no others, have their seat in the pure understanding. It was an enterprise worthy of an acute thinker like Aristotle to make search for these fundamental concepts. But as he did so on no principle, he merely picked them up as they came his way, and at first procured ten of them, which he called categories (predicaments). Afterwards he believed that he had discovered five others, which he added under the name of post-predicaments. But his table still remained defective. Besides there are to be found in it some modes of pure sensibility (*quando*, *ubi*, *situs*, and also *prius*, *simul*), and an empirical concept (*motus*), none of which has any place in a table of concepts that trace their origin to the understanding. Aristotle's list also enumerates among the original concepts some derivative concepts (*actio*, *passio*), and of the original concepts some are entirely lacking. (KdrV A81=B107)

With respect to these complaints against Aristotle, the key issue is whether or not Aristotle was proceeding on some principle. I have shown that he was, and that the principle is a thoroughly respectable one within linguistics. I suspect Kant would reply that this is not really a *principle*, since (a) it proceeds empirically rather than *a priori*, and (b) it provides no closure, no criterion for saying the list is complete. Rather than challenging these retorts I propose instead to note them as points of contrast with Aristotle, and to further delineate the contrast by considering Kant's categories and Wittgenstein's more Aristotelian language-games.

Kant's alternative, as is well known, is to base the table of categories on his table of judgments (KdrV A70=B95), in which he claims to have presented all the logically possible functions of judgment. The table of categories (A80=B106) is claimed to be derived from the table of judgments, and to be complete because that table is complete. His alternative to Aristotle, like Katz's interpretation of Aristotle, presupposes a full-blown logical apparatus within which and by means of which the categories can be defined. Kant is unquestionably correct in saying that this way of proceeding makes for a neat table and a closed system of categories. It also, alas, changes the project hopelessly, from an empirical description of the ways of words to a logical prescription of what they *must* be. The system is not only neat and in tabular form but is also complete. But this is achieved by completely changing the nature of the enterprise, as Ryle acerbically notes:

> Kant's doctrine of categories starts from quite a different quarter from that of Aristotle, and what he lists as categories are quite other than what Aristotle puts into his index. Kant quaintly avers that his purpose is the same as Aristotle, but in this he is, save in a very broad and vague sense, mistaken. (Ryle 1971, II.176)

Kant's purpose could be described as an ideal language project based on analysis of propositions or judgments, rather than an empirical project based on their use.

Unfortunately Ryle proceeds in a way that has some of the same defects as that of Kant, including the crucial defect of basing categories on logic. He

does not begin with a table of judgments. He begins instead with "sentence-factors" and "proposition-factors", where a factor is some part of a sentence or proposition, whether a word or a phrase. By means of this beginning Ryle avoids having a closed system. But he identifies categories with logical types, and says that a categorial sentence is one that specifies the logical type of some factor. Like Kant, therefore, he identifies categories analytically rather than contextually, and within the framework of logic (which is just assumed) rather than within some prepropositional domain of language such as rhetoric.

Wittgenstein uses the word 'category' sparingly, but it does occur both in the early lectures after his return to Cambridge and in the very late remarks on certainty (M 295; OC 308). Though it is easier to associate these uses with Aristotle than with Kant, Wittgenstein did not wish to resurrect the concept. Instead he developed the concept of language-games, and he placed his discussion of them at the beginning of his *Philosophical Investigations*, as a kind of *organon*. He summarizes some of his points in PI 23:

> But how many kinds of sentence are there? Say assertion, question, and command? – there are countless kinds: countless different kinds of use of what we call "symbols", "words", "sentences". And this multiplicity is not something fixed, given once for all; but new types of language, new language-games, as we may say, come into existence, and others become obsolete and get forgotten.
>
> Here the term "language-game" is meant to bring into prominence the fact that the speaking of a language is part of an activity, or of a form of life.

Here Wittgenstein insists on open-endedness in the list of language-games, one of the features of Aristotle's categories that Kant most roundly condemned. He also emphasises that what is primarily at issue is doing something rather than simply thinking or intending. This emphasis on action is what he had in mind when he quoted (as he did frequently) Goethe's line (from *Faust*, part 1), "Im Anfang war die Tat" (In the beginning was the deed).[5]

Wittgenstein's later work seems centered around "grammar," and by grammar he means the description of these language-games, or the rules implicit in them. As in Aristotle this grammar (description of "things that are said") leads into metaphysics:

> Essence is expressed by grammar. (PI 371)
> Grammar tells what kind of thing anything is. (PI 373)

Perhaps the metaphysical dimension of language-games comes into play most clearly in connection with sensations:

[5] See Winch 1987 for useful further discussion of the significance of this quotation from Goethe for Wittgenstein's later philosophy.

> "And yet you again and again reach the conclusion that the sensation itself is a nothing." – Not at all. It is not a something, but not a nothing either! The conclusion was only that a nothing would serve as well as a something about which nothing could be said. We have only rejected the grammar which tries to force itself on us here.
>
> The paradox disappears only if we make a radical break with the idea that language always functions in the same way, always serves the same purpose: to convey thoughts – which may be about houses, pains, good and evil, or anything you please. (PI 304)

Wittgenstein refers over and over to discourse conditions and discourse possibilities in distinguishing different ways in which language functions. The range of cases he discusses is different from Aristotle's, but there is nothing in what Wittgenstein says or in the way that he proceeds that casts doubt on Aristotle's work. If there are categories of which Wittgenstein would be critical, they are those of Kant rather than those of Aristotle.

Wittgenstein's relation to linguistics is as puzzling as Aristotle's. Both of them discuss language and language use in general terms, and do so with methods that are familiar to linguists. But neither of them discusses the aspects of dimensions of language with which linguists and grammarians are traditionally concerned. They further, and more significantly, depart from classical linguists in that the features to which they refer are inextricably embedded in contexts of human action and interaction, resulting in a contextualism that is only superficially like the analytic methods of traditional linguistics. In the long run the principal advantage of a linguistic reading of Aristotle's *Categories* may be that it opens the way to Wittgenstein's grammar as a generalization of that work.

It is useful to note how Wittgenstein's naturalization of categories in his concept of language-games, although initially a rebuke to Kant, has the additional benefit of allowing a perspicuous rendition of Kant's project for critical philosophy. The central point of critical philosophy is that the critical criteria, which must themselves be subject to criticism (to avoid dogmatism), must be self-referential, or self-referentially certified, if critical philosophy is to avoid the Scylla and Charybdis of dogmatism and endless regress. Kant made a stab in the right direction by distinguishing transcendental from formal logic and by claiming to provide a "deduction" in the jurisprudential rather than the logical sense. Rüdiger Bubner has claimed, usefully, that the key to the transcendental deduction is self-referentiality; but it is none too clear just what is self-referential. Is the rule that concepts that occur in *a priori* judgments must always be schematized itself the schematized use of *a priori* concepts? Perhaps. Schwyzer, however, has concluded, on the basis of detailed examination of the key sections in the *Critique of Pure Reason*, that Kant did not succeed in providing a unified conception of the understanding. If Schwyzer is right, as he seems to be, self-referentiality also fails.

Like Kant, Wittgenstein recognized the importance of self-referentiality. The famous passage at the end of the *Tractatus*, about everyone who understands him recognizing his propositions as nonsense, is an extremely poignant and impressive nod in the direction of Kant. But it is even more obvious in the *Tractatus* than in the *Critique* that what we get is a *failed* self-reference, and that won't do.

The naturalized categories, and the conception of philosophy as a kind of grammar that goes along with them, open the way to a solution of this crucial dilemma of critical philosophy. Describing language-games or categories is a kind of grammar, in the sense that it is a description of forms of language. The critical criteria that emerge from this conception of philosophy suffice to clip the wings of speculative metaphysics; no one who reads Wittgenstein's work can fail to appreciate the trenchant critical tool he makes of grammar. Since describing grammar is itself a language-game, grammar is self-referential, and so is Wittgenstein's critical philosophy. This is a very considerable achievement, sufficient in itself to win Wittgenstein a prominent place in the history of philosophy.

Wittgenstein's view of philosophy as an extension of grammar is a generalization of his earlier view of philosophy as "logic and metaphysics, the former its basis"; logic is still inherent in truth-claims, but truth-claims are only one of countless uses of language. Wittgenstein's grammar is descriptive, and differs from linguistics by focussing on uses of language (language-games), by not being systematic, and by being self-referential. Wittgenstein's view therefore implicitly incorporates elements of naturalism, Kant, and linguistic methodology; but it forces us to rethink what we have traditionally thought about Aristotle, Kant, naturalism, and linguistics, to see their contribution to Wittgenstein's thought and the use that he makes of them.[6]

References

Ackrill, J. L. 1963 *Aristotle's 'Categories' and 'De Interpretatione'*, Oxford: Clarendon.

Austin, J. L. 1975 *How to Do Things with Words.* Ed. J. O. Urmson and M. Sbis. Second edition; Cambridge MA: Harvard University Press.

Benveniste, E. 1971 *Problems in General Linguistics.* Coral Gables: University of Florida Press.

Bubner, R. 1975 "Kant, Transcendental Arguments, and the Problem of Deduction", *Review of Metaphysics* XXVIII: 453-467.

Chomsky, N. and Halle, M. 1968 *Sound Patterns of English.* New York: Harper and Row.

Garver, N. 1994 *This Complicated Form of Life: Essays on Wittgenstein.* Chicago: Open Court.

Householder, F. W. 1971 *Linguistic Speculations.* London: Cambridge University Press.

Husserl, E. 1969 *Formal and Transcendental Logic.* The Hague: Martinus Nijhoff.

Jakobson, R., Fant, C. G. M. and Halle, M. 1952 *Preliminaries to Speech Analysis.* Cambridge MA: MIT Press.

[6]Portions of this paper also appear in my essay "From Categories to Language-games", in Garver 1994; copyright 1993 by Newton Garver and used by permission.

Kant, I. (KdrV) *Critique of Pure Reason.* Tr. N. Kemp Smith. London: Macmillan.

Katz, J. J. 1966 *Philosophy of Language.* New York: Harper and Row.

Malinowski, B. 1947 "The Problem of Meaning in Primitive Languages." Appendix to Ogden and Richards. *The Meaning of Meaning.* 8th edition. London: Routledge, pp. 296-336.

Ryle, G. 1971 *Collected Papers.* Two volumes. London: Hutchinson.

Schwyzer, H. 1990 *The Unity of Understanding: A Study in Kantian Problems.* Oxford: Clarendon Press.

Searle, J. R. 1969 *Speech Acts.* Cambridge: Cambridge University Press.

Winch, P. 1988 *Trying to Make Sense.* Oxford/New York: Blackwell.

Wittgenstein, L. 1958 (PI) *Philosophical Investigations.* Second edition. Tr. G. E. M. Anscombe. Oxford: Blackwell.

Wittgenstein, L. 1969 (OC) *On Certainty.* Oxford: Blackwell. References are to numbered sections.

Wittgenstein, L. 1959 (M) "Wittgenstein's Lectures in 1930-33." In G. E. Moore, *Philosophical Papers*, pp. 252-324. London: Allen and Unwin.

Thinking with a Word Processor

J. C. Nyíri

In a well-known passage of the *Blue Book* Wittgenstein remarks: "We may say that thinking is essentially the activity of operating with signs. This activity is performed by the hand, when we think by writing; by the mouth and the larynx, when we think by speaking." We may, he continues, legitimately employ the expressions "'we think with our mouths', or 'we think with a pencil on a piece of paper'".[1] When one of Wittgenstein's favourite authors, Friedrich Nietzsche, started to use a typewriter and sent some rhymes he produced on it to a friend, the latter – a composer – commented upon the robust language. "Perhaps you will through this instrument even take to a new idiom", the friend wrote; "with me at any rate this could happen; I do not deny that my 'thoughts' in music and language often depend on the quality of pen and paper".[2] To which Nietzsche replied: "You are right – our writing equipment takes part in the forming of our thoughts."[3]

The question I am here asking is: In what ways, if any, are our thoughts affected by the shift from the pen or the typewriter to a word processor? My question is not whether thinking about computers changes the image we have of ourselves; nor indeed whether computers do or do not think.[4] What I *do* ask is: With the word processor becoming our writing instrument, what changes do there occur, if any, in the ways and content of our thinking?[5] In particular, what changes can there be discerned, or expected, in terms of the organization of our ideas, in terms of the organization of our memory – our access to, and summary view of, the ideas available to us – in terms of our concept of time, and in terms of the perception we have of the place and role of our thoughts in relation to the thoughts of others.

I am indebted to the Caledonian Research Foundation and the Royal Society of Edinburgh for a European Visiting Research Fellowship at the Centre for Philosophy and Public Affairs, Department of Moral Philosophy, University of St. Andrews, in Spring 1993, during which time the present paper was written. I am especially indebted to the Director of the Centre, John Haldane, for helpful and stimulating conversations. Also, I would like to acknowledge the encouragement and support extended to me by Stevan Harnad (Princeton).

[1] Wittgenstein, *Blue and Brown Books*, pp.6f.

[2] Heinrich Köselitz to Nietzsche, February 19, 1882: Nietzsche, *Briefwechsel* III/2, p.229.

[3] Nietzsche, *Briefwechsel* III/1, p.172.

[4] I have touched on this problem in my paper "Wittgenstein and the Problem of Machine Consciousness".

[5] See Nyíri, "Historisches Bewußtsein im Informationszeitalter", Heim, *Electric Language*, Bolter, *Writing Space* and *Turing's Man.*

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 63-74

The notion that thinking – both how we think and what we think – is not independent of the concrete linguistic medium in which it unfolds is of course very much in accordance with Wittgenstein's position. Not only does Wittgenstein say: "When I think in language, there aren't 'meanings' going through my mind in addition to the verbal expressions: the language is itself the vehicle of thought", and not only does he point out that what we are concerned with is "the spatial and temporal phenomenon of language",[6] but he also repeatedly stresses, and indeed this is one of his central insights, that the meaning of a linguistic sign depends on the circumstances, the spatial and temporal surroundings[7] in which it occurs; that intention depends on context. However, Wittgenstein does not seem to have been alert to the fact that contexts change with the medium; that "thinking by writing" creates linguistic surroundings radically different from those created by "thinking by speaking". Let me come to my main topic by touching on these differences first.

In order to highlight some characteristic features of thinking by speaking we have to bring to mind conditions where literacy is altogether absent. Our knowledge of such conditions comes from ethnological studies of primitive peoples, and from classical philology reconstructing the poetry of Homeric Greece. If speaking has no recourse to written texts, verbal recollection will have to rely on specific mnemotechnic devices: on rhythm, rhyme, and formulaic repetition. Presenting texts will mean: the stitching together of songs, the producing of poetry out of traditional elements. Verbal knowledge will be preserved, as far as necessary and as far as possible, in handed-down rhythmic formulas.

The singing of oral epic does not, and cannot, amount to the reproduction of a fixed text. There is no original version and no authorship; there is no correctness or incorrectness of recollection, only a greater or lesser mastery in the handling of traditional patterns. Phrases, not words, are the threads out of which texts are woven; phrases are remembered, adjusted; they are not, however, compared to each other; a juxtaposition of spoken utterances is impossible, the idea of textual identity makes no sense, notions of logical coherence and contradiction do not emerge. This is not to suggest that in the Greece of Homer questions as to the accuracy or deceitfulness of utterances, or as to the accord between utterances and deeds, did not arise. It was not feasible, however, to preserve the exact wording of extended texts over longer periods. Records of bygone events did not survive unchanged as time went by; past and present merged.

He who thinks by speaking learns by hearing. His thoughts do not belong to him, they belong to everyone. Homeric Greek has no words to represent mental events; thinking is dialogue, thinking to oneself a dialogue between parts of one's body. There is no vocabulary to express abstract

[6] Wittgenstein, *Philosophical Investigations* §§329, 108.
[7] Wittgenstein, *Philosophical Investigations* §539.

cognitive states or processes. Of Homeric heroes it is not predicated that they possess wisdom, or knowledge; what is said of them is that they are good at counsel, skilled at action, sharp–sighted in spotting danger.

Learning by hearing precludes criticism. The listener's task is to absorb what is being said; doubt should not impede memorizing. Faithful attending is safeguarded by the belief that the text delivered has been handed down unchanged through the generations and is, ultimately, of divine origin. With the development of alphabetic writing and the spread of literacy in Greece, critical thinking emerges. In their written-down form, oral narratives invite reflection and interpretation. Reflection: that is, bending back words and taking a look at them. Interpretation: that is, making one word stand for another. Texts are now compared, contradictions detected. The transition from oral to written formulation gives rise to the genre of the philosophical aphorism: Heraclitus and Parmenides still compose for an illiterate audience, they have to think up memorable formulas, but their thinking already shows the influence of the logic of writing. Written language permits a syntax of abstractions: Plato is dissatisfied with talking about beautiful things, he wishes to know what the beautiful itself is; meanings become an issue. If what is known are abstract objects, the knower, too, is construed as an abstract entity: the notion of a soul with cognitive capacities makes its appearance.

Still, thinking in ancient Greece and Rome, and all through the Middle Ages, remains, predominantly, thinking by speaking and hearing. Silent reading is almost unknown; the written text, devoid of intervals and punctuation, has to be read out loud in order to be understood. Copying by hand is laborious: texts are rare, learning still amounts, on the whole, to listening, the authority of the teacher is scarcely weakened. Texts are interspersed by comments if copied by an expert scholar, impaired by mistakes if copied by an unqualified clerk: copies of the same work increasingly differ from each other, the notion of authorship remains blurred. The decay of texts evokes the notion of ancient truths now lost; the disfigured references to dates, names, and places in historical narratives result in an intermingling of fact and legend.

Only with the advent of the printed book will the cognitive consequences of literacy fully unfold. Printing produces thousands of identical copies; mistakes are, with every new edition, progressively eliminated; a community of scholars all over Europe works on the same texts, gradually establishing a firm framework of categories, names, of historical time and geographical space; descriptions, findings, discoveries can be increasingly compared with each other, maps, diagrams, illustrations, figures and calculations reproduced; the ideal of a unified knowledge, symbolized by the metaphor *library*, emerges.[8] The past is now articulated along a stable chronology, antiquity is seen as forever gone, as something entirely different

[8] See Mersch, *Ariadne im Labyrinth der Zeichen*, especially pp. 106ff.

from the present; modern historical consciousness comes into being. The biographies of different personalities cease to be merged with each other, portraits showing characteristic features are reproduced unchanged over the time, the framework of the modern individual is created. The printed page is easily scanned, silent reading becomes the rule, writing does not have to be transformed into sound in order to be intelligible, texts are *there*, ready to be looked up, knowledge is available, objective, contained in books. At the same time, knowledge is centered around the knower: the scholar is surrounded by his books, they belong to him, they even accompany him on his journeys, an image enhanced by that astounding new invention, the portable book; thinking ceases to be heard, to be public, it proceeds, as philosophy now sees it, in the privacy of the thinker's mind. In Locke's telling formulation the mind is at first a "white paper, void of all characters", onto which experience will "imprint" ideas, while the subsequent operations of the mind then consist in "viewing" its own ideas, in "reflecting" upon them.[9]

Words on the printed page appear clearly and distinctly. The handwritten manuscript however which the author prepares is, as a rule, rather convoluted. If one adds to this that authors tend to experience feelings of intimacy and elation at the sight of their handwriting, the conclusion is difficult to avoid that the writer thinks in a medium less clear than the one in which his readers will think. Fair copies offer only partial remedy; author's corrections in proof are not just second thoughts, but thoughts in a more externalized, more objective setting. With the introduction of the typewriter much of this, of course, becomes past history. The typewritten text is impersonal and perspicuous; ideas generated on a typewriter will easily look distant, even strange, inviting critical scrutiny. Heidegger acts in accordance with his grand strategy of dissolving the subject-object dichotomy in philosophy when he crusades against the typewriter. He admits that typing does have some limited use when serving as transcription and preserving handwriting, but abhors the machine taking the place of the hand, depriving script, as he says, of its essential source, degrading the word to a mere means of communication.[10]

For Nietzsche of course the typewriter had an entirely different significance. He was, from his early youth on, extremely short-sighted, his eyes quickly getting tired and painful, causing terrible headaches. By the end of the 1870s he was practically unable to formulate his ideas in writing. His daily routine was to go for extended walks, immersed in dialogues with himself and thinking up short aphorisms, spoken aloud as Hollingdale assumes,[11] jotting them down in small notebooks, and then at home, from time to time, trying to decipher the scribble and prepare some clean text. There is much he can in the end not copy, especially when it comes to longer remarks.

[9]See Locke, *An Essay concerning Human Understanding*, Book II, ch.1, and Book IV, ch.2.

[10]Heidegger, *Parmenides*, pp. 119 and 125.

[11]Hollingdale, *Nietzsche*, p. 141.

He curses his unwilled telegraphic style and when he learns about the new invention, the typewriter, he jumps at this possibility of writing without looking[12] – "touch typing" we call it, but in Nietzsche's case the German term, "Blindschreiben", is more apt. He receives the machine early in 1882, in Italy; he is delighted, but of course the initial difficulties are frustrating, his spoken-out-loud thoughts are now being written down in a style even more terse. "Wann werde ich es über meine Finger bringen, einen langen Satz zu drücken", he complains to a friend – "When will my fingers enable me to type a long sentence!"[13] Eventually the machine turns out to be a disappointment; the ribbon gets torn, becomes even wet and sticky when the weather is damp, and then one can't see the letters at all. Nietzsche goes on thinking-out-loud a philosophy which, significantly, detects in grammar and language the source of metaphysics, of that "prejudice of reason" forcing us to assume "unity, identity, permanence, substance"; and argues against the "old conceptual fiction that posited a 'pure, will-less, painless, timeless knowing subject'".[14] In Nietzsche's thinking the intuitions of written language are gradually stifled, and the intuitions of spoken language come to the fore.

As I have attempted to show in some previous papers of mine,[15] Wittgenstein's later philosophy, too, reflects the spirit of spoken language, of language voiced and heard. Of course Wittgenstein had no difficulties with writing; indeed he was an obsessive writer.[16] But he did have a problematic relation to written language, especially to written language in its fully developed form: the printed book. Already in the preface to his *Wörterbuch für Volksschulen* Wittgenstein had complained about the distorting effects of typography; and his reluctance to publish his writings is of course notorious. I also have in mind his poor orthography, his anachronistic predilection for having people read texts to him out loud, the common observation that his favourite readings he really knew by heart, the aphorism and the dialogue as conspicuous stylistic features of his writing, and his inability or unwillingness to put together what one would call a treatise in the modern sense. "It was my intention at first", he writes in the preface to his *Philosophical Investigations*, "to bring all this together in a book", with the aim "that the thoughts should proceed from one subject to another in a natural order and without breaks". But he had to realize, he continues, that he would never succeed "to weld [his] results together into such a whole", that even the best he could write "would never be more than philosophical remarks" and that "this was, of course, connected with the very nature of the investigation. For this compels us to travel over a wide field of thought criss-cross

[12]Janz, *Nietzsche* vol. 2, p. 27.

[13]Nietzsche, *Briefwechsel* III/1, p.172.

[14]Nietzsche, *Götzen-Dämmerung*, *Werke* 6, pp. 77f., and *Zur Genealogie der Moral*, *Werke* 5, p. 365.

[15]Nyíri, "On Esperanto" and "Heidegger and Wittgenstein".

[16]Sluga, "Thinking as Writing".

in every direction. – The same or almost the same points were always being approached afresh from different directions, and new sketches made." Wittgenstein's method of composition, as you know, was a rather peculiar one. He first wrote down his remarks in small pocket notebooks. Subsequently he copied them into large manuscript books, making selections and changes in the process. These manuscripts were then again edited, some remarks left out, but the bulk of them, once in a while, dictated to a typist. What now followed was thinking with scissors: Wittgenstein cut up the typescripts, arranged and rearranged the cuttings, having some of the arrangements typed again at some later stage. If at the time the word processor had already been introduced, Wittgenstein could have made good use of it. Or, if I may put it more pointedly: lack of a single perspective in Wittgenstein's later thinking invited a method of composition which today calls to mind the operations of a word processor.

There is a view according to which word processing is just a "refinement of printing", a kind of "glorified typing".[17] I think this view is, ultimately, misleading; but it provides a convenient initial perspective. What are, then, the characteristics of the word processor, regarded as a typing and printing instrument? Bear with me while I summarize the obvious. A text composed on a word processor is revised, edited, formatted and re-formatted, printed, and even published, with very little effort. Writing on a word processor is easy both in the sense of permitting for the provisory, the draft, the experiment, and in the sense of allowing for ready use of bits of texts already there – of one's own texts, or of texts written by others, the latter effortlessly amalgamated with the former. Huge masses of writings, contemporary and classical, become available either on tape, CD-ROM, and disk – like dictionaries and encyclopaedias, the Greek and Latin corpus, English poetry in its entirety, the Musil *Nachlaß*,[18] now even Wittgenstein[19] – or through networks, providing access to databases of various kinds, among them to electronic editions of a growing number of scholarly journals.[20] Networking becomes, increasingly, a matter of course, especially since joining the e-mail community is an unavoidable necessity.

Let me spell out this last point by saying that when we inquire about

[17]This is the view, for instance, of Robert Sokolowski, whose formulations I am here quoting; Sokolowski, "Natural and Artificial Intelligence", p.49. Sokolowski is contrasting the effects of the word processor with the possible future effects of artificial intelligence; in his opinion the former does not, but the latter could, change our ways of thinking. As he puts it: "Thinking is shaped by writing; intelligence is modified when it takes on the written form; writing permits us to identify and differentiate things in ways that were not possible when we could speak but not write. If artificial intelligence can in turn transform writing, it may be able to embody a kind of intelligence that cannot occur in any other way". Sokolowski, "Natural and Artificial Intelligence", p.50.

[18]For some details and assessments see, for instance, the *Times Literary Supplement*, April 30 1993, pp.7ff., and *Die Zeit*, Dec. 4 1992, p.65.

[19]Stern, "Toward a Complete Edition of the Wittgenstein Papers".

[20]For some interesting recent developments see, for instance, the *Times Literary Supplement*, May 14, 1993, p.17.

the cognitive consequences of using a word processor, our questions ultimately relate to the word processor as enmeshed in a network. And let me make the observation that the practice of networking undermines the habit of producing printouts. When paper is not needed to mediate between the writer and his reader, it will be less and less used to mediate between the writer and himself. Clearly, we have come a long way by now from the idea of the word processor as a glorified typewriter. But even the isolated word processor, even with the documents written on it regularly printed out, will give rise to patterns of linguistic behaviour, and indeed to patterns of thinking, that are significantly different from the patterns created by typing and book printing.

The text of a printed book or article is a finished product, is there to be referred to, looked up, read, reflected upon, criticized. The reader may add marginal notes to the text, but he cannot rewrite it. Even its author cannot rewrite it, though of course he may prepare a new edition. The old and the new editions will then exist side by side, ready to be compared with each other. The typewritten text might be a definitive one, as for instance a legal document bearing the necessary signatures, or might have a provisional character, waiting to be revised by its author or a reader. The corrections and alterations will be in handwriting, with the old text usually still discernible under the old. The revised text might, or might not, be re-typed; if it is re-typed, the old type-script is usually kept, and as a rule there are carbon copies both of the old and the new versions.

By contrast, a text on the word processor's display is there to be updated – to be altered, revised. As Richard Dimler has put it: For the user of a word processor, language has "become dynamic rather than static, malleable rather than fixed, soft rather than hard, plastic rather than rigid. As a consequence language never seems to reach a finished stage;;.[21] When a text is changed, the original wording usually vanishes without a trace. It is not there, anymore, on the display; and if the corrections were made in a printout in the first place, the printout is subsequently thrownaway. Of course one can keep old printouts, and of course one can save the older versions of one's files – but there would have to be a special reason for one to do so. With old typescripts by contrast, one has to have a special reason to throw them away. Texts stored in a word processor bear no marks of their history, they are ageless, they possess no temporal existence of their own. And by being subject to continual re-writing, they possess a merely limited objectivity not just of meaning, but of form as well. When deliberated over, they will not be interpreted, they will be altered. Thinking about them is, partly, changing them. They cease to be pure objects of thought; they become the thoughts themselves, thoughts in flux. As Colette Daiute observes, the word processor can "blur the distinctions between thinking, talking, and writing in a way

[21]Dimler, "Word Processing and the New Electronic Language", p. 463.

that the pencil and typewriter [does] not".[22] Or as Michael Heim puts it, "The immediacy of formulation in digital writing is akin to the immediacy of speaking. ... word processing reclaims something of the direct flow of oral discourse."[23] Relieved, to some extent, of the "constraints imposed by the linear order of writing on paper",[24] the writer "can begin anywhere in the text",[25] just like, at the dawn of literacy, Parmenides could: it is all one to me where I begin, he said – "faithfully reproducing", as Havelock puts it, "the plunge that the bard takes into [his] medium".[26] Also, as Dimler already observes, word processing "fosters a modular style in writing", "the writer will be tempted to repeat set formulas and phrases, a linguistic throwback to the ancient 'singers of tales' who used oral formulas as mnemonic devices in recounting the great historic epics".[27]

Just as speaking, as a rule, is less coherent than writing, a text composed on screen tends to be less coherent than a text composed in handwriting or on the typewriter. The reason for this is obvious. Maintaining coherence is a matter of comparing texts with each other, as well as of comparing one bit of a text with other bits of the same text. On screen such comparisons can be executed to a very limited extent only. Depending on the system used and the kind of display available, one, two, or even more documents can be viewed simultaneously; but of each document only a small segment will be exposed at a time. Comparison of segments of texts – their juxtaposition – is of course becoming less awkward as programs allowing for a flexible use of so-called "windows" are increasingly available. Working with windows does indeed resemble working with sheets of paper – but the resemblance is confined to narrow limits. A synoptic view of all accessible and relevant documents, or even of a single extended document, is not possible to attain. Contradictions become difficult to spot; the unity of a text difficult to sustain. A decrease in logical rigor is the inevitable consequence. On a pedestrian level, publishers now learn to be prepared for novel types of authors' mistakes, generated by the use of word processors – like, for example, paragraphs having been moved in such a manner that the result is nonsensical,[28] or like the same paragraph repeatedly occurring, having been copied to more than one place in the text. This is the surface. At a deeper level one might perhaps formulate the preliminary conclusion that thinking with a word processor combines the characteristics of both pre-literal and literal thought patterns. It is fluid, fragmentary, formulaic, with no unity of perspective, and a diminishing sense of the self. At the same time it can rely on texts – on an immense mass of texts – that are there to be looked up.

These characteristics are vastly amplified when the word processor

[22] Daiute, *Writing and Computers*, p.vi.
[23] Heim, *Electric Language*, pp. 154, 209.
[24] Daiute, *Writing and Computers*, pp. 97f.
[25] Heim, *Electric Language*, p. 207.
[26] Havelock, "The Alphabetization of Homer", p. 179.
[27] Dimler, "Word Processing and the New Electronic Language", p. 464.
[28] Hodgkin, "New Technologies in Printing and Publishing".

becomes connected to a network. The basic form of networking is e-mail, and it is fascinating to observe how closely the style of e-mail messages tends to resemble that of spoken language. E-mail texts abound with false starts and incoherent sentences. The apparent reason for this is that e-mail letters are, commonly, written in software surroundings which allow for corrections only to a very limited degree. The resulting text then contains any number of mistakes, mostly innocent, but sometimes amounting to a truly Freudian spectacle – and still the message gets dispatched, because one is in a hurry and does not want to start all over again. And here I think the element of hurrying constitutes the essential reason, and the limitations in editing are merely a corollary. After all, as you of course know, e-mail messages can indeed be composed in subtle word processing surroundings, permitting any degree of careful consideration; but switching to an appropriate text editor takes time. Instead, the technical possibility of sending off an e-mail letter on the spot will be seized on. You read your mail and answer it – on the spur of the moment, just like in a conversation. Incidentally, e-mail exchanges tend to represent a particularly rude type of conversation. As a brief survey in 1985 formulated: "Electronic mail promotes a confrontational style – you get angry and you whip off a message – something you would not say face to face because it is impolite". [29]Or, as a newsletter of the University of Pittsburgh put it: "A good thing to remember is that there is a human being behind that network username and e-mail address. It is easy to tear into a faceless target, especially when you can hide behind an impersonal computer interface."[30]

With networking, the body of information accessible, of texts there to be scanned and used, is growing at a tremendous rate and will in time, no doubt, be all-inclusive. However, knowledge yielded by search processes is, inevitably, made up of disconnected elements. Typically, one has an incomplete notion of what one is looking for; and what one then finds is again incomplete, lacking context. The metaphor of the library will not fit here, the images of outline, orientation, browsing are not applicable. As Heim puts it: "Textual database searches conceal as well as reveal what it is we learn."[31] In a way this is true of computer memory generally: it buries as well as stores what we have learned.

Networking radically blurs the notion of individual authorship. Already at the level of simple word-processing, cooperative writing is easy, co-authors can readily revise and complement each other's texts. With networking, one's ideas emerge and evolve in surroundings in which the ideas of many other persons are incessantly and actively present, affecting one's ideas, and themselves being affected by them. As Stevan Harnad, the editor of *Behavioral and Brain Sciences* emphasizes with prophetic zeal, the pursuit, as well

[29] *Chronicle of Higher Education*, October 2 1985, p. 32.
[30] *Connections!, a University of Pittsburgh Computing and Information Services Newsletter*, November/December 1992, p. 17.
[31] Heim, *Electric Language*, p. 214.

as the dissemination, of scientific knowledge is thereby substantially restructured. "The whole process of scholarly communication", Harnad writes, "is currently undergoing a revolution comparable to the one occasioned by the invention of printing... The potential role of electronic networks in scientific publication... goes far beyond providing searchable electronic archives for electronic journals."[32] Harnad lists the cognitive revolutions of the emergence of speech, the advent of writing, and then the invention of the printing press. "All three", he says, "had a dramatic effect on how we thought as well as on how we expressed our thoughts, so arguably they had an equally dramatic effect on what we thought."[33] But to–day, Harnad stresses, a "fourth cognitive revolution" is imminent:

> e–mail networks becoming the carriers of that vast prepublication phase of scientific inquiry in which ideas and findings are discussed informally with colleagues (currently in person, by phone and by regular mail), presented more formally in seminars, conferences and symposia, and distributed still more widely in the form of preprints... It has now become possible to do all this in a remarkable new way that is not only incomparably more thorough and systematic in its distribution, potentially global in scale, and almost instantaneous in speed, but... unprecedentedly interactive...[34]

Scholarly inquiry in this new medium, called "scholarly skywriting" by Harnad, "is likely to become a lot more participatory, though [it will become] perhaps also more depersonalized, with ideas propagating and permuting on the net in directions over which their originators would be unable (and indeed perhaps unwilling) to claim proprietorship."[35]

These are, then, some of the ways in which our thinking changes when we are thinking with a word processor. But what is it really, I would like to ask by way of conclusion, we think "with" when we think with a word processor? Here the Wittgenstein passage I quoted in the beginning was, if you will forgive me, chosen not so much to enlighten, as rather to set the stage. Thinking, in the view of the mature Wittgenstein, is not done with anything; it is not an activity at all. Hacker is entirely right when he says that "[the] parallels between the grammar of thinking and the grammar of activities are misleading, for they induce us to overlook important differences. Nothing need go on when one thinks."[36] This is, after all, one of the fundamental Wittgensteinian discoveries: that mental phenomena cannot be identified independently of *Umstände*, of the broad story within which they occur; that, as I stressed earlier, intention depends on context. So what are the characteristics of the context, of the circumstances, under which we say

[32]Harnad, "Scholarly Skywriting".
[33]Harnad, "Post-Gutenberg Galaxy".
[34]Harnad, "Scholarly Skywriting".
[35]Harnad, "Scholarly Skywriting".
[36]Hacker, *Wittgenstein: Meaning and Mind*, p.304.

that we are thinking – with a word processor? What kind of language game is: "thinking with a word processor"? I have tried to outline an answer in the foregoing: When we think with a word processor it is a synchronous intellectual exchange with fellow thinkers all over the world we are, ultimately, engaged in. So what are we thinking with when we think with a word processor? The word "with" here, I conclude, does in the last analysis point not to instrumental application – but to human companionship.

References

G. Baumann (ed.), *The Written Word: Literacy in Transition*, Oxford: Clarendon 1986.

Jay David Bolter, *Turing's Man: Western Culture in the Computer Age*, Chapel Hill: University of North Carolina Press 1984.

Jay David Bolter, *Writing Space: The Computer, Hypertext, and the History of Writing*, Hillsdale, N.J.: Lawrence Erlbaum 1991.

Roberto Casati and Graham White (eds.), *Philosophy and the Cognitive Sciences: Papers of the 16th International Wittgenstein Symposium*, Kirchberg am Wechsel: The Austrian Ludwig Wittgenstein Society 1993.

Colette Daiute, *Writing and Computers*, New York: Addison–Wesley 1985.

G. Richard Dimler, S.J., "Word Processing and the New Electronic Language", *Thought* vol. 61 no. 243, December 1986.

P.M.S. Hacker, *Wittgenstein: Meaning and Mind. An Analytical Commentary on the Philosophical Investigations* 3, Oxford: Blackwell 1990.

Rudolf Haller (ed.), *Wittgenstein – Towards a Re-Evaluation*, Wien: Hölder-Pichler-Tempsky 1990.

Stevan Harnad, "Scholarly Skywriting and the Prepublication Continuum of Scientific Inquiry", *Psychological Science* 1 (1990).

Stevan Harnad, "Post-Gutenberg Galaxy: The Fourth Revolution in the Means of Production of Knowledge", *Public–Access Computer Systems Review*, 1991.

Eric A. Havelock, *The Literate Revolution in Greece*, Princeton: Princeton University Press 1982.

Martin Heidegger, *Parmenides*, Frankfurt: Vittorio Klostermann 1982.

Michael Heim, *Electric Language: A Philosophical Study of Word Processing*, New Haven: Yale University Press 1987.

Adam Hodgkin, "New Technologies in Printing and Publishing: The Present of the Written Word", in Baumann (ed.), *The Written Word* pp.151–169.

R.J. Hollingdale, *Nietzsche: The Man and His Philosophy*, London: Routledge & Kegan Paul 1965.

Curt Paul Janz, *Friedrich Nietzsche: Biographie*, vol. 2, München: Carl Hanser 1978.

John Locke, *An Essay concerning Human Understanding*, Oxford: Clarendon 1894.

Dieter Mersch, *Ariadne im Labyrinth der Zeichen: Semiotik, Rationalität und Rationalitätskritik bei Umberto Eco*, Dissertation, Darmstadt 1993.

Dieter Mersch, J.C. Nyíri (eds.), *Computer, Kultur, Geschichte: Beiträge zur Philosophie des Informationszeitalters*, Wien: Edition Passagen 1991.

Friedrich Nietzsche, *Briefwechsel*, in *Kritische Gesamtausgabe*, Berlin: Walter de Gruyter 1975–.

Friedrich Nietzsche, *Samtliche Werke: kritische Studienausgabe*, ed. Giorgio Colli, Mazzino Montinari, Berlin: de Gruyter, 1988.

J. C. Nyíri, "On Esperanto: Usage and Contrivance in Language", in Haller (ed.), *Wittgenstein* II, pp. 303–310.

J. C. Nyíri, "Heidegger and Wittgenstein", in Nyíri, *Tradition and Individuality.*

J. C. Nyíri, "Historisches Bewußtsein im Informationszeitalter", in Mersch, Nyíri (eds.), *Computer, Kultur, Geschichte*, pp. 65–80. English translation in Nyíri, *Tradition and Individuality.*

J. C. Nyíri, *Tradition and Individuality: Essays*, Dordrecht: Kluwer 1992.

J. C. Nyíri, 'Wittgenstein and the Problem of Machine Consciousness", *Grazer Philosophische Studien* 33/34, 1989.

Robert Sokolowski, "Natural and Artificial Intelligence", *Daedalus* Winter 1988.

Hans Sluga, "Thinking as Writing", *Grazer Philosophische Studien* 33/34 (1989), pp.115–141.

David G. Stern, "Toward a Complete Edition of the Wittgenstein Papers: Prospects and Problems", in Casati and White (eds.), *Philosophy and the Cognitive Sciences*, pp.501–505.

Ludwig Wittgenstein, *The Blue and Brown Books*, Oxford: Basil Blackwell 1958.

Ludwig Wittgenstein, *Philosophical Investigations*, second edition Oxford: Blackwell 1959.

Wittgenstein, Computationalism, and *Qualia*

Georges Rey

1 Introduction

There can be little doubt that, *temperamentally*, Wittgenstein would deplore recent computational theories of the mind. If his early and late work have nothing else in common, it is their extreme antipathy to what he regards as the shallow scientism that has the modern mind in its thrall:[1] the *Tractatus* sought to show that "what was important" lay on the other side of the expressible; the *Investigations*, that it at least lay in language games other than the standard scientific ones. And certainly he was inclined to include a great deal of what we are ordinarily wont to say about the mind among the things that were important and to be protected in these ways. Computational theories are unlikely to be so reverent.

It is not difficult to sympathize with Wittgenstein's aversion to the hegemony of science in contemporary *culture*, or even with his sense that in our ordinary relations with people we are often better guided by traditional wisdom than by the latest fashions in psychological theory. But it is one thing to think that science shouldn't play certain roles in our lives, quite another to think that there is no place for a science of the mind at all. In the present paper, I want to insist on a form of scientism that Wittgenstein loathed, and argue that he is not entitled to dismiss the kind of

This paper is a short version of a much longer paper, (Rey forthcoming c), that will deal with Wittgenstein's attacks on a science of the mind in much greater detail than was possible here. As well as at the Kirchberg conference, versions of this paper have been presented at the University of Maryland, Baltimore County; King's College, University of London; and at the Australian National University, Canberra. I am greatful to audiences at those institutions, as well as to George Bealer, Frank Dring, Paul Horwich, Naomi Scheman, Julia Tanney, and the editors of the present volume for their reactions and advice. I am especially indebted to lectures and conversations of Rogers Albritton: his reading of Wittgenstein greatly influenced the view defended in the last, if not quite in the earlier, sections of this paper.

[1] His loathing of scientism –and sometimes of science– is vivid, but never so visible as in his recently published *Vermischte Bemerkungen*: "it isn't absurd, e.g. to believe that the age of science and technology is the beginning of the end for humanity... that there is nothing good or desirable about scientific knowledge and that mankind, in seeking it, is falling into a trap". (VB:56) At one point he even thinks the development of the atom bomb is to be welcomed as an antidote to "our disgusting soapy water science" adding that "the people presently opposing the development of the bomb are intellectual scum" (VB:49). Perhaps the title of this volume is part Yiddish? See also the preface to PR, as well as (Baker and Hacker 1983:4) and (Monk 1990:300,484-6,490).

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 75-87

computational theory of the mind that is presupposed in much current cognitive science. To the contrary, some of what he says actually *invites* it. As a critique of a certain common, naive conception of mental phenomena (objects, properties) as phenomena in a peculiarly "private", introspective realm, Wittgenstein's remarks are exceptionally acute and useful in motivating a more theoretical psychology. Taken, however, as he seems to have intended them, as a critique of *any* psychology of inner mental processes, they are stunningly inadequate. Their most glaring omission is of any indication of how, without internal mental states, there is any chance of explaining intelligent regularities in human and animal behavior.

In Section 2 below, I shall very briefly sketch a functionalist account of mental states, specifically a computational/representational theory of thought (CRTT), and argue that, straightforward scientific hypothesis that it is, Wittgenstein has no basis for rejecting it. In Section 3, however, I will focus upon what strikes me as the strongest part of Wittgenstein's discussion, his discussion of our ordinary talk about sensations, which notoriously does seem to resist explication by CRTT and invites precisely the kinds of ghostly philosophy of mind that he was trying to resist. I will argue that his remarks in this regard are not only compatible with CRTT, but actually contribute importantly to its interest; in any case, it is only against the background of a functionalist theory like CRTT that those remarks are actually plausible.

2 Wittgenstein and a Computational/Representational Theory of Thought

I take the heart of Wittgenstein's view about the mind to be contained PI:§308 in which he claims that the common error of both Cartesians and Behaviorists is to "talk of processes and states and leave their nature undecided", which seems to force us "to deny the yet uncomprehended process in the yet unexplored medium". His solution is to regard mental talk as gaining its meaning not by reference to such mysterious processes, but by its use in ordinary talk. Indeed, we should reject the "misleading parallel: psychology treats of processes in the psychical sphere, as does physics in the physical". (PI:§571) Lest this be thought merely an attack on conceptions of processes in some ghostly, non-material "psychical sphere", it is noteworthy that thoughout the later writings, and especially the recently published writings on the philosophy of psychology, Wittgenstein repeatedly inveighs against specific postulations of internal processes involving e.g., vision (Z:§614; RPP-1:§918), color perception (RC:pp37-40), reading (PI:§160), language processing (PI:§1), understanding of other people (Z:§220), aesthetic reactions (L&C:17-20), animal thought (PI:§25; Z:117, RPP-2:§192). That is, he pretty clear is intent on rejecting the very sorts of hypotheses that are the stock-in-trade of contemporary cognitive science.

But perhaps the postulations of cognitive science can be read as non-

referential in just the way that Wittgenstein would like. Perhaps such postulations are only expressions of an "attitude" towards creatures – "My attitude towards him is an attitude towards a soul. I am not of the *opinion* that he has a soul", (PI:p178), rather like Dennett's "intentional stance", an attitude or stance that doesn't involve any claims about any literally internal processes. I think this view is untenable for the following reason.

To a first approximation, cognitive science is committed to an explanatory framework that I shall call "Minimal Mentalism", which might be defined thus:

Minimal Mentalism ("MM"): there are two basic kinds of mental states, *informational* ones (such as belief, judgment) that *represent* the world as being one way or another, and *directional* ones (such as desire, wishing) that *direct* their agent towards some represented state. These states combine in the following ways: stimuli cause informational states (perception), some of which are retained (memory) and subsequently combine with one another in sometimes rational patterns to produce other informational states (thought, reasoning), and with directional states to produce further directional states and actions (decision making, action).

Why think anything has a minimal mind? As Wittgensteinians are often fond of pointing out, traditional answers have often been question-begging: transcendental arguments, introspection, appeals to intentional action, all presuppose and so cannot be used to defend mentalism (which is what leads Wittgensteinians often to regard mentalism more as an attitude than an opinion). It is important to see that MM can be supported by non-tendentiously described data.

Consider the "standardized tests" (such as the SAT and GRE) that are administered in the United States to millions of people each year. These are tests in which the "input" (printed pages) and the "output" (pages of patterns of graphite-filled rectangles) can be so physically described that they can be produced and "read" by simple machines (no "interpretation", hence "standardized"). The statistical correlations and counterfactual supporting generalizations relating the inputs and outputs of the millions of individuals are staggering, and could be made more so were the tests intended to establish commonalities instead of differences.

Now, we can suppose that there would be no problem in principle in explaining, *in any particular case*, the physical causal connections between the input and the output; but how are we to begin to explain the *correlations*? The physics and/or neurophysiology alone won't be enough: we have no reason to think it would be anything but an accident were the distinguising causes to involve the *same* specific physically characterized mechanisms.

Of course, Methodological (for example, Skinnerian) Behaviorism does purport to have an account. And, given Wittgenstein's frequent appeals to "teaching" and "training" (see e.g. PI:§5,6,86,185, 630; RPP-1:§131, RPP-2:§6,139) in accounting for human comptencies, it is hard to resist the im-

presssion that he is presuming that that's what will explain what regularities there are. But it is notorious that theories of "training" and conditioning are empirically entirely inadequate for explaining the behavior of rats and chimpanzees alone, not to mention language and the above standardized regularities.[2] If not conditioning, then precisely what sort of "training" *does* Wittgenstein have in mind? In any event, Wittgenstein needs to show how such appeals are serious alternatives to explanation in terms of inner mental states.[3]

So far as we know, MM is the only theory-sketch that begins to be remotely plausible as an explanation of standardized regularities (to say nothing of most ordinary processes of training). To a first approximation: the test takers get into certain familiar informational and directional states as a result of reading the input sheets, are reasonably intelligent (and gullible) so as to think out the answers (and/or fall for the fallacies) to the questions they take themselves to have been asked, and are able to execute the responses that are recommended as a result of that reasoning.

Now, the crucial question a Wittgensteinian needs to address is how this plausible theory-sketch could begin to be genuinely explanatory without its terms referring to literal internal processes of e.g. thinking and decision making as causes of the explained behavior. Just how do these millions of students do it? Of course, the correlations *could* just possibly be entirely an artifact of the environment of the tests, the "contexts" in which they are described, and the previous lives of the students – just as it *could* turn out, as Wittgenstein imagines on behalf of biologists, that the structure of a plant might correspond to "*nothing* in the seed" (Z:§608). But this would be a monumental coincidence – in botony, psychology and demography! To rely on it would not be mere antipathy to scientism; it would be bad science. Wittgenstein would be in the position of those tobacco lobbyists who keep insisting that the correlations between smoking and lung cancer have nothing to do with what's in cigarettes.

Now someone might reasonably complain that functionalism and even MM are so sketchy that they're hardly in a better explanatory position. Towards answering this challange, many cognitive scientists have proposed a computational-representational theory of thought (CRTT). The idea is that we should regard the brain as a computer performing operations in real time on logically complex internal representations, the primitives of which stand in certain co-variational relations with either stimulus patterns or with

[2]See not only Chomsky 1957 for the problems that such theories have with language, but also Gallistel 1990 for a rich summary of their failures even with rats and birds.

[3]Of course, he might simply forswear explanation of even these regularities altogether. Malcolm, for example, rejects both behaviorism and "the myth of cognitive processes and structures", claiming that "we could, just as rationally, have said that the man or child [recognizing a dog] just knows (without using any model, pattern, or Idea at all) that the thing he sees is a dog". (Malcolm 1977:168). I suppose we could "just as rationally" never ask for any explanations of anything. But the question is whether we *can* rationally ask for them, and, if we can, what the correct explanations might be.

phenomena in the environment. Propositional attitudes are to be explicated as computationally defined relations to such representations that, by virtue of their logical structure and those co-variant relations, express particular propositions. Mental processes are causal processes in the brain that, as in a computer, also instantiate computational ones. The details of such an account are complex, much discussed elsewhere, and so needn't detain us here.[4] It is not even important, for these purposes, that this story be *true*. All that is important is that it be a *possible* explanation of the working of some machine, and thereby a viable hypothesis about the brains of people and animals. In any case, it should be clear that there is no reference to any "uncomprehended process in [some] yet unexplored medium" (PI§308, quoted earlier). Detailed examination of neural pathways, response-time experiments, startle and eye-fixation, patterns in fallacies and other patterns in reasoning can provide plenty of evidence that bears on the question of how someone represents the world, the character of their computations, and whether they think one way rather than another.

MM and CRTT need to be distinguished from stronger claims that are often also made about the mind, but are not clearly justifiable in the same way: these are claims about consciousness, qualia, first-person authority, free will, a soul. It is these phenomena, of course, that have been the most troublesome for the philosophy of mind, partly because it has not been clear how they could fit into the causal order of the world, but also because it has not been completely clear what non-question-begging evidence could be adduced for their postulation. Notice, however, that there is no reason to burden either MM or CRTT with them. None of the states invoked by MM and CRTT to explain the standardized regularities need be conscious, or involve some special first-person relation, free will, or a soul. This is not to say in itself that these are not real phenomena (or that they are). Only that one could do an immense amount of psychology without them.

Of course, it is likely that it just these stronger phenomena that really concern Wittgenstein. In a telling passage in RPP-1, he writes:

> Ought I call the whole field of the psychological that of 'experience'? And so all psychological verbs 'verbs of experience' ('Concepts of experience.') Their characteristic is this, that their third person, but not their first person is stated on grounds of observation. (RPP-1:§836)

[4]See Fodor 1975,1987,1991, Stich 1983. Loewer and Rey 1991 provides a summary of much of this along the lines of the present paper. I should mention that, in my view, Kripke's (1980) interpretation of Wittgenstein properly belongs to this theoretical discussion (it is what gets known in it as the "disjunction" problem; see Fodor 1987:102ff), In the end his argument has to do with the resources of idealization – what counts as "science fiction", what can be included in a *ceteris paribus* clause, what is an adequate theoretical reduction to physics (see especially Fodor 1980:26-30) – issues that are very difficult to see occupying Wittgenstein (indeed it is a serious failing of Wittgenstein that he doesn't consider *theoretical* issues at all). See Pietroski and Rey (forthcoming) and Rey (forthcoming b) for discussion.

That is, Wittgenstein's remarks should be taken to be aimed not at empirically motivated theories such as MM or CRTT, but at these stronger mentalisms that have been motivated by traditional philosophical agenda, which do tend to arise precisely those cases in which there seems to be a serious discrepancy between first- and third-person-present ascription.

The cases that have received the most attention in this regard are those of sensations and consciousness. It is in regard to philosophical conceptions of those states that I think Wittgenstein's remarks are most perceptive. In the next section, I shall not try to *solve* the problems raised by these phenomena – this I have attempted in other papers[5] – but rather only set out what I think Wittgenstein's valuable contributions to this debate have been, and, particularly, how they can better be understood against the background of a theory like CRTT than against the vaguely behavioral and contextualist background emphasized by Wittgenstein.

3 Cartesianism and Qualia Reversals

I take "qualia" to be the purported properties we seem to be aware of when we have sensory experiences, different from the properties of the objects in the external spatio-temporal world that may cause those experiences: e.g. pain, the taste of pinapple, looking red. A good number of philosophers have argued that there are such properties and that they (and consequently genuine sensory experiences) cannot be captured by any functionalist proposals, but especially not by CRTT. With Dennett (1991), let us call such philosophers "qualiaphiles". Most notably, Ned Block (1978/80) argues that no functionalist theory can capture qualitative states, since one can imagine *qualia reversals* with respect to functionally equivalent systems.

One natural response is straightforwardly dualist: qualia just *are* further properties that do not supervene on any physical/functional ones. Others seek refuge in as yet unspecified physical properties of the brain. Block and even another leading functionalist, Sydney Shoemaker, look to purely physiological states at least to ground the particular qualities of sensory experience. Now, I think Wittgenstein has the makings of an interesting argument against both these positions. It will take, however, some sorting out to see it, as well as a replacement of his implausible behavioral/contextual framework by a version of CRTT.

As mentioned earlier, a good deal of Wittgenstein's argument against literal internal mental states turns on his criticisms of an excessively referential theory of meaning. Now, someone might concede those criticisms to him and still insist that there are "private" qualia nonetheless: mental terms may not get their *meaning* from referring to private objects, but they refer to them all the same. Against this position, Wittgenstein raises a number of further arguments, the most famous being the argument against the possibility of a "private language". I mercifully do not propose to add to the

[5] Rey 1992, 1993.

mountains of material on that topic.[6] Rather, I'm interested in a somewhat different argument that is intertwined in Wittgenstein's discussion of that issue, but is, I think, independent of it. It might be called the "otiosity argument": appeals to essentially private phenomena serve absolutely no interesting explanatory or introspective purpose; they are like "a wheel that can be turned though nothing else moves with it" (PI:§271). One doesn't need to claim that there *couldn't* be such wheels, i.e. such peculiar phenomena; all that needs to be established for scientific purposes is that there's absolutely no good reason to think there are. What I think is particularly interesting about this argument is that it seems to me to apply not only to "private" objects, but to *any* non-CRTT properties that might be recruited (as they are by Block) to underwrite appeals to "qualia".

I think that the best way to appreciate its force is to consider precisely the hard cases of inverted qualia emphasized by Block: purported examples of people who are functionally identical, one of whom, however, sees, for example red wherever the other sees green.[7] These cases present an especially important problem for Wittgenstein, since they would appear to provide a purchase on the notion of a private quality without clearly invoking an excessively referential theory of meaning.[8] Indeed, they can seem to invite a fairly strong form of dualism, since it can be difficult to see how one could settle the doubts they raise, short of somehow "entering into" – telepathizing? – the "private" world of someone else's mind to check things out. Even if one is a materialist, it can seem less than perfectly clear how and why comparing one person's physiological properties with another's should settle the issue. *Which* physical properties should one look for? Merely ones that happen to be correlated with a particular kind of sensory experience? Suppose there are none; should we conclude that peoples' experiences therefore differ? Suppose there are some: what seriously argues that they are therefore the same?[9]

As Wittgenstein seems to have been the first to notice,[10] the possibility

[6] I am persuaded by, *inter alia*, Judith Thomson 1964 that Wittgenstein presumes an extreme verificationism that Malcolm 1954 made more explicit and which we have absolutely no reason to endorse.

[7] The literature on this topic is considerable. An excellent discussion of it is in Shoemaker 1981/84 and Block 1991.

[8] Wittgenstein notices this supposed possibility at PI:§272, but his remarks are too brief. It's unclear whether he thinks it's a serious possibility only on the referential assumptions he's attacking. However, abandoning that assumption won't by itself dissolve the puzzle: it can be raised, particularly within Wittgenstein's behavioral framework, by considering e.g. the possibility of secretly implanting spectrum inverting lenses in an infant, who then grows up to speak and act as we do. The problem is what we should say about ordinary cases in view of such a possibility, quite apart from any semantic theory.

[9] Joseph Levine 1983 has discussed the "explanatory gap" that seems to result from the fact that physiological properties don't seem to be capable of necessitating any qualitative ones in that way that, for example, physical properties of water molecules necessitate the macro-properties of water.

[10] Shoemaker (1981/84) provides a history of the topic, and remarks on some of the ironies of Wittgenstein's own positions.

of *qualia* reversals between people invites that possibility in the case of a *single* person over time. It's important to see how this might arise. At one place, Wittgenstein writes:

> someone says, "I can't understand it, I see everything red today that was blue yesterday and *vice versa*". We answer "It must look queer!" He says it does and, e.g. goes on to say how cold the glowing coal looks and how warm the clear (blue) sky... (PESD:284)

He goes on to claim that we should be inclined to say that the person had undergone qualia inversion; but, so far as I can see, he presents no argument for this over other hypotheses that on the face of it seem perfectly possible – e.g. that she misremembers her experience, that she understands old words in a new way, that she's in an entirely novel qualitative state. She is roughly in the position that many of us are in when we notice, say, in regard to beer, that "our tastes have changed" since we were children and we wonder whether this is due to beer actually tasting differently or it tasting the same and our evaluating the taste differently. The wonder certainly seems perfectly intelligible, yet it's quite unclear as to what further considerations should settle the matter.

Of course, to wonder about hypotheses that one can't imagine settling in principle is anathema to Wittgenstein. And he thinks he can avoid them by denying that such terms as 'the taste of beer' refer to a (literal) inner process. Indeed, he seems to think that these sorts of hypotheses are so obviously idle that he advises us to:

> Always get rid of the idea of the private object in this way: assume that it constantly changes, but that you do not notice the change because your memory constantly deceives you. (PI:p207; see also §271)

a point that even more pointedly emerges in the famous passage in which he supposes that everyone had a box whose contents, which they call 'a beetle', were observable only by the owner. "One can", he claims "'divide through' by the thing in the box; it cancels out whatever it is" (PI:§293).

Now, I think, so stated, this view can reasonably be thought to beg the question against the qualiaphile, who plausibly insists that you *cannot* just "divide through": *precisely* what he is worried about is that although 'red', 'green' and 'the taste of beer' – or for that matter 'beetle' – may well have a use in language, a sceptical problem can still arise as to whether they actually refer to the same thing. Merely pointing out that it might make no difference to peoples' ordinary *practice* is entirely beside the point. The qualiaphile concedes *that*. What he wants is an *argument* for thinking that this practice is justified, particularly given the possibility of reversals, and as well as the possibility raised, but blithely dismissed by Wittgenstein, of our constantly forgetting ever-changing states.

I think that something along the lines of the argument Wittgenstein sketches here is correct, but only so long as one leaves the excessively behavioral framework in which he raises the issue. It's crucial to distinguish

qualia reversal problems as they arise for behavioral as opposed to functionalist accounts. In a standard *behavioral* case (say, of reversing lenses secretly implanted at birth) all that has happened is that the behavior caused, for example, by red is instead caused by green. The actual and (*ex hypothesi*) irrelevant events in the brain – the relations of retinal stimulation to other events in the nervous system – are to remain unchanged. But for a *functionalist*, mental states are identified by the patterns in those internal events. Behaviorism, after all, tends to be motivated by largely epistemological issues – for example, by extreme verificationism. Functionalism is more *metaphysical*: for example, what *sort* of state could play the explanatory role that mental states are supposed to play? Indeed, it is precisely because it seems a very reasonable metaphysical presumption that qualitative states supervene on the states of the agent's brain that we are entitled to take seriously the possibility of reverse qualia in a behavioral case: given what we know about ourselves, it just seems improbable in the extreme that qualitative states supervene on *behavior* alone. This is why it seems particularly inappropriate to "divide through" in the cases that Wittgenstein describes.

When the reversal is, however, "taken inside" this very same presumption ought to have the *opposite* effect: if qualitative states supervene on the brain, then a change in the brain should entail a change in them.[11] So already the functionalist is in a better position to answer the qualiaphile. But, as I mentioned earlier, functionalism itself tends to be excessively vague and abstract about what precisely the relevant internal roles might be. CRTT is sharper. Indeed, with regard to the question about the taste of beer, it can supply a potentially perfectly good answer: it can quite reasonably posit sub-systems of the mind that are dedicated to specific processing: for example, one for (gustatory) perception, another for memory, another for the ordering of preferences, still others for the abilities to attend and compare. The question about the taste of beer would then be a question about just which of these (or other computationally defined) sub-systems is responsible for the net change in attitude towards beer. *Pace* Dennett, I see no reason to think that evidence couldn't be provided one way or the other along lines mentioned earlier with regard to CRTT generally. But I also see no reason to think that that evidence need be either introspective or strictly "behavioral": subtle differences in response time, or perhaps just internal examination of the computational organization of the brain could turn out to be the best one could do.[12]

A serious qualiaphile might claim that still all that wouldn't be enough. It may solve the beer example, but why think that it would solve the harder questions about, for example, red and green? It is worth pressing CRTT

[11] Which, one should note, is a reason to begin not to trust one's intuitions about functionalist cases as about behavioral ones. Lycan 1987: 59-61 makes a similar point.

[12] See Fodor 1983:76-7 for particualrly subtle empirical evidence for hypotheses of this very sort with regard to perception of phonemes.

further. Elsewhere[13] I have tried to show how CRTT could be understood in a way that accounts for what I take to be a variety of non-tendentious facts that are associated with sensory experience (e.g. its involuntariness, ineffability, even an aspect of its "privacy"). Roughly, I claim that such experiences are states involving a particular computational relation to specially restricted predicates in a creature's system of internal representation. It is of a piece with the claim that sensory experience can be understood as a species of *cognition*, broadly understood (i.e. understood as those states involving computational relations to representations *à la* CRTT). The details need not concern us here. Call this resulting supplementation of CRTT, "CRTQ". So the qualiaphilic worry here is that there could be qualia reversals even for CRTQ-equivalent systems.

Let us suppose that CRTQ is otherwise an adequate theory of all of an agent's propositional attitudes *except* tendentious ones (like "having the *genuine* thought that one is now in pain") that might depend upon one's view of the qualia issue. And let us suppose that there is an elaborate systematization of the mind such as that sketched above with regard to the taste of beer. The qualiaphile is then suggesting that there *still* might be differences in qualia even though all of *that* is fixed. But now this possibility *does* seem obscure. The suggestion entails that it would be possible for there to be a change in qualia while the agent stares fixedly in front of her at some colored patch, entirely convinced that she is in a constant qualitative state with respect to allthe non-tendentious properties, but at the same time there is absolutely nothing otherwise different or defective about her systems of perception, memory, attention, reasoning, comparison, or any other cognitive sub-system of the mind. Unlike the intra-personal cases imagined by Wittgenstein (RC:28, quoted above), the agent wouldn't be in the slightest confused or doubtful about her sensory experience. She'd be cognitively indistinguishable (actually and counterfactually) from a perfectly normal person. For all you and I know we could be undergoing just such qualitative reversals every other minute, but never have an inkling of it until the dualist in some way finally reveals the real workings of qualia, or the physiologist finally provides a convincing story about why just *those* non-cognitive physiological changes are changes in real feel. Until we are provided with some serious reason to believe otherwise, such postulations seem entirely gratuitous. Of what possible interest would they be? It's not just that, as Wittgenstein assures us, language and ordinary life could proceed quite happily without such entities: *everything* could! Such cognitively transcendent qualia would serve absolutely no practical, introspective or serious theoretic purpose whatsoever. Insofar as "what it's like" has anything to do with how things seem, the comparisons we make, what we think and love and hate, they would be entirely otiose, no better than absolute space

[13]Rey 1991, 1992. Some of what I say here I said there. A similar view is also advanced by Lycan 1990 and Leeds (forthcoming).

or undectable angels rescuing immaterial souls.[14]

Indeed, the qualiaphile, unlike the computationalist, is in the awkward position of having absolutely *no* non-question-begging evidence for his claims. His position seems no better than that of the theist who claims God exists because, after all, he has a "direct experience" of Her ("Explain that!" he and the qualiaphile exclaim). It is *here*, at an explanatory point provided by a complete theory of cognition, that one *can* "divide through" and say that any such further property "cancels out whatever it is", private or public. Wittgenstein's (in)famous remark that "an inner process stands in need of outward criteria" (PI:§580), so implausible read *behaviorally*, is entirely plausible if the "outer" includes computational facts about our brains: at any rate, an inner process stands in need of *some* sort of cognitive criteria.

Wittgenstein's real insights emerge, however, in his diagnosis of our strong inclination to think otherwise. The qualiaphile is "held captive" by a certain "picture" (PI:§115) of the "inner world" of the mind, but without having any idea about how it is to be seriously *applied*. Perhaps he thinks that the phenomena in such a world somehow carry with them their own, self-evident rules of application: "I know how the color green looks to *me*", Wittgenstein has the interlocuter exclaim; to which he replies, "Imagine someone saying: 'But I know how tall I am!' and laying his hand on top of his head to prove it" (PI:§§278-9). As Wittgenstein is at pains to argue throughout the *Investigations*, without the possibility of some objective comparisons, mere ostensions are empty gestures. And nothing – not an image, not a sentence, not a thought, certainly not the picture of the "inner world" – brings with it its own criteria of application.

Lacking such a principle of self-application, the naive picture leaves us precisely in the kind of position Wittgenstein describes in PI:§308, the passage that I earlier took to be his central (if exaggerated) claim: "we talk of processes and states and leave their nature undecided". "But that", of course, "is just what commits [the qualiaphile] to a particular way of looking at the matter. For we have a definite conception of what it means to know a process better" – for example, what it is to learn about metabolism by investigating the workings of someone's body. But images of somehow entering another person's mind "non-physically" are (so far as we know) unintelligible; and possibilities of connecting brains for the purpose of any kind of cognitive comparison are limited in the extreme: even were we to be clear about what counted as a "connection", what reason in the world would we have to take seriously the judgments that might result? Such a procedure would be as silly as determining whether two computers are computing the same program by seeing what happens when you plug one into the other. Connect my Toshiba to your Macintosh and I expect what would happen is what happens whenever the hardware isn't perfectly matched: nothing.

[14]Dennett 1991:392-3 tries to make a similar point, but unfortunately traps himself in precisely the same kind of behavioral framework as Wittgenstein. I discuss his difficulties in Rey (forthcoming a).

Certainly whatever happened would not be evidence one way or another for similarity of program unless we already had a theory that indicated to us what counted as *computing the same program* (which, of course, we do). Similarly, imagined interpersonal comparisons of qualia presuppose and so cannot ground our accounts of qualitative identity.

In desperation, of course, the qualiaphile just insists upon inspecting the actual physical properties that realize CRTQ. But as I already mentioned that others have reasonably complained, this is entirely inconclusive: even if we did in some way fix upon some sensory correlated properties, we could still quite reasonably wonder just what they really had to do with anything: How much change in them can be tolerated? Just how are they incorporated into a person's *mental* life in a way that, e.g. the person's distance from the Eiffel Tower is not? Whether physiological or dualistic, such properties are

> like pontificals which we may put on, but cannot do much with, since we lack the effective power that would give these vestments meaning and purpose. (PI:§427)

And so "now the analogy which was to make us understand our thought falls to pieces", and we wonder how *possibly* we could ever find out about one another's qualia, or even our own.

Of course, if the application *is* provided, then there can be no such objection nor any such wonder. That is precisely the purpose of CRTQ. It provides a relatively clear basis for applying the "analogy" quite literally: the having of a sensation is the undergoing of a certain computationally defined process that is realized in the brain precisely as such processes are now standardly realized inside computers. Thus may we save talk of *literal* inner processes. The only suggestion that remains lacking clear application is the qualiaphile's further conception of properties transcending all those computational facts. It is in this way that Wittgenstein's insights benefit from a CRTQ – just as a CRTQ in turn benefits from them. CRTQ may not, itself, be exactly *required* for those insights; but one needs *some* story with that the kind of psychologically plausible detail it provides.

Abbreviations for Wittgenstein's Works

L&C *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*, ed. Cyril Barrett, Oxford: Blackwell 1978.

PESD "Notes for Lectures on 'Private Experience' and 'Sense Data'", ed. Rush Rhees, *The Philsophical Review* 77 (1968), pp. 275–339.

PI *Philosophical Investigations*, Oxford: Blackwell 1953.

PR *Philosophical Remarks*, ed. Rush Rhees, Oxford: Blackwell 1975.

RC *Remarks on Colour*, ed. G.E.M. Anscombe, Oxford: Blackwell 1977.

RPP-1 *Remarks on the Philosophy of Psychology* I, ed. G.E.M. Anscombe and G.H. von Wright, Oxford: Blackwell 1980.

RPP-2 *Remarks on the Philosophy of Psychology* II, ed. G.E.M. Anscombe and G.H. von Wright, Oxford: Blackwell 1980.

VB *Vermischte Bemerkungen*, translated as *Culture and Value*, ed. G.H. von Wright and Heikki Nyman, Oxford: Blackwell 1980.

Z *Zettel*, ed. G.E.M. Anscombe and G.H. von Wright, Oxford: Blackwell 1981.

References

Dennett, D.C. 1991 *Consciousness Explained*, Boston: Little Brown.

Baker, G.P and Hacker, P.M.S. 1983 *An Analytical Commentary on Wittgenstein's Philosophical Investigations* I , Oxford: Blackwell.

Block, N.J. 1980 "Troubles with Functionalism", in Block, N.J. (ed.), *Readings in the Philosophy of Psychology*, Cambridge MA: Harvard University Press, pp. 268–306.

Dretske, F. 1987 *Explaining Behavior: Reasons in a World of Causes*, Cambridge MA: MIT Press.

Fodor, J. 1975 *The Language of Thought*, New York: Thomas Crowell.

Fodor, J. 1983 *The Modularity of Mind*, Cambridge MA: MIT Press.

Fodor, J. 1987 *Psychosemantics*, Cambridge MA: MIT Press.

Fodor, J. 1991 *A Theory of Content*, Cambridge MA: MIT Press.

Jackson, F. 1982 "Epiphenomenal Qualia", *Philosophical Quarterly* 32, 127–32.

Kripke, S. 1982 *Wittgenstein on Rules ad Private Language*, Cambridge MA: Harvard University Press.

Leeds, S. 1992 "Qualia, Awareness and Sellars", *Nous*, pp. 303–29.

Levine, J. 1983 "Materialism and Qualia: the Explanatory Gap", *Pacific Philosophical Quarterly* 64, 354—61.

Loewer, B. and Rey, G. 1992 Introduction to Loewer, B. and Rey, G. (eds.), *Meaning in Mind: Fodor and his Critics*, Oxford: Blackwell.

Lycan, W. 1987 *Consciousness*, Cambridge MA: MIT Press.

Lycan, W. 1990 "What is the 'Subjectivity of the Mental'?", in Tomberlin, J. (ed.) *Philosophical Perspectives*, vol. 4: *Action Theory and the Philosophy of Mind*, Atascadero CA: Ridgview.

Malcolm, N. 1954 "Wittgenstein's *Philosophical Investigations*", *Philosophical Review* 63, 530–559.

Malcolm, N. 1977 *Thought and Knowledge* Ithaca NY: Cornell University Press.

Monk, R. 1990 *Ludwig Wittgenstein: the Duty of Genius*, London: Cape.

Pietroski, P. and Rey, G. forthcoming "When Other Things Aren't Equal: Saving *Ceteris Paribus* from Vacuity", *British Journal for the Philosophy of Science.*

Rey, G. 1992 "Sensational Sentences Switched", *Philosophical Studies* 67, 73–103.

Rey, G. 1993 "Sensational Sentences", in Davies, M. and Humphyries, G. (eds.), *Consciousness*, Oxford: Blackwell, pp. 240–57.

Rey, G. forthcoming a "Dennett's Unrealistic Psychology", in *Philosophical Topics.*

Rey, G. forthcoming b "Keeping Meaning More in Mind", in Vision, J. (ed.) *Cognition and Information* (to appear).

Rey, G. forthcoming c "Why Witgenstein Ought to Have Been a Computationalist and What a Computationalist Can Gain from Wittgenstein" in Gottlieb, D. and Odell, J. (eds.) *Wittgenstein and Cogntive Science*, Ithaca NY: Cornell University Press.

Shoemaker, S. 1981/84 "The Inverted Spectrum", in S. Shoemaker *Identity, Cause and Mind*, Cambridge: Cambridge University Press.

Stich, S. 1983 *From Folk Psychology to Cognitive Science*, Cambridge MA: MIT Press.

Thomson, J. 1964 "Private Language", *American Philosophical Quarterly* 1.

Wittgenstein's Private Language Argument and Reductivism in the Cognitive Sciences

Dale Jacquette

1 The Privacy of Experience

The private language argument in Wittgenstein's *Philosophical Investigations* is standardly interpreted as refuting the privacy of experience. Wittgenstein in this light is often seen as having cleared away a heavy underbrush of conceptual confusions in the philosophy of mind, discrediting the distinction or explanatory force of the distinction between public and private phenomena, and allowing a reductivist psychological or cognitive science to flourish in place of unsupportable phenomenological assumptions about the privacy of thought.[1]

D.M. Armstrong, in explaining the historical precedents for his own reductive materialism-cum-logical-behaviorism in *A Materialist Theory of the Mind*, offers this now familiar reading of Wittgenstein's claims about the need for criteria of correctness in the private language argument, as upholding analytical behaviorism:

> Gilbert Ryle's book *The Concept of Mind* seems to be a defence of Analytical Behaviourism. I think the same is true of Wittgenstein's *Philosophical Investigations*, although this interpretation is hotly denied by many disciples. The problem of interpreting Wittgenstein's book may perhaps be reduced to the problem of interpreting a single sentence:
>
> 580 An 'inner process' stands in need of outward criteria.
>
> When Wittgenstein speaks of 'outward criteria' he means bodily behaviour... But there is one difficulty in interpreting Wittgenstein and Ryle as Behaviourists. Both writers deny that they hold this doctrine! I think, however, that the only reason that these philosophers denied that they were Behaviourists was that they took Behaviourism to be the doctrine that there are no such things as minds. Since they did not want to deny the existence of minds, but simply wanted to give an account of the mind in terms of behaviour, they denied that they were Behaviourists.[2]

[1] See Paskow 1974; Reeder 1979.
[2] Armstrong 1968: 54-55

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 89-99

If Armstrong's interpretation is correct, it puts Wittgenstein's private language argument in the forefront of disagreement between the proponents of reductivist cognitive science and phenomenology. Yet Wittgenstein indicates in §128 that he would no more affirm nor deny any positive psychological theory, on the grounds that he regards substantive theorizing as falling outside the scope of philosophy.[3] The fact that Wittgenstein in §307 seems to distance himself from behaviorism should give us pause in attributing to him any reductivist theory of mind. If we are to engage in the risky business of comparing single sentences from Wittgenstein's complex and highly mannered text with that chosen by Armstrong, we might alternatively refer to §248, in which Wittgenstein indicates that the privacy of sensation is an analytic truism or redundancy:

> 248 The proposition "Sensations are private" is comparable to: "One plays patience [the card game of solitaire] by oneself".

This suggests that Wittgenstein at least does not unequivocally reject the privacy of sensation. But then the standard account of the private language argument by which the hard scientific reductivism of cognitive science is supposed to triumph over the phenomenology of private first-person introspective experience must be reconsidered. The passage unfortunately can also be understood in just the opposite way, as disputing the significance or intelligibility of the thesis that sensations are private, depending on how we are to understand the later Wittgenstein's attitude toward tautology and necessary truth. He might be saying, as the *Tractatus Logico-Philosophicus* would have it, that the privacy of thought is literally senseless, though perhaps not nonsensical (as the distinction in 4.461–4.4611 allows). But in the *Investigations*, with the rejection of logical atomism and the picture theory of meaning, Wittgenstein on the contrary might regard a statement like the above in §248 as expressing a rule of philosophical grammar governing use of the words 'private' and 'sensation'.

The questions posed by such conflicting interpretations can only be approached by clarifying the purpose and content of Wittgenstein's private language argument. This in turn requires a careful exposition of how the argument is situated in the overall structure of Wittgenstein's project in the book.

[3] PI §128. The passage in §307 to which Armstrong refers does not unequivocally reject, but rather indicates Wittgenstein's unwillingness to endorse behaviorism, or to classify himself as a behaviorist. He diverts the question about his own substantive views concerning the nature of mind to his preferred focus on the philosophical grammar of psychological language games. His reply is ponderous, conditional, and thetically noncommital. "'Are you not really a behaviourist in disguise? Aren't you at bottom really saying that everything except human behaviour is a fiction?' – If I do speak of a fiction, then it is of a grammatical fiction."

2 Wittgenstein's Private Language Argument

To begin, let us review what Wittgenstein says in presenting the private language argument. As a thought experiment, we are to try to imagine a diarist using a private sensation language to record recurring pains. The diarist records sensations in a symbolism that cannot be translated or decoded into a public language, but that only he as first-person experiencer can understand. Wittgenstein argues that, despite appearances, the conceptual requirements for naming sensations presupposed by a private sensation language are unsatisfiable. From this he concludes, contrary to the thought experiment's assumption, that there cannot be a private sensations diary written in a private sensation language.

The reason is that the philosophical grammar for naming particulars requires what Wittgenstein calls criteria of correctness, a requirement which cannot be met in the case of private sensations. Criteria of correctness are reliable procedures by which name-users can determine in principle when the same numerically identical object is reidentified and referred to by the same name on different occasions. If this condition is not satisfied, then, according to Wittgenstein, putative name-users are not really naming, but only going through the motions of naming. The criteria must be strong enough to distinguish instances in which the rule always to use the same name for the same object is actually followed, from those in which the name-user merely believes that the rule is being followed, because he merely believes that he is in a position to apply the same name to the same object. Wittgenstein concludes that criteria of correctness obtain only in naming public particulars, such as my body or the Eiffel Tower. For these, there is always (supposedly) a way of deciding whether the same name is used on different occasions to refer to the same object. Wittgenstein believes that in such cases we can distinguish actually following the rule from merely believing that we are following it by checking up in various ways on the object's spatio-temporal continuity. In trying to name particular private sensations, by contrast, there simply are no criteria of correctness. Wittgenstein maintains that there is no test independent of my belief that I am following the rule always to use the same name for the same pain, by which I can determine whether I am actually following the rule, or whether I merely believe that I am doing so. The belief would have to justify itself, with no possibility of being overruled by external correction or check.

If I am a private sensation diarist, I have no reliable way to confirm that I have correctly reidentified the very same recurring sensation, or merely a similar but strictly different one. In the latter situation, I will not have followed the 'Same pain, same name' sensation particulars naming rule. I will merely believe that I am following it, because I merely believe that the same sensation has recurred. As a putative private name-user, I can only rely on belief in the form of impression and memory. But if this is my only resource, then I am limited to the closed circle of my beliefs in whatever

identity test I may try to apply. If mere belief cannot be confirmed or disconfirmed by something beyond or outside itself, then in Wittgenstein's sense there are no criteria of correctness for naming private sensations. If criteria of correctness are required for naming particular private sensations, but are strictly unavailable, then private sensations cannot be named, and in that sense there cannot be a private sensation language.[4]

Wittgenstein fashions some of the most colorful images in all of his philosophical writing to describe the logical predicament of the private sensation diarist. These have entered into the philosophical vocabulary and exerted an unparalleled influence on contemporary thinking about the relation between language and mind. Wittgenstein says that the attempt to verify by impression and memory that the same sensation is being referred to by the same name on different occasions is, "As if someone were to buy several copies of the morning paper to assure himself that what it said was true". (§265) Elsewhere, he refers to the private naming of sensations as like a language game involving a beetle in a box, in which box holders can examine only their own beetles, and not another's.

> 293 ... Here it would be quite possible for everyone to have something quite different in his box. One might even imagine such a thing constantly changing. – But suppose the word "beetle" had a use in these people's languages? – If so it would not be used as the name of a thing. The thing in the box has no place in the language-game at all; not even as a something: For the box might even be empty. – No, one can 'divide through' by the thing in the box; it cancels out, whatever it is. That is to say: if we construe the grammar of the expression of sensation on the model of 'object and designation' the object drops out of consideration as irrelevant.

But if the private language argument is sound, does it show any more than that private sensations cannot be named as particulars? Does it show that there are no private mental objects, or that sensation is not private? Does it have substantive implications for the dispute between reductive cognitive science and "phenomenology"?

3 Context and Purpose of Wittgenstein's Private Language Argument

The role of the private language argument in Wittgenstein's later philosophy can only be understood in terms of the work's larger context. In the *Investigations*, Wittgenstein is primarily concerned with the requirements of meaningfulness in language, especially in the aftermath of his dissatisfaction with his own previous attempt to explain the semantics of logic and language in the *Tractatus*. The picture theory of meaning postulates juxtapositions of simple objects in one-one correspondence with concatenations

[4]PI §§243-271. See Jacquette 1994, especially Chapter 5.

of their names. But while in the early work Wittgenstein takes the naming of particulars for granted, in the later philosophy he raises the most subtle and penetrating difficulties about the conditions under which an object is named.

Wittgenstein rejects the idea that naming can be achieved by purely behavioral ostension in his extended criticism of the passage from Augustine's *Confessions* in the opening sections of the *Investigations*. Pointing by itself inevitably underdetermines what the name-user intends. To point toward a child and say 'Rebecca' could be taken to mean the child, the space she occupies, her hair color, or many other things. The direction, and even the fact of pointing, cannot be inferred merely from a finger's being held out. Wittgenstein observes that background cultural conventions must be presupposed as a kind of stage setting, without which simple ostension merely in extending a finger in space cannot be understood unambiguously as indicating a particular intended object. The necessary background conventions in turn cannot all be communicated by simple ostension, on pain of infinite regress. The temptation then, in view of the failure of purely behavioral ostension, is to turn from body movement or behavior to mind, or, as Wittgenstein says, in more traditional Cartesian terms, from *body* to *spirit*.

> 36 And we do here what we do in a host of similar cases: because we cannot specify any one bodily action which we call pointing to the shape (as opposed, for example, to the colour), we say that a spiritual [mental, intellectual] activity corresponds to these words.
>
> Where our language suggests a body and there is none: there, we should like to say, is a spirit.[5]

The point of the private language argument, after the naive Augustinian behavioral or ostensive theory of naming has been rejected, is to disallow what Wittgenstein calls the alternative 'spiritual' explanation of meaning, in which objects are named by an internal private ostension, as a kind of mental pointing. The private language argument discredits this solution too, leaving a pragmatic or instrumentalist account of meaning as the only remaining choice. This account is rooted in the actions of language-users in shared practices or common forms of life, as complex as human societies themselves, in which forms of life provide the extra-semantic foundations of semantics. The instrumentalist theory is signaled again by some of Wittgenstein's most interesting metaphors. Words in a language are compared to a toolbox of tools with different purposes, and to the handles in a locomotive cabin, which do different things, even if they look more or less alike.[6]

[5]The English equivalents 'mental, intellectual' in brackets are supplied by the translator, and do not appear in the original German text. They are legitimate variations in a language that has no straightforward equivalent of the word 'mind', but may also indicate contemporary uneasiness with or embarrassment about Wittgenstein's use of the more Cartesian-sounding terms '*geistige*' and '*Geist*'.

[6]PI §§11–12, 14–15.

The main thrust of the private language argument is to reject the possibility that meaning can be explained in terms of psychological ostension, or private mental pointing to objects. This is ruled out by the consideration that sensation particulars cannot be privately named if the philosophical grammar of naming particulars on the model of individual object and designation requires criteria of correctness, and if criteria of correctness are necessarily unavailable in the case of private sensations. But how exactly is the private language supposed to show this? What is the connection between the private language argument and the failure of mental pointing or private psychological ostension as a spiritual account of naming?

Wittgenstein's text leaves us dangling precisely here, and we can only proceed by plausible interpolation to complete the argument Wittgenstein has sketched. (Wittgenstein in the Preface explains that he does not want to spare his readers the trouble of thinking for themselves.[7]) There are at least two possibilities for filling in the blanks.

The first solution is to argue that psychological ostension does not explain naming generally if it cannot explain the naming of particular private sensations. Yet if private ostension explained naming, it should otherwise be expected to give the best or easiest explanation in the case of private sensations, which are immediate in experience or right before the mind, ready for the naming. On this interpretation, the private language argument refutes mental pointing or private ostension by showing that it provides an inadequate account of naming, because pains or private sensations generally are not sufficiently identifiable to apply the 'Same pain, same name' rule. The suggestion has some merit, but it may be unsatisfactory in presupposing the possibility of naming particular private sensations, concluding only that private ostension cannot do the job. The private language argument on the contrary seems to contradict the naming of private sensations by any means.

The second hypothesis is more persuasive. By this account, pain is just a useful noteworthy example of a private experience from which Wittgenstein draws a more general moral about the impossibility of naming by mental pointing or private ostension. The private language argument precludes naming by private psychological ostension because it (supposedly) shows that private experiences like mentally pointing to an object cannot constitute namings. It does this (supposedly) by showing that we cannot follow the generalized 'Same object, same name' rule. And it does this (supposedly, on this interpretation) by showing that we cannot satisfy the requirement of knowing as opposed to merely believing on distinct occasions that we are using the same name, or naming the same object with the same name, if naming is itself an unnameable private psychological occurrence of mental pointing or private ostension.

The second analysis completes Wittgenstein's argument against mental pointing or private ostension by maintaining that private sensations are not sufficiently identifiable to constitute names or namings in applying the 'Same pain, same name' rule. The problem is not, as with the first interpretation, that pains will not hold still, so that we cannot be sure when we have the same pain to which we may then apply the same name. Rather, the second interpretation emphasizes the fact that mental pointings or private psychological ostensions construed as naming experiences providing names for pains and other objects will not hold still, so that we cannot be sure when we have the same name to apply to pains and other private sensations, or to any other object. Thus, the hypotheses offered to complete this part of Wittgenstein's criticism focus in different ways on the two sides (sameness of pain and sameness of name) of the 'Same pain, same name' equation for their explanations of his objection to mental pointing or private ostension as an inadequate account of naming based on the private language argument.

Far from concluding that sensation is not private, each of these interpretations of Wittgenstein's private language argument presupposes rather than refutes the privacy of thought. If sensation is not private, then there is no failure of naming private sensations by mental pointing or private ostension, nor can strictly unreidentifiable private mental pointings or ostensions fail as satisfactory namings. The consequence on either account is that Wittgenstein's private language argument offers no support whatsoever to reductivist cognitive science as against the phenomenology of private first-person introspection.

[7] PI, Preface: vi: "I should not like my writing to spare other people the trouble of thinking. But, if possible, to stimulate someone to thoughts of his own."

4 Pragmatic Justification for Naming Particular Private Sensations: Wittgenstein's Manometer Problem in §§270–271

Is it possible, if the correct theory of meaning is based on action rather than body or spirit, to provide a pragmatic justification for naming particular private sensations? Wittgenstein eliminates this instrumentalist defense of private sensation language in §§270–271, in his discussion of the manometer problem.

> 270 Let us now imagine a use for the entry of the sign 'S' in my diary. I discover that whenever I have a particular sensation a manometer [blood pressure gauge] shews that my blood-pressure rises. And what is our reason for calling "S" the name of a sensation here? Perhaps the kind of way this sign is employed in this language-game. – And why a "particular sensation," that is, the same one every time? Well, aren't we supposing that we write "S" every time?

Wittgenstein argues that the pragmatic justification for sign 'S' in the example is apparent only. Relying on sensation sign use in the diary to determine blood pressure in the absence of the manometer would be just as effective even if it were not exactly the same particular sensation that was

identified on separate occasions, but other, so to speak, S-like sensations, provided they coincide with blood pressure fluctuations measurable (perhaps counterfactually) by the manometer. That is, Wittgenstein regards sign 'S', which is supposed (for purposes of indirect proof) to name a particular private sensation, as serving instead as a generic type or kind term for a class of similar sensations, related by their dispositional public correlations with manometer readings. But if the pragmatic payoff for naming sensations obtains equally on the assumption that similar kinds of sensations rather than identical particulars are designated, then there is no pragmatic justification for naming particular as opposed to similar types or kinds of sensations. It is pointless from a pragmatic or instrumental point of view in that case to try assigning names to pains and other sensations as particulars, for the attempted use of names as opposed to predicates does no real work. It has, after all, no use, and as such it constitutes what Wittgenstein likes to call a useless gesture or idle ceremony.[8]

5 Private Language and the Privacy of Sensation

Those who interpret Wittgenstein's private language argument as disproving the privacy of thought seem to do so implicitly on the basis of a simple inference. The argument is so uncomplicated that it has not been formulated with sufficient care to expose the fallacy and commitment to false assumptions it contains. The proof has this form:

1 If thought is private, then there is a private language of thought

2 But there is no private language of thought

3 Therefore, Thought is not private.

But while the inference is deductively valid, its premises are not easily defended. There is nothing in Wittgenstein's discussion to establish the strong claim in (1) that thought is private only if there is a private language of thought. The private language argument shows at most that private sensations cannot be named as particulars. By itself this does not entail that unnameable private sensations do not exist beyond the reach of designation. The second assumption in (2) moreover seems too general. Wittgenstein's argument does not imply that sensation is public, nor that private sensations cannot be referred to generically in a private sensation language. These could be collectively rather than individually designated, as the manometer problem itself suggests, by predicates or generic terms for types or kinds of sensations in a language that does not try to name particular private sensations. Wittgenstein's sensation diarist, insofar as he successfully achieves

[8]PI §270: "And now it seems quite indifferent whether I have recognized the sensation right or not. Let us suppose I regularly identify it wrong, it does not matter in the least. And that alone shews that the hypothesis that I make a mistake is mere show. (We as it were turned a knob which looked as if it could be used to turn on some part of the machine; but it was a mere ornament, not connected with the mechanism at all.)"

anything at all, can and presumably does use the generic name 'S' for similar sensations. The diarist merely believes he is naming a particular private sensation, when in fact he is applying a generic predicate term like those found in ordinary language for such categories as 'pain', 'shooting pain', 'stabbing pain', and the like. These are not names, but descriptions for unnameable particular private sensations belonging to nameable kinds.[9]

When Wittgenstein says, in passage §580, which Armstrong quotes, that inner processes stand in need of outward criteria, he undoubtedly means that they need such criteria in order to be named as particulars, not in order to be explained. This is clear when it is recalled that scientific descriptions and explanations of mental occurrences, of which Wittgenstein offers no critique, are typically concerned with general principles, in which sensations and other psychological events are classed together under predicates as possessing certain properties, rather than designated as particulars. The same is true of microphysical particles, which quantum physics has no need to describe or explain as particulars, but only as kinds or types. The analogy moreover is instructive, since, according to the Heisenberg indeterminacy principle, there are no criteria of correctness in Wittgenstein's sense for the reidentification of particular quantum particles, whose position and momentum can never be simultaneously determined. If quantum particles can exist without being named as particulars, then so can particular private sensations. If this conclusion is correct, then Wittgenstein at most calls attention to a peculiarity of the limitations of attempts to name private sensations, and does not disprove the private or internal nature of sensation, nor the nonexistence of private sensations. Then, since psychology, like quantum mechanics, is meant to be a generalizing taxonomic, predictive, and explanatory science, it does not need to name or refer to mental occurrences as particulars, but only collectively and descriptively by predicates as types or kinds.[10]

6 Privacy of Generic Sensation Types

Wittgenstein nevertheless rejects psychological privacy in the epistemic sense, even for generic sensations. He reduces what otherwise passes for *epistemic* to *proprietary* privacy, and denies the phenomenological thesis that first-person experiencers can have knowledge of thought content by introspection or acquaintance. He maintains:

[9]The existence of unnameable particulars is anathema to the picture theory of meaning. If the interpretation is correct, it lends credence to accounts of Wittgenstein's later philosophy as a radical departure from the semantics of logical atomism in the *Tractatus Logico-Philosophicus* specifically with respect to the concept of naming. For a more complete discussion of Wittgenstein's early philosophy of mind, see Jacquette 1992-93.

[10]The covering law model of scientific explanation no more requires the naming of quantum particles or particular psychological experiences than does the formulation of general laws. It is enough to refer to particulars as falling under general descriptions in order to satisfy a law's conditional antecedents in order relevantly to detach conclusions about whatever nameable or unnameable particulars satisfy the description.

246 In what sense are my sensations *private*? – Well, only I can know whether I am really in pain; another person can only surmise it. – In one way this is wrong, and in another nonsense. If we are using the word 'to know' as it is normally used (and how else are we to use it?), then other people very often know when I am in pain. – Yes, but all the same not with the certainty with which I know it myself! – It can't be said of me at all (except perhaps as a joke) that I *know* I am in pain. What is it supposed to mean – except perhaps that I *am* in pain?

Other people cannot be said to learn of my sensations *only* from my behavior, – for *I* cannot be said to learn of them. I *have* them.

The truth is: it makes sense to say about other people that they doubt whether I am in pain; but not to say it about myself.

The conclusion follows from the knowing-doubting polarity, as a feature of the philosophical grammar of each of these concepts. If it does not make sense to doubt that I am in pain, then it does not make sense to say that I know it, and conversely. Yet the argument rests on considerations that seem to entail the impossibility of having any kind of knowledge *with certainty*, since to be certain by definition precludes the possibility of doubt. Perhaps it is only knowledge with less than absolute certainty that implies the possibility of doubt. Then the knowing-doubting polarity need not apply to introspective knowledge of first-person mental states, if, as some philosophers believe, introspection yields absolutely certain incorrigible knowledge of mental occurrence and content. In that case, the phenomenological epistemic privacy thesis once again escapes Wittgenstein's objection.

The private language argument in any event is independent of his argument from the knowing-doubting polarity. The private language argument by contrast with the polarity argument offers no aid and comfort to the hard psychological or reductive cognitive sciences against phenomenology, and seems on the contrary to presuppose the privacy of thought, minimally in the proprietary or private ownership sense, but also in the more interesting sense of private epistemic access or knowledge by direct acquaintance. As an objection to the naming of particular sensations, it provides no basis for rejecting private experience in preference to what is exclusively public and external, and is therefore neutral with respect to the truth of hard psychological sciences concerned exclusively with behavior, neurophysiology, or information flow and control functionalities.

Wittgenstein regards both affirming and denying substantive scientific psychological theories as beyond the proper scope of philosophy. His point in the private language argument is linguistic and semantic rather than ontological; it is about language, which is always his central concern, and the limitations of language, rather than the nature of mind. If there is a route from the philosophical grammar of naming to the metaphysics of thought,

Wittgenstein at least does not try to show the way.[11]

References

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

Jacquette, D. 1992-93 "Wittgenstein's Critique of Propositional Attitudes and Russell's Theory of Judgment", *Brentano Studien* 4, 193-220.

Jacquette, D. 1994 *Philosophy of Mind*, Englewood Cliffs: Prentice-Hall.

Paskow, A. 1974 "A Phenomenological View of the Beetle in the Box", *New Scholasticism* 48, 277-304.

Reeder, H. P. 1979 "Language and the Phenomenological Reduction: A Reply to a Wittgensteinian Objection", *Man and World* 12, 35-46.

Wittgenstein, L. 1922 *Tractatus Logico-Philosophicus*, ed. C.K. Ogden, London: Routledge & Kegan Paul.

Wittgenstein, L. 1953 *Philosophical Investigations*, ed. and trans. G.E.M. Anscombe, 3rd ed. 1968, New York: The Macmillan Company.

[11]I am grateful to The Pennsylvania State University for a Melvin and Rosalind Jacobs Research Fellowship in the Humanities in support of this project.

Finding the Mind in the Natural World

Frank Jackson

Conceptual analysis played a prominent role in the defence of materialism mounted by the Australian materialists and their American ally David Lewis. It was how they found a place for the mind within the material world. The leading idea is encapsulated in the following argument schema:

1. Mental state M = occupant of functional role F. (By conceptual analysis)
2. Occupant of role F = brain state B. (By science)

   Therefore, $M = B$. (By transitivity)

This schema gives the role of conceptual analysis in the Australian defence. But it does not tell us why conceptual analysis had to have a role in the defence. Indeed, the schema positively invites the thought that conceptual analysis was not needed. For to get the conclusion that $M = B$, all that is needed is the truth of the two premisses. It is not necessary that one of them be a conceptual truth. And I think, speaking more generally, that the Australian materialists left it unclear why materialists need to do some conceptual analysis. Nevertheless, I think that they were right that materialists need to do some conceptual analysis. This paper is a defence of this view. In a nutshell my argument will be that only by doing some conceptual analysis can materialists find a place for the mind in their naturalistic picture of the world. In a final section we will note the implications of our discussion for the knowledge argument.

In arguing for the necessity of conceptual analysis I am swimming against the tide. Current orthodoxy repudiates the role of conceptual analysis in the defence of materialism for at least three reasons. First, materialism is a doctrine in speculative metaphysics. And, runs the first reason, though conceptual analysis has a role in the philosophy of language and the study of concepts, it has no essential role when our subject is what the world is, at bottom, like. The second reason is that the history of conceptual analysis is the history of failure. For any proffered analysis someone clever always finds a counter-example. The final reason turns on the claim that we have learnt from Hilary Putnam and Saul Kripke about the necessary *a posteriori*, and that tells us that there can be necessary connections that, precisely by virtue of being *a posteriori*, are not revealed by or answerable to conceptual analysis. The materialist should, according to this line of thought, hold that the connection between the mental and the material or

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 101-112

physical is a necessary *a posteriori* one, and so not a matter accessible via conceptual analysis. During the course of the discussion we will see how to reply to each of these objections to the need for conceptual analysis in the defence of materialism.

The first step in our defence of the materialists' need for conceptual analysis is to note that materialism is a piece of what I will call serious metaphysics, and that, like any piece of serious metaphysics, it faces the location problem.

1 The Location Problem

Metaphysics is about what there is and what it is like. But it is concerned not with any old shopping list of what there is and what it is like. Metaphysicians seek a comprehensive account of some subject matter – the mind, the semantic, or, most ambitiously, everything – in terms of a limited number of more or less fundamental notions. Some who discuss the debate in the philosophy of mind between dualism and monism complain that *each* position is equally absurd. We should be *pluralists*. Of course we should be pluralists in some sense or other. However, if the thought is that any attempt to explain it all, or to explain it all as far as the mind is concerned, in terms of some limited set of fundamental ingredients is mistaken in principle, then it seems to me that we are being, in effect, invited to abandon serious metaphysics in favour of drawing up big lists. And we know we can do better than that. At least some of the diversity in our world conceals an underlying identity of ingredients. The diversity is a matter of the same elements differently selected and arranged. But if metaphysics seeks comprehension in terms of limited ingredients, it is continually going to be faced with the problem of location. Because the ingredients *are* limited, some putative features of the world are not going to appear explicitly in the story told in the favoured terms. The question then will be whether the features nevertheless figure *implicitly* in the story. Serious metaphysics is simultaneously discriminatory and putatively complete, and the combination of these two facts means that there is bound to be a whole range of putative features of our world up for either elimination or location.

What then is it for some putative feature to have a place in the story some metaphysic tells in its favoured terms? One answer is for the feature to be entailed by the story told in the favoured terms. Perhaps the story includes information about mass and volume in so many words, but nowhere mentions density by name. No matter – density facts are entailed by mass and volume facts. Or perhaps the story in the favoured terms says that many of the objects around us are nothing but aggregations of molecules held in a lattice-like array by various inter-molecular forces. Nowhere in the story in the favoured terms is there any mention of solidity. Should we then infer that nothing is solid, or at any rate that this particular metaphysic is committed to nothing being solid? Obviously not. The story in the

favoured terms will, we may suppose, tell us that these lattice-like arrays of molecules exclude each other, the inter-molecular forces being such as to prevent the lattices encroaching on each others' spaces. And *that* is what we understand by solidity. That's what it takes, according to our concept, to be solid. Or at least it is near enough. Perhaps pre-scientifically we might have been tempted to insist that being solid required being everywhere dense in addition to resisting encroachment. But resisting encroachment explains the stubbing of toes quite well enough for it to be pedantic to insist on anything more in order to be solid. Hence, solidity gets a location or place in the molecular story about our world by being entailed by that story, and we see this by asking ourselves about our concept of solidity in the sense of asking what it takes to be solid.

Thus, one way materialists can show that the psychological has a place in their world view is by showing that the psychological story is entailed by the story about the world told in the materialists' favoured terms. We will see, however, that it is not just one way; it is the one and only way.

2 Completeness and Supervenience

Materialism is the very opposite of a 'big list' metaphysics. It is highly discriminatory, operating in terms of a small set of favoured particulars, properties and relations, typically dubbed 'physical' – hence its other name, 'physicalism'; and it claims that a complete story, or anyway a complete story of everything contingent, including everything psychological, about our world can in principle be told in terms of these physical particulars, properties and relations alone. Only then is materialism interestingly different from dual attribute theories of mind.

Now what, precisely, is a complete story? We can make a start by noting that one particularly clear way of showing *incompleteness* is by appeal to independent variation. What shows that three co-ordinates do not provide a complete account of location in space-time is that we can vary position in space-time while keeping any three co-ordinates constant. Hence, an obvious way to approach completeness is in terms of the lack of independent variation. But, of course, lack of independent variation is supervenience: position in space-time supervenes on the four co-ordinates. So the place to look when looking for illumination regarding the sense in which materialism claims to be complete, and, in particular, to be complete with respect to the psychological, is at various supervenience theses.[1]

Now materialism is not just a claim about the completeness of the physical story concerning certain individuals or particulars in our world. It claims completeness concerning the world itself, concerning, that is, the total way things are. Accordingly, we need to think of the supervenience base as consisting of possible worlds – complete ways things might be. We need,

[1] What follows is one version of a familiar story. See, for example, T. Horgan 1982 and D. Lewis 1983.

accordingly, to look to global supervenience theses, an example of which is

(I) Any two possible worlds that are physical duplicates (physical property, particular and relation for physical property, particular and relation identical) are duplicates *simpliciter*.

But (I) does not capture what the materialists have in mind. Materialism is a claim about our world, the actual world, to the effect that its physical nature exhausts all its nature, whereas (I) is a claim about worlds in general. A more restricted supervenience thesis in which our world is explicitly mentioned is:

(II) Any world that is a physical duplicate of our world is a duplicate *simpliciter* of our world.

However, materialists can surely grant that there is a possible world physically exactly like ours but which contains as an addition a lot of mental life sustained in non-physical stuff, as long as they insist that this world is not our world. Consider the view of those theists that hold that materialism is the correct account of earthly existence but it leaves out of account the after-life. When we die our purely material psychology is reinstated in purely non-physical stuff. Surely materialists can grant that these theists are right about some world, some way things might be, as long as they insist that it is *not* our world, not the way things actually are. Hence, materialists are not committed to (II).

The trouble with (II) is that it represents materialists' claims as more wide ranging than they in fact are. What we need is something like (II) but that limits itself to worlds more nearly like ours, or at least more nearly like ours on the materialists' conception of what our world is like. I suggest

(III) Any world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world.

What is a minimal physical duplicate? Think of a recipe for making scones. It tells you what to do, but not what *not* to do. It tells you to add butter to the flour but does not tell you not to add whole peppercorns to the flour. Why doesn't it? Part of the reason is that no one would think to add them unless explicitly told to. But part of the reason is logical. It is impossible to list all the things *not* to do. There are indefinitely many of them. Of necessity the writers of recipes rely on an intuitive understanding of an implicitly included 'stop' clause in their recipes. A minimal physical duplicate of our world is what you would get if you used the physical nature of our world (including of course its physical laws) as a recipe in this sense for making a world.

We arrived at (III) by eliminating alternatives. But we can give a positive argument for the conclusion that the materialist is committed to (III). Suppose that (III) is false; then there is a difference in nature between our world and some minimal physical duplicate of it. But then either our world contains some nature that the minimal physical duplicate does not, or the minimal physical duplicate contains some nature that our world does not. The second is impossible because the extra nature would have to be non-physical (as our world and the duplicate are physically identical), and the minimal physical duplicate contains no non-physical nature by definition. But if our world contains some nature that the duplicate does not, this nature must be non-physical (as our world and the duplicate are physically identical). But then materialism would be false, for our world would contain some non-physical nature. Hence, if (III) is false, materialism is false – that is to say, materialism is committed to (III).

3 From (III) to Entry by Entailment

Given that (III) follows from materialism, there is a straightforward and familiar argument to show that if materialism is true, then the psychological story about our world is entailed by the physical story about our world.

We can think of a statement as telling a story about the way the world is, and as being true inasmuch as the world is the way the story says it is. Let $\mathbf{\Phi}$ be the statement which tells the rich, complex and detailed physical story that is true at the actual world and all and only the minimal physical duplicates of the actual world, and false elsewhere. Let $\mathbf{\Pi}$ be any true statement entirely about the psychological nature of our world: $\mathbf{\Pi}$ is true at our world, and every world at which $\mathbf{\Pi}$ is false differs in some psychological way from our world. If (III) is true, every world at which $\mathbf{\Phi}$ is true is a duplicate *simpliciter* of our world, and so a fortiori a psychological duplicate of our world. But then every world at which $\mathbf{\Phi}$ is true is a world at which $\mathbf{\Pi}$ is true – that is, $\mathbf{\Phi}$ entails $\mathbf{\Pi}$.

We have thus derived what we might call the *entry by entailment thesis*: a putative psychological fact has a place in the materialists' world view if and only if it is entailed by the physical story about the world. The one and only way of getting a place is by entailment.

4 From Entry by Entailment to Conceptual Analysis

How does entry by entailment show the importance of conceptual analysis? If $\mathbf{\Phi}$ entails $\mathbf{\Pi}$, what makes $\mathbf{\Phi}$ true also makes $\mathbf{\Pi}$ true (at least when $\mathbf{\Phi}$ and $\mathbf{\Pi}$ are contingent). But what makes $\mathbf{\Phi}$ true is the physical way our world is. Hence, the materialist is committed to each and every psychological statement being made true by a purely physical way our world is. But it is the very business of conceptual analysis to address which matters framed in terms of one set of terms and concepts are made true by which matters framed in a different set of terms and concepts. For instance, when we seek an analysis of knowledge in terms of truth, belief, justification, causation and so on, we seek an account of how matters described in terms of the latter notions make true matters described in terms of the former. When we seek an account of reference, we seek an account of the kinds of causal and

descriptive facts which make it true that a term names an object. When and if we succeed, we will have an account of what makes it true that 'Moses' names Moses in terms of, among other things, causal links between uses of the word and Moses himself. And so on and so forth.

How could the *a priori* reflections on, and intuitions about, possible cases so distinctive of conceptual analysis be relevant to, for instance, the causal theory of reference? Well, the causal theory of reference is a theory about the conditions under which, say, 'Moses' refers to a certain person. But that is nothing other than a theory about the possible situations in which 'Moses' refers to that person, and the possible situations in which 'Moses' does not refer to that person. Hence, intuitions about various possible situations – the meat and potatoes of conceptual analysis – are bound to hold centre stage. (This is particularly true when the test situations cannot be realised. We cannot, for instance, make twin earth to check empirically what we would say about whether XYZ is water.)

The alternative is to *invent* our answers. Faced with the question, say, of whether the physical way things are makes true the belief way things are, we *could* stipulate the conditions under which something counts as a belief in such a way as to ensure that there are beliefs, or, if we preferred, that there are no beliefs. But that would not bear on whether beliefs according to *our* concept have a place in the materialists' picture of things, only on whether beliefs according to the stipulated concept have a place. In order to address the question of whether beliefs as we understand them have a place, what else can we do but consult and be guided by our honed intuitions about what counts as a belief? Would it be better to invent, or to go by what seems counter-intuitive?

I should emphasise, though, that a sensible use of conceptual analysis will allow a limited but significant place for *a posteriori* stipulation. We mentioned earlier the example of finding a place for solidity in the molecular picture of our world, and the fact that what the molecular picture vindicates is the existence of solid bodies according to a conception of solidity cashed out in terms of mutual exclusion rather than in terms of the conjunction of mutual exclusion and being everywhere dense. For our day to day traffic with objects, it is the mutual exclusion that matters, and accordingly it is entirely reasonable to rule that mutual exclusion is enough for solidity. The role of conceptual analysis of K-hood is not always to settle on a nice, neat, *totally a priori* list of necessary and sufficient conditions for being a K – indeed, that is the task that has so often been beyond us. It is rather to guide us in dividing up the cases that clearly are not cases of a K, from the cases that a principle of charity might lead us to allow as cases of a K. Then, armed with this information, we are in a position to address the question of whether some inventory of fundamental ingredients does, or does not, have a place for Ks.

I should also emphasise that the contention is not that *a priori* reflection on possible cases gives us new information, let alone some sort of infallible new information, about what the world is like. The reflection is *a priori* in the sense that we are not consulting our intuitions about what would *happen* in certain possible cases – it is not like the famous thought experiments in science – rather we are consulting our intuitions about how to *describe* certain possible cases. And what we learn (in the sense of making explicit) is not something new about what the world is like, but something about how, given what the world is like as described in one set of terms, it should be described in some other set of terms. Perhaps the point is clearest in the example about finding solidity in the molecular account of the objects around us. Reflection on our concept of solidity tells us that the molecular account includes solidity, but it does not tell us that solidity is an addition to what appears in the molecular account of objects, let alone an infallible one.

5 The Objection from the Necessary *a posteriori*

It might well be urged that the argument given above from (III) to the conclusion that $\mathbf{\Phi}$ entails $\mathbf{\Pi}$ is undermined by the existence of necessary *a posteriori* truths. The objection can be put in two different ways. Consider

> Over 60% of the Earth is covered by H_2O.
>
> Therefore, over 60% of the Earth is covered by water.

One way of putting the objection is that although every world where the premise is true is a world where the conclusion is true, the argument is not valid because the premise does not entail the conclusion in the relevant sense. It is not possible to move *a priori* from the premise to the conclusion. The premise fixes the conclusion without entailing it, as it is sometimes put. Likewise, for all we have shown by the considerations based on (III), $\mathbf{\Phi}$ fixes $\mathbf{\Pi}$ but does not entail it.

This way of putting the objection makes it sound like a quarrel over terminology. It invites the response of distinguishing entailment *simpliciter*, the notion cashed out simply in terms of being necessarily truth-preserving, from *a priori* or, as it is sometimes called, conceptual, entailment, the latter being the notion tied to *a priori* deducibility. But the real objection, of course, is that the necessarily truth-preserving nature of the passage from 'Over 60% of the Earth is covered by H_2O' to 'Over 60% of the Earth is covered by water' is not one that can in principle be revealed by conceptual analysis. Reflection on, and intuitions about, possible cases and concepts, unless supplemented by the *a posteriori* information that water is H_2O, will get you nowhere. Materialists, it seems, can allow that (III) forces them to admit a necessarily truth-preserving passage from $\mathbf{\Phi}$ to $\mathbf{\Pi}$, without allowing a role for conceptual analysis. They can simply insist that the entailment from $\mathbf{\Phi}$ to $\mathbf{\Pi}$ is an *a posteriori* one.

We will see, however, that acknowledging the necessary *a posteriori* does not alter matters in any essential respects as far as the importance

of conceptual analysis goes. The argument to this conclusion turns on a negative claim about the nature of the necessity possessed by the necessary *a posteriori*, and a consequent view about the role of conceptual analysis, in the sense of intuitions about possibilities, in the detection of the necessary *a posteriori*.

6 The Necessity of the Necessary *a posteriori*

There are two different ways of looking at the distinction between necessary *a posteriori* statements like 'Water = H_2O' and necessary *a priori* ones like 'H_2O = H_2O' (all necessary modulo worlds where there is no water, of course). You might say that the latter are analytically or conceptually or logically (in some wide sense not tied to provability in a formal system) necessary, whereas the former are metaphysically necessary, meaning by the terminology that we are dealing with two senses of 'necessary' in somewhat the way that we are when we contrast logical necessity with nomic necessity. On this approach, the reason the necessity of water's being H2O is not available *a priori* is that its necessity is not the kind that is available *a priori*.

I think, as against this view, that it is a mistake to hold that the necessity possessed by 'Water = H_2O' and 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by water' is different from that possessed by 'Water = water' and 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by H2O'. Just as Quine insists that numbers and tables exist in the very same sense, I think that we should insist that water's being H_2O and water's being water are necessary in the very same sense.

My reason for holding that there is one sense of necessity here relates to what it was that convinced us that 'Water = H_2O' is necessarily true. What convinced us were the arguments of Saul Kripke and Hilary Putnam about how to *describe* certain possibilities, rather than arguments about what is possible *per se*. Kripke and Putnam convinced us that a world where XYZ plays the water role – that is, satisfies enough of (but how much is enough is vague): filling the oceans, being necessary for life, being colourless, being called 'water' by experts, being of a kind with the exemplars we are acquainted with, and so on – did not warrant the description 'world where water is XYZ', and the stuff correctly described as water in a counterfactual world is the stuff – H_2O – which fills the water role in the actual world. The key point is that the right way to describe a counterfactual world sometimes depends in part on how the actual world is, and not solely on how the counterfactual world is in itself. The point is not one about the space of possible worlds in some newly recognised sense of 'possible', but instead one about the role of the actual possible world in determining the correct way to describe certain counterfactual possible worlds – in the sense of 'possible' already recognised.

All this was, it seems to me, an exercise in conceptual analysis. We had an old theory about the meaning of 'water', namely, that it meant 'that which fills the water role', a theory that was refuted by appealing to our intuitions about how to describe possible worlds in which something different from that which actually fills the water role fills the water role. We became convinced of a new theory – again by reflection on possible cases, the meat and potatoes of conceptual analysis – according to which 'water' is a rigid designator of the stuff that fills the water role in the actual world. At no time did we have to recognise a new sort of possibility, only a new way for something in some counterfactual situation to count as a K, namely, by virtue not solely of how things are in that counterfactual situation, but in part in virtue of how things actually are.

If this is right, the inference

Over 60% of the Earth is covered by H_2O.

Therefore, over 60% of the Earth is covered by water.

is not an example of an a posteriori entailment that shows the irrelevance of conceptual analysis to the question of whether an a posteriori entailment holds. For it is conceptual analysis that tells us, in light of the fact that H_2O fills the water role, that the entailment holds.

7 Two-Dimensionalism and the Knowledge Argument

I have argued that materialists must hold that the complete story about the physical nature of our world given by Φ entails everything about our psychology, and that such a position cannot be maintained independently of the results of conceptual analysis. But it is quite another question whether they must hold that Φ *a priori* entails everything about our psychology, including its phenomenal side, and so quite another question whether they must hold that it is in principle possible to deduce from the full physical story alone what it is like to see red or smell a rose – the key assumption in the knowledge argument that materialism leaves out *qualia*. I will conclude by noting how the two dimensional treatment of the necessary *a posteriori* – the obvious treatment of the necessary *a posteriori* for anyone sympathetic to the view that such necessity is not a new sort of necessity – means that materialists are committed to the *a priori* deducibility of the phenomenal from the physical.

If the explanation of the *a posteriori* nature of the necessary *a posteriori* does not lie in the special necessity possessed, where does it lie? Two dimensionalists insist that the issue is an issue about sentences, and not about propositions, or at least not propositions thought of as sets of possible worlds. For, by the conclusion that we are not dealing with a new sort of necessity, the set of worlds where water is water is the very same set as the set where water is H_2O, and so, by Leibnitz's Law, there is no question of the proposition that water is water differing from the proposition that water

is H_2O in that one is, and one is not, necessary *a posteriori*. Their contention is that there are sentences such that the proposition expressed by them depends on the context of utterance.[2] We understand them in that we know how the proposition expressed depends on the context, but if we do not know the relevant fact about the context, we will not know the proposition expressed. (In Robert Stalnaker's terminology, we know the propositional concept but not the proposition; in David Kaplan's, we know the character but not the content.[3]) Consider 'Over 60% of the Earth is covered by water'. Because 'water' is a rigid designator whose reference is fixed by 'the stuff that fills the water role', someone who does not know what that stuff is does not know which proposition the sentence expresses, but they understand the sentence by virtue of knowing how the proposition expressed depends on how things actually are, and, in particular, this being the relevant contextual matter in this case, on what actually fills the water role. The explanation of the necessary *a posteriori* status of 'If over 60% of the Earth is covered by H_2O, then over 60% of the Earth is covered by water' then runs as follows. The proposition expressed by the sentence 'Over 60% of the Earth is covered by H_2O', is the same as the proposition expressed by 'Over 60% of the Earth is covered by water', and so the proposition expressed by the conditional sentence is *a priori* and necessary. But consistent with what is required to count as understanding the conditional sentence, it is contingent and *a posteriori* that it expresses a necessary *a priori* proposition.

I should emphasise that this does not mean that people who fully understand a sentence like 'Over 60% of the Earth is covered by water' but do not know that water is H_2O do not, in some perfectly natural sense, know the conditions under which what they are saying is true.[4] True, full understanding of the sentence does not in itself yield which proposition is expressed by the sentence, but knowledge of the way in which the proposition expressed depends on context, combined with knowledge of the truth conditions of the various propositions, does enable them to say when the sentence they produce is true. For their knowledge about how the proposition expressed depends on context together with the conditions under which the various propositions are true is given in the following array:

> If H_2O fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true iff over 60% of the Earth is covered by H_2O.
>
> If XYZ fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true iff over 60% of the Earth is covered by XYZ.
>
> If — fills the water role, then 'Over 60% of the Earth is covered by water' expresses a proposition that is true iff over 60% of the Earth is covered by —.

For each distinct, context-giving, antecedent, a distinct proposition is expressed by the sentence. Nevertheless, simple inspection of the array shows that the sentence is true iff over 60% of the Earth is covered by the stuff that fills the water role. That is the sense in which the fully understanding producer of the sentence knows when the sentence is true.[5]

Now, to return to the main plot, although understanding alone does not necessarily give the proposition expressed by certain sentences – that is how they can be necessary and yet this fact be in principle not accessible to understanding plus acumen alone, that is how they can be necessary a posteriori – understanding alone does give us the way the proposition expressed depends on context; and that fact is enough for us to move a priori from, for example, sentences about the distribution of H_2O combined with the right context-giving sentences, to information about the distribution of water. Consider, for instance, a supplementation of our earlier inference:

(1) Over 60% of the Earth is covered by H_2O.

(2) H_2O fills the water role.

(3) Therefore, over 60% of the Earth is covered by water.

Although, as noted earlier, the passage from (1) to (3) is necessarily truth-preserving but *a posteriori*, being an *a posteriori* entailment, the passage from (1) and (2), to (3) is *a priori*. And it is so because, although our understanding of 'Over 60% of the Earth is covered by H_2O' does not in itself yield the proposition expressed by the sentence, it yields how the proposition depends on context, and (2) gives that context. (2) gives the relevant fact about how things are "outside the head". We did not know that (1) entailed (3) until we learnt (2), because we did not, and could not, have known that (1) and (3) express the same proposition until we learnt (2). But as soon as we learn (2), we have the wherewithal, if we are smart enough, to move *a priori* to (3).

The point, then, is that the necessary *a posteriori* nature of 'Water = H_2O' does not mean that the fact that the H_2O way things are entails the water way things are is not answerable to our grasp of the relevant concepts plus acumen. It means, rather, that we need to tell a rich enough story about the H_2O way things are, a story that includes the crucial contextual information, before we can move from the H_2O way things are to the water way they are using our grasp of the concepts alone.

[2] I take it that what follows is a sketch of the approach suggested by the version of two-dimensionalism in Stalnaker1978 .

[3] Stalnaker 1978 and Kaplan 1978.

[4] I am indebted here to David Lewis and David Chalmers.

[5] This observation bears on the dispute about whether Earthians and Twin Earthians believe alike. Although the sentence 'Water is plentiful' expresses different propositions in the mouths of the Earthians and the Twin Earthians, they agree about when the sentence is true, and so in *that* sense agree in belief.

More generally, the two-dimensional way of looking at the necessary *a posteriori* means that even if the entailment the materialist is committed to from some physical story about the world to the full psychological story is *a posteriori*, there is still an *a priori* story tellable about how the story in physical terms about our world makes true the story in psychological terms about our world. Although understanding may not, even in principle, be enough to yield the proposition expressed by the physical story, understanding and logical acumen is enough to yield how the proposition expressed depends on context. But, of course, the context is, according to the materialist, entirely physical. The context concerns various matters about the nature of the actual world, and that nature is capturable in entirely physical terms according to the materialist. Hence, the materialist is committed to there being an *a priori* story to tell about how the physical way things are makes true the psychological way things are. But the story may come in two parts. It may be that one part of the story says which physical way things are, Φ_1, makes some psychological statement true, and the other part of the story, the part that tells the context, says which different physical way things are, Φ_2, makes it the case that it is Φ_1 that makes the psychological statement true. What will be a priori accessible is that Φ_1 and Φ_2 together make the psychological statement true.[6]

References

Terence Horgan 1982 "Supervenience and Microphysics", *Pacific Philosophical Quarterly* 63, 29–43.

David Kaplan 1978 "Dthat", in P. Cole, (ed.), *Syntax and Semantics* Vol. 9, New York: Academic Press.

David Lewis 1983 "New Work for a Theory of Universals", *Australasian Journal of Philosophy* 61, 343–377.

Robert C. Stalnaker 1978 "Assertion" in P. Cole (ed.), *Syntax and Semantics* Vol. 9, New York: Academic Press, pp. 315 - 332.

[6]I am indebted to Lloyd Humberstone, David Chalmers, David Lewis, Michael Smith, and Philip Pettit.

Logic and Physicalism

Neil Tennant and Frank Jackson

Neil Tennant

1 Introduction

Frank Jackson sets out, in his paper 'Finding the Mind in the Natural World', to show that "only by doing some conceptual analysis can materialists find a place for the mind in their naturalistic picture of the world." To do this he needs a formulation of materialism as a supervenience thesis. The supervenience thesis Jackson favours is his:

(III) Any world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world.

Jackson then makes two important entailment claims, backed by argument:

(A) Materialism entails (III).

(B) (III) entails that the physical story about our world entails the psychological story about our world.

By transitivity, then, if materialism were true then the physical story would tell it all. Finally, he argues for what I shall call his *a prioricity thesis*:

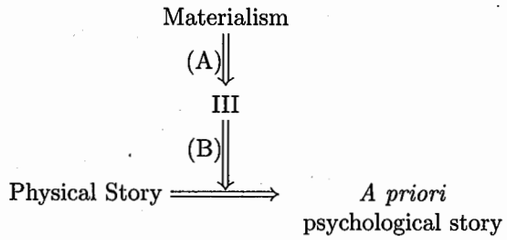
(C) The materialist is committed to there being an *a priori* story to tell about how the physical way things are makes true the psychological way things are.

(B) is an arresting claim. It is rejected by many writers on materialism, supervenience and reductionism. The orthodox view is that materialism, explicated as some form of supervenience claim, does *not* entail reductionism. That is, it does not entail any claim to the effect that the mental story can be obtained from the physical story by entailment. (It does not matter whether the entailment is direct, or mediated by appropriate identifications and definitions.)

I disagree with each of (A), (B) and (C). My aim here is to set out criticisms of the arguments that Jackson gives in support of each of these false claims.

In summary, Jackson's aim is to establish the following entailments (A), (B) and (C):

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 113-126



I shall show that (A) fails. Then I shall show that even if (A) holds, (B) fails. Finally, I shall show that even if (B) holds, (C) fails.

2 The Failure of (A)

A minimal physical duplicate of our world would be obtained by putting into it just those physical features of our world, and nothing more. Jackson illustrates this idea by appeal to a recipe for making scones:

> A minimal physical duplicate of our world is what you would get if you used the physical nature of our world (including of course its physical laws) as a recipe... for making a world.

Note, however, that Jackson can't be intending the phrase "for making a world" to cover the world in all its possibly *non*-physical respects as well as its physical respects. For on *that* wide a reading a world that was physically just like ours, but in which all creatures were zombies, would (according to (III)) be a duplicate *simpliciter* of our world. Hence all the creatures in our world would be zombies. But materialism, properly conceived, cannot be committed to *that*. Thus Jackson's elucidation of (physical world)-recipes should read:

> A minimal physical duplicate of our world is what you would get if you used the physical nature of our world (including of course its physical laws) as a recipe... for making a world *in its physical respects.*

We shall consider first an example showing that the converse of (A) fails. Imagine a world, which I shall call World, in which there is just one physical thing, a rock. It has no mental life. World also contains exactly one *non*-physical thing: a disembodied philosophical intelligence, which happens to be pondering the truth, in World, of the thesis of materialism (as it concerns World). If this philosophical intelligence were to take Jackson's line, it would think that materialism about World entails:

(III$_{\text{World}}$) Any world that is a *minimal* physical duplicate of World is a duplicate *simpliciter* of World.

Consider a minimal physical duplicate of World, namely Otherworld. Otherworld contains exactly one rock, just like the one in World. That, after all, is all that the *physical* World-recipe calls for. But the truth of (III$_{\text{World}}$) requires that Otherworld should be a duplicate *simpliciter* of World. Hence Otherworld also contains a disembodied philosophical intelligence, which is *non*-physical. Provided only that all worlds that were minimal physical duplicates of World were to contain such non-physical disembodied philosophical intelligences, the truth of (III$_{\text{World}}$) would be sustained. But, in the nature of these cases imagined while yet sustaining (III$_{\text{World}}$), materialism would be dramatically false. For a materialist, there could be no such thing as a disembodied philosophical intelligence.

We have shown that the *converse* of (A) above is false. (III$_{\text{World}}$) could be true in such a way as to falsify materialism. But (A) itself, Jackson would say, is still true: materialism about World entails (III$_{\text{World}}$).

Not, unfortunately, so. To see this, we shall presently vary the thought experiment slightly. But first, let me explain what I mean by the *laws of epiphenomenal emergence* for a given world. These are the laws that tell us what non-physical, epiphenomenal traits arise, within that world, out of the physical bases within it. The emergent traits are epiphenomenal in that they do not "feed back" causally into the *physical* happenings within the world in question.

Here now is the slightly varied thought experiment promised. World is the world with one rock, as before. The rock has no mental life. But that is *all* that there is in World. So World has no non-physical things in it. The minimal physical duplicate Otherworld of World contains the rock, and *nothing else*; but the rock, in Otherworld, *does* have a mental life. Let it be ever so spasmodic: just once, let us say, it runs through Descartes' *cogito ergo sum.* (Do not ask *how we would know this*; one is free, by the rules of this game, to stipulate that this, metaphysically, is how things actually *are* in the world Otherworld). The laws of epiphenomenal emergence in Otherworld are different, then, from those in World. So we have the minimal physical duplicate Otherworld of World *not* being a duplicate *simpliciter* of World, because in World, but not in Otherworld, the rock has no mental life.

World may be dull, but it is no embarrassment to the materialist. Otherworld is a little more exciting, *but still need not be an embarrassment to the materialist*! Why? Well, for the materialist, all that's important is that the physical should determine the non-physical, in the sense that any difference in the non-physical (such as the mental) facts would have to be subtended by some difference in the physical facts. So if, say, the rock in Otherworld were running through the ontological argument rather than the *cogito*, this would be because of some physical difference in the rock: it was made of quartz, say, instead of granite.[1] All that is important for the

[1]Note that it is no difficulty for my argument that this would make Otherworld an inexact physical duplicate of World. I can settle for its being impossible for the rock in World to be running through any philosophical argument but the *cogito*. Its being made

materialist is that, once the physical facts are in, then all the non-physical (i.e. mental) facts are fixed. It's the rock's being made of granite that would make it the case that it would be the *cogito* that it ran through; and it's the rock's being made of quartz that would make it the case that it would be the ontological argument that it ran through. Do not ask me why; that's just how it happens to be. (Remember, I get to make up these worlds; and I'm a materialist in doing so.)

The upshot is: materialism can happily countenance the failure of (III). Therefore materialism does not entail (III), contrary to what Jackson claims. (A) is false.

If the worlds with the rocks are too far-fetched for my philosophical audience, let me make the same case with a pair of worlds nearer to home: one as close as one can get, the other not too many Lewisons away.[2] World is the actual world. Otherworld is a minimal physical duplicate of World. That is, Otherworld contains the same (sorts of) physical things, distributed the same way through space and time, as World contains, and also has the same physical laws as World. Consistently with this requirement of minimal physical duplication, however, I stipulate that Otherworld has different laws of epiphenomenal emergence from World. The laws of epiphenomenal emergence in Otherworld make it the case that the Doppelgänger, in Otherworld, of Jack Smart in World, has telepathic, empathetic insight into what it is like to be a cricket. This insight is into precisely those aspects of cricket-being that manifest themselves anyway in crickets' observable behaviour. Thus Doppelgänger-Jack Smart's physical dealings with crickets are, in Otherworld, just as they are in World.

So we have a minimal physical duplicate Otherworld of the actual world World, but Otherworld is not a duplicate *simpliciter* of World, because in World Jack Smart has no telepathic cricket-empathy. That is, (III) is false. If Jackson were right about materialism entailing (III), then our thought experiment would have refuted materialism.

But this, the materialist would say, cannot be; it is all a little too swift. Swift it is indeed, but only by courtesy of Jackson's claim that materialism entails (III)! Better to hang on to materialism, I would say, and regard as fishy the alleged entailment taking one from materialism to (III).

For the materialist *about World* need not be at all put out by the unusual goings-on (by World's standards) in *Otherworld.* And the materialist about *Otherworld* need not be at all put out by Doppelgänger-Jack Smart's cricket-empathy within Otherworld. It still supervenes on the physical happenings within Otherworld! – or so she could claim. If that empathy were different, it would have to be because of some difference in the physical nature of Otherworld. ("No empathy-change without some physical change" is how the supervenience slogan would specialize to the case at

of granite still fixes that *that* is what it's thinking.

[2]The Lewison: the basic unit for measuring distances between possible worlds.

hand.) If Doppelgänger-Jack Smart empathized with a given cricket some way other than the way he actually does in Otherworld, this could only be because his neurological firings, or those of the cricket, were in some respect different from the way they actually are in Otherworld. I am maintaining materialism-via-supervenience; the difference is just that I don't understand supervenience to be captured by Jackson's (III).

Jackson was right to look to some form of supervenience claim to capture the essence of the thesis of materialism. But (III) is the wrong sort of supervenience claim to plump for. (III) is far too strong as an explication of materialism in possible-worlds terminology. (III) can be counterexemplified without violation to one's materialist convictions, properly conceived. Hence (III) is not entailed by materialism, properly conceived. And we saw earlier that (III) does not entail materialism, properly conceived. (III) and materialism are at logical cross-purposes. (III) is neither necessary nor sufficient for materialism, properly conceived.

3 The failure of (B)

We have seen, then, that Jackson's entailment claim (A) is false. I proceed now to his entailment claim (B). Jackson endorses a "straightforward and familiar argument" which purports to show that (III) entails that "the psychological story about our world is entailed by the physical story about our world". With no misrepresentation of its essential structure, but with a little extra detail supplied, Jackson's argument is as follows:

> Let Δ be the statement which tells the true physical story about our world.
> Let Π be any true statement entirely about the psychological nature of our world.
> Let W be an arbitrary world at which Δ is true.
> Then W is a minimal physical duplicate of our world.
> By (III), W is a duplicate *simpliciter* of our world.
> Hence W is a psychological duplicate of our world.
> Thus Π is true in W.
> But W was arbitrary. Hence any world making Δ true makes Π true also. That is, Δ entails Π.

Unfortunately, this argument can be faulted at two of its steps. One of the faults can be corrected; the other cannot. First, we have no guarantee that there is a unique sentence Δ that tells the full physical story about our world. The full physical story may not be finitely axiomatizable. Suppose, however, that we get round this objection by appealing instead to some (possibly infinite) *theory* Q instead of a single sentence Δ. The corrected argument would then read:

> Let Q be the theory which tells the true physical story about our world.

Let Π be any true statement entirely about the psychological nature of our world.
Let W be an arbitrary world at which Q is true.
Then W is a minimal physical duplicate of our world.
By (III), W is a duplicate *simpliciter* of our world.
Hence W is a psychological duplicate of our world.
Thus Π is true in W.
But W was arbitrary. Hence any world making Q true makes Π true also. That is, Q entails Π.

This argument, however, is still defective. The fallacious step is the move

W is an arbitrary world at which Q is true;
So, W is a minimal physical duplicate of our world.

The reason why this is defective is that the theory Q need not be *categorical*. That is, Q might have distinct non-isomorphic models, even of the same cardinality. Our world may be but one among many distinct, non-isomorphic models of Q. So another of them, chosen as W, need not be a minimal physical duplicate of our world. This objection is fatal to the argument in its present form, and I do not see how to correct it.

Jackson is therefore deprived of his so-called "entry by entailment" thesis (B). It follows that his subsequent efforts are misdirected. These efforts are directed to maintaining (C) in the face of the anticipated objection that necessary but *a posteriori* truths (such as 'Water is H_2O') may be involved as premisses of the entailments in question.

4 The Failure of (C)

By opting for the so-called two-dimensional treatment of the necessary *a posteriori*, Jackson hopes to establish (C), the claim that "materialists are committed to the *a priori* deducibility of the phenomenal from the physical". Even though we have seen already a fatal objection to Jackson's derivation of entry by entailment, let us set that objection aside for the time being in order to look more closely at his proposed defence of his (false) a prioricity thesis (C) against the objection that he anticipates would be based on the necessary *a posteriori*.

In his discussion of the objection he deals exclusively, by way of illustration, with a case involving the necessary *a posteriori* truth 'Water is H_2O'. Consider Jackson's example inference

(1) Over 60% of the Earth is covered by H_2O

(2) H_2O fills the water rôle on Earth

(3) Therefore, over 60% of the Earth is covered by water.

(1) is contingent *a posteriori*. (2) is necessary *a posteriori*. (3) is contingent *a posteriori*. For Jackson, (1) on its own does entail (3), in that truth is necessarily preserved from (1) to (3). But the entailment of (3) by (1) is not *a priori*. The entailment of (3) by (1) *and* (2), however, is *a priori*. And that's the nub of (C). That (2), if true, is necessary, is a matter revealed to us, according to Jackson, by conceptual analysis in the philosophy of language. Jackson describes himself as having argued that materialists must hold that the complete story about the physical nature of our world given by Δ entails everything about our psychology, and that such a position cannot be maintained independently of the results of conceptual analysis.

That is, in the context of his example inference above, (1) entails (3), but this cannot be maintained independently of the results of that conceptual analysis that guarantees necessity for truths such as (2).

Jackson's example, however, was about water and H_2O, and 'water' is hardly a *psychological* term. Nevertheless, his example is suggestive enough to lead one to see how others might be constructed for suitably psychological terms. I would venture the following as an exact analogue of Jackson's inference above, designed to make rather more to the point his description of what he takes himself to have argued for:

(1*) Sticking pins into people causes their C-fibres to fire

(2*) C-fibre-firing fills the pain rôle for human beings on Earth

(3*) Sticking pins into people causes them pain.

(1*) is contingent *a posteriori*. (2*) (if true) is necessary *a posteriori*. (3*) is contingent *a posteriori*. For Jackson, (1*) on its own would entail (3*), in that truth would necessarily be preserved from (1*) to (3*), should (2*) be true (hence necessary). But the entailment of (3*) by (1*) would not be *a priori*. The entailment of (3*) by (1*) *and* (2*), however, would be *a priori* (as (C) contends). That (2*), if true, is necessary, is a matter revealed to us, according to Jackson, by conceptual analysis in the philosophy of language.

I think we have here an analogue that Jackson would admit as adequate for the purpose of making his general point about the rôle for conceptual analysis. But now let us look at his closing description of the general situation he has sketched:

> ... the materialist is committed to there being an *a priori* story to tell about how the physical way things are makes true the psychological way things are. But the story comes in two parts. It may be that [the] one part of the story say[ing] which physical way things are, Δ_1, makes some psychological statement [Y] true, and the other part of the story, the part that tells the context, say[ing] which different physical way things are, Δ_2, makes it the case that it is Δ_1 that makes the psychological statement [Y] true. What will be *a priori* accessible is that Δ_1 and Δ_2 *together* make the psychological statement [Y] true.

This general gloss obviously calls for the correspondences

$$(1)/(1^*) - \Delta_1 \quad (2)/(2^*) - \Delta_2 \quad (3)/(3^*) - \Upsilon$$

with reference to the examples given above. Note that the unstarred ones came from Jackson; while the starred ones were the ones that I supplied by faithful analogy, as more apt for making the point about *psychological* terms.

But now something objectionable emerges upon this clarification. Jackson's gloss on Δ_2, "the other part of the story, the part that tells the context, say[ing] which different physical way things are" is fine when applied to his own sentence

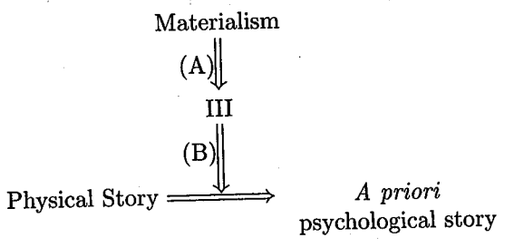
(2) H_2O fills the water rôle on Earth.

But what about the sort of sentence he should have been dealing with in connection with *psychological* terms? The sort of sentence in question would be one like

(2*) C-fibre-firing fills the pain rôle for human beings on Earth.

Does *this* sentence really deserve to be glossed as "say[ing] which different *physical* way things are"? Isn't the whole point that "pain" is a *psychological* term, not a *physical* term? It would appear that Jackson's use of the water-H_2O examples (in his discussion of the rôle of conceptual analysis via the theory of rigid designation) has misled him into thinking that the sentences *really* needed for slot (2) in his generic example would be telling a purely *physical* story. In this he is mistaken. Even if we grant him his "entry by entailment" thesis (B), Jackson has failed to establish his a prioricity thesis (C): "the materialist is committed to there being an *a priori* story to tell about how the physical way things are makes true the psychological way things are."

To repeat my earlier summary: Jackson's aim was to establish the following entailments (A), (B) and (C):



I have shown that (A) fails. But even if (A) holds, (B) fails. Moreover, even if (B) holds, (C) fails.

The root problem underlying these failures, I believe, is that Jackson characterized the concept of supervenience inadequately at the outset. This is not the place to rehearse an alternative explication of supervenience; I shall only mention that there is a model-theoretic explication, due to Hellman and Thompson, which to my mind is much more promising.[3] It gives rise to a technically more demanding problematic: to establish whether, given supervenience, it would follow that one could reduce the theory of the supervening level to the theory of the subvening level. Beth's Theorem was invoked by George Bealer in attempt to show that reductionism would indeed follow from supervenience.[4] Hellman and Thompson tried to block the argument via Beth's Theorem. Their strategy, however, does not work; but it turns out that there are other, better reasons why reductionism is not entailed by supervenience.[5] In my view the orthodox position (supervenience plus anti-reductionism) can be defended. But to rest secure with it, the debate has to move from the relatively crude devices of possible worlds semantics to the more satisfactory intricacies of model-theoretic semantics. The resulting philosophical conviction (as to the correct view) is worth the technical effort; but that is another story.[6]

Frank Jackson

Neil Tennant correctly notes that my discussion can be divided into three parts: in the first I argue that materialists are committed to a certain supervenience thesis (III); in the second I argue that (III) commits them to the entailment of the psychological by the physical; in the third I argue that this entailment in turn commits materialists to the possibility of *a priori* deducing the psychological way things are from the physical way things are. He objects to each part. I think that his objections involve misunderstandings of the (difficult) issues and of the detail of my arguments. I will presuppose familiarity with both my original paper and his reply, but it will be necessary to cover a certain amount of old ground.

1 Does Materialism Entail (III)?

As Tennant notes, I argue that materialism amounts to

(III) Any world that is a minimal physical duplicate of our world is a duplicate *simpliciter* of our world.

[3] G.Hellman and F.Thompson, "Physicalism, Ontology and Reduction", *The Journal of Philosophy* 72 (1975), 551–564.

[4] G.Bealer, "An inconsistency in functionalism", *Synthese* 38 (1978), 333–372.

[5] N.Tennant, "Beth's Theorem and Reductionism", *Pacific Philosophical Quarterly* 66 (1985) 342–54.

[6] I am grateful for the editorial invitation, after the Colloquium discussion of Frank Jackson's provocative paper, to write up my comments on it; and to Frank himself for providing me with a draft so that I could do so.

although the argument in the second two parts only uses the weaker claim that materialism entails (III). Tennant argues that (III) does not entail materialism, and, more to the point for the argument in the second two parts, that materialism does not entail (III).

The arguments he offers, however, involve a misunderstanding of what 'a minimal physical duplicate' means in (III). A minimal physical duplicate of our world (or of any world w) is a world that (a) is exactly like our world (or w) in every physical respect (property for property, particular for particular, law for law,relation for relation), and (b) contains nothing else in the sense of nothing more than it *must* to satisfy (a). Clause (b) is a kind of "no gratuitous additions" clause that I sought to give intuitive expression to with the recipe metaphor. Tennant's misunderstanding of the notion of a minimal physical duplicate runs right through his discussion in section 2, but I will focus on his main argument against my claim that materialism entails (III).

I offered (III) as an expression of what it takes for materialism to be true at our, the actual, world. It is though clear how to generalise it to give an account of what it takes for materialism to be true at a world w, namely

($\mathbf{III}_w$) Any world that is a minimal physical duplicate of w is a duplicate *simpliciter* of w.

(III_w) will be false of many worlds, but those it is true of are the worlds where materialism is true. Tennant in effect gives his objection as one to the view that materialism at w entails (III_w) for a nice simple w. This is fair enough. I agree that a corollary of what I say is that the truth of materialism at w entails (III_w).

He describes two distinct possible worlds, called 'World'and 'Otherworld'. World contains just one physical, inanimate rock; Otherworld is physically exactly like World – its physical nature is exhausted by its containing a rock exactly like the rock in World – but it has in addition some laws of "epiphenomenal emergence" that ensure that this single rock has a little bit of mental life. His objection is that materialism is true at World, but it is not true that every minimal physical duplicate of World is a duplicate *simpliciter* of World, for Otherworld is a minimal physical duplicate of World but is not a duplicate *simpliciter* of World. Tennant is right that materialism istrue at World but wrong that Otherworld is a minimal physical duplicate of World. It contains gratuitous extras, namely, the laws of epiphenomenal emergence and the mental life secured by them. Indeed, it is explicit in his discussion that Otherworld is obtained from World by *addition*.

He makes the same mistake in his follow up argument given in terms of a pair of worlds more like our own (the example involving Jack Smart and cricket). He adds laws of epiphenomenal emergence but thinks that he is still dealing with a minimal physical duplicate of the world he *added* these laws to.

How did Tennant manage to misunderstand the notion of a minimal physical duplicate? As far as I can tell the answer is contained in the first paragraph of his section 2. Here he seems to think that I could not have meant what I did mean by 'minimal physical duplicate' on the ground that this reading would mean that

> ... a world that was physically just like ours, but in which all creatures were zombies, would (according to (III)) be a duplicate *simpliciter* of our world. Hence all the creatures in our world would be zombies. But materialism, properly conceived, cannot be committed to that.

But what (III) commits a materialist to holding is that *any* minimal physical duplicate of our world is a duplicate *simpliciter*, and so that *if* there is a possible world which is a minimal physical duplicate of our world except that all the creatures in it are zombies, then that world is a duplicate *simpliciter* of our world. And sensible materialists – ones who do not think that we are all zombies (all materialists as far as I know, for even eliminativists hold that we are conscious in enough of a sense to count as not being zombies) – hold that the antecedent of this conditional is false; that is, they deny that there is a possible world which is a minimal physical duplicate of our world except that all the creatures in it are zombies. They think that being conscious is a *physical* feature of many creatures in our world and so that any physical duplicate of our world contains consciousness and hence creatures that are not zombies. Of course many *non-materialists* have held that there is a possible world, W, which is a minimal physical duplicate of our world and yet which contains only zombies. They argue: our world has consciousness; W lacks it; our world and W agree in all physical respects and so what our world has that W lacks is non-physical; ergo, our world has some non-physical nature and materialism is false. The materialist reply to this argument is, and must be, to deny that there is such a W.

2 Does (III) Entail That There is a Statement About the Physical Way Things Are That Entails the Psychological Way Things Are?

Tennant does not criticise my argument for the answer yes to this question. He criticises an argument that he wrongly says is mine. My argument is of course available in my original paper but it helps to bring out the crucial difference between the argument I offered and the one he thinks I offered if I start by giving mine in the same style as Tennant sets out the argument he thinks I offered. My argument can be set out as follows:

(1) Let Φ be the statement true at our world and all and only the minimal physical duplicates of our world.

(2) Let Π be any true statement entirely about the psychological nature of our world.

(3) Let w be an arbitrary world at which Φ is true.

(4) Then w is a minimal physical duplicate of our world.

(5) By (III), w is a duplicate *simpliciter* of our world.

(6) Hence, w is a psychological duplicate of our world.

(7) Thus, Π is true at w.

(8) But w was arbitrary, and so any world making Φ true makes Π true; that is, Φ entails Π.

Tennant notes that we have no guarantee that there is a unique sentence that tells the true physical story about our world. But the point is unimportant. We can either develop the argument in terms of an arbitrary sentence ('Let Φ be an arbitrary sentence...'); or use the notion of a *statement*, where statements are individuated by their truth conditions, so that there can only be one statement true at any given set of worlds including, of course, the set consisting of our world and all and only the minimal physical duplicates of our world (this was in fact what I had in mind, and was why I did not use the term 'sentence'). Tennant prefers to use the term 'theory' instead of 'sentence' or 'statement'. I am not sure why. What he says suggests that he does not want to use 'sentence' when we are dealing with something that may be infinite. In any case, one thing we are agreed upon is that the point is a minor one. I will continue to use 'statement' but everything I say in what follows could be expressed using his term 'theory'.

It is worth highlighting the role of the notion of a statement being entirely about the psychological nature of our world. This means, as I explained in the original paper, that if the statement is false at a world, then that world must differ in psychological nature from our world – the statement is not about anything but our world's psychology, so there is no other way for it to be false. It is this notion that secures the step from (6) to (7).

The difference between my argument and Tennant's is that in place of my (1) Tennant has

(1*) Let Φ be the statement which tells the true physical story at our world.

This difference is crucial. Tennant faults the step in the argument from

(3) Let w be an arbitrary world at which at which Φ is true.

to

(4) Then w is a minimal physical duplicate of our world.

He would be right to do so had my argument used (1*) instead of (1). In Tennant's version of the argument Φ is "the statement [or theory] which tells the true physical story at our world" (quoting from (1*)), and there are worlds at which the true physical story about our world is true but which have additional non-physical stuff (extra "angels", say) and so are not minimal physical duplicates of our world. However, my argument actually used (1), and in it Φ is "the statement true at our world and all and only the minimal physical duplicates of our world" (quoting from (1)), and on this account of Φ the step is clearly valid.

3 Does the Physical *a priori* Entail the Psychological?

Suppose that you are a materialist and (unlike Tennant) accept that I have shown that you are committed to a certain physical story about the world entailing each and every psychological detail about our world. You might very well point out that my argument uses the necessary truth preserving account of entailment, and so, in view of the now widely accepted existence of necessary *a posteriori* truths, that it shows nothing as such about the *a priori* deducibility of the psychological from the physical. The aim of the last part of my paper was to show (sketchily – the issue is complex and space and time were limited) that, nevertheless, you are committed to an *a priori* deducibility claim – or, more precisely, you are if you accept what I (unoriginally) regard as the most appealing account of the necessary *a posteriori*, that provided by two dimensional modal logic. I will not repeat the argument here, but I should point out that there is a serious error in Tennant's account of my argument.

I sought to convey my basic point in terms of the best known example of a necessary *a posteriori* truth, the identity of water with H_2O. I argued that in spite of the fact that this is *a posteriori*, the H_2O way things are *a priori* entails the water way they are. In particular I discussed the argument

(1) Over 60% of the Earth is covered by H_2O

(2) H_2O fills the water role on Earth

(3) Therefore, over 60% of the Earth is covered by water.

Tennant points out, correctly, that I say a) that the step from (1) to (3) is necessarily truth preserving and so an entailment on *that* account of entailment, but the step is not *a priori*, and b) that the step, from (1) and (2) combined, to (3) is both necessarily truth preserving and *a priori*, and so the conjunction of (1) and (2) *a priori* entails (3). The error comes when he says (on his own behalf and mine) that (2) is necessary *a posteriori*, and that I hold that the fact that if (2) is true, it is necessarily true is revealed by conceptual analysis.

(2) is not necessary *a posteriori* because it is not necessary. Something other than H_2O might have filled the water role on Earth. It is not a necessary truth that the clear, potable, etc. liquid on Earth is H_2O anymore than it is a necessary truth that Einstein fills the role of being the most famous scientist of the twentieth century on Earth.

The reason I hold that the step from (1) and (2) combined to (3) is *a priori* as well as necessarily truth preserving is quite different from the reason Tennant attributes to me. It turns on the point that

(4) Water fills the water role on Earth

is, in my view, *a priori* true (provided that the water role is spelt out in the right way so as to include causal connections with certain uses of the word 'water' and all the rest of it) though contingent. But this means that we can see *a priori* that if (1) and (2) are both true, then so is (3). For from (1) and (2) we can infer *a priori* that what fills the water role covers over 60% of the Earth; but then the *a priori* nature of (4) enables us to make the final step to (3) also *a priori*. But we already know that the step from (1), and *a fortiori* from (1) and (2) combined, to (3) is necessarily truth preserving, so we have the desired result that the step from (1) and (2) combined to (3) is necessarily truth preserving and *a priori*. But (1) and (2) are both about the H_20 way things are. Hence, we have shown that the H_2O way things are *a priori* entails thewater way things are; or at least we have for a single case, and as the argument does not depend on the particular details of the case, it is plausible that the point generalises.[7]

[7]I am indebted to comments from David Braddon-Mitchell and Philip Pettit.

Remarks on Machines and Rule-Following

John Haugeland

Whether machines can be said to follow rules depends on the senses of 'machine', 'follow', and 'rule'. By '*machine*', I shall mean the sort of system that is studied by artificial intelligence these days. That's generic with regard to good old fashioned artificial intelligence – what I call *GOFAI* – or connectionism, dynamical models, genetic algorithms, or whatever is presently trendy. On the other hand, I don't mean to include within the purview of 'machine', people–though in a broader sense, of course, one could say we are machines, material objects of some special sort.

I take it for granted that people can follow rules; and the problem is what that amounts to. By contrast, it is an open question as to whether machines, in this narrower sense I have described, can follow rules. This leaves untouched a further ticklish question as to whether animals – cows, dogs, monkeys, and so on – are machines or can follow rules. I happen to think that animals are more like AI systems than they are like people, when it comes to interesting questions about rule-following. But I know a lot of people feel differently about animals; and I'm uncomfortable talking to those people about it, because some of them get emotional. So I just mention this by way of coming clean with my own prejudices, and other than that I don't talk about animals at all.

Back now to 'rule' and 'follow' which will be the rest of the talk. There are lots of senses of 'rule' and 'follow', and what we've got to do is sort them out. In at least two senses, two clear, simple senses, it's obvious that machines *can* follow rules. First, there are rules in the sense of describable regularities. Follow a rule like that is just exhibiting the regularity, that is, behaving in the way that the description of the regularity describes, being regular. Planets follow such rules, chemicals follow such rules. Laws of nature are regularities like this, though they aren't the only ones. That machines follow rules in this sense is not very surprising.

The second sense is more interesting. A rule can be a prespecification, in some appropriate formulation, some formalism, of what is to be done: a list of instructions, a recipe, or, of course, a program. And a machine can follow a rule in that sense too. Following such a rule is responding to a token of the specification. That is, you have a physical token of it, a copy of the program, and following the rule is doing what the specification specifies. This 'doing what is specified' carries counterfactual force. In other words, were the token to have been different, were it to have been a token of a

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 127-138

specification of some other behavior, then the system would have done that instead.

That's what I mean by responding to a token; and surely a machine can respond to rule-tokens in that sense. Programmable machines are the essential technology of our era. The Babylonians had agriculture, we've got programs. This development is so important, so unprecedented and striking, that a lot of people have supposed – have mistaken it – to be the essence of rule following at all. It does have the important advantage of being clearly compatible with naturalism and materialism, something that all good people want now.

Of course, GOFAI holds that we ourselves *are* programmable or programmed machines in that sense; so that's how far some reckless folk will go in taking this to be the essential notion of rule-following. But I think that's moving too fast. These are not the only senses of rule and following, and I'm going to talk about several more. But, by way of preparation, let me observe that neither of the two senses just characterized can be what Wittgenstein was worried about when he talked about rule-following. So, to acknowledge that machines can follow rules in those senses is not at all to acknowledge that they can do whatever it was that bothered Wittgenstein. That question remains open; and there are some other senses, yet, to be distinguished.

Now, it's not my main purpose – I wouldn't dare – to interpret Wittgenstein or enter any of the many disputes about him. But I want to say a little bit, by way of orientation, because it will, so to speak, locate what I want to say in other discussions that interest many philosophers. The little that I will say is loosely based on the writings of John McDowell – particularly his rule following paper of 1984. As he organizes the issues in that paper (actually I've transformed it somewhat, and McDowell might not like the way I'm going to put it) we can think of Wittgenstein's discussion in the *Investigations* as having the structure of a dilemma within a dilemma – that is, nested dilemmas.

Each dilemma has a presupposition. So, in principle, a way to avoid the dilemma is to deny the presupposition. The outer dilemma presupposes that the only way to understand a rule with normative force is to interpret it. By *understanding* a normative rule, I mean knowing how to follow it correctly – the ability to tell the difference between correct and incorrect performance, and then do the former. So, the outer presupposition is that the only way to be able to follow a normative rule correctly it is by means of an interpretation. That is, the difference between following it correctly and incorrectly is always relative to how it is interpreted. Why does this create a dilemma?

Well, one option is to infer that all normative rules need to be interpreted. (That's only one option; we'll get to the other in a minute.) First, we ask: What is an interpretation of a rule? Asking this question presents us with the inner dilemma. What are its alternatives? One is to say that an interpretation of a rule is a specification of how to follow the rule correctly, a specification of what to do. But, in that case, the interpretation itself is, in effect, another rule. Hence, it too would have to be interpreted to be understood – that is, interpreted in order to be followed correctly – and that launches the famous regress. Some readers construe Wittgenstein as arguing that there is no way out of this regress, creating a kind of paradox. But I think that's just to concede defeat on one horn of a dilemma, and that Wittgenstein's purpose is rather to show that the dilemma itself is ill-founded.

The second alternative (for the inner dilemma) is to allow that at least some interpretations are not, themselves, rules. Either an interpretation itself is a rule and requires an interpretation, or else some interpretations are *not* rules. In the latter case, the interpretations need not themselves be interpreted to be understood. These interpretations, the ones that aren't rules, must, so to speak, wear their intelligibility on their sleeves – "transparently". That's why they don't need to be interpreted, and, hence, can serve as regress stoppers. What this is supposed to mean is that there is no possibility of misunderstanding such interpretations. They have some "queer" or "super-rigid" direct connection with their application: a super-rigid machine (or "rails"). This, I take it, is what Wittgenstein mocks as a mythology: the mythology of *meaning* as the last interpretation.

So, the inner dilemma is a choice between these two options, regress (or paradox) and mythology, neither of which is tolerable. The presupposition of this dilemma is that every normative rule needs an interpretation. This presupposition is itself one horn of the outer dilemma. In other words, the whole inner dilemma is one horn of the outer dilemma – hence the inner/outer structure. The way to avoid the inner dilemma is to deny its presupposition – namely, that all normative rules need interpretations. But, to say that some normative rules don't need to be interpreted will just land us on the second horn of the outer dilemma. The presupposition of that dilemma is that the only way to understand a normative rule is by means of an interpretation. So, if we deny interpretation, we deny understanding. In other words, according to this option, some normative rules don't depend on any ability to tell the difference between following them correctly and incorrectly.

What could such rules be? Well, the obvious candidate is dispositions. Dispositions are simply *triggered* by their respective antecedent conditions, and once triggered they – well, to personify – they carry themselves through to completion automatically. No understanding, hence no interpretation, is necessary. There is no need to tell the difference between correct and incorrect performance for a disposition to execute. It just happens. Simple dispositions could be regarded as versions of the first sort of rule that I mentioned at the outset, that is, natural regularities. Then, the second sort, namely computer programs, could be thought of as ways of implementing complex dispositions.

The trouble with dispositions, however, is that the distinction between

right and wrong, correct and incorrect, collapses. Natural regularities, whether simple or complex, are not in themselves normative. Now this is not to deny that we, ourselves, can impose a kind of normativity on them, relative to our own purposes. Obviously, lots of dispositions are normative in that indirect, derived sense. For instance, when we strike a match, it's not only *disposed* to light, it's also *supposed* to light. But why? Well, because that's what we designed it to do – it's our match, we built it, and we intend it to light. So it's supposed to. But that normativity is imposed on the match from us, from the outside.

Now, I will add a claim that's possibly contentious: the undeniable normativity of (current) computer programs, the normativity in terms of which things like bugs and malfunctions are intelligible, as surely they are, is also of this imposed external sort. It's the normativity that comes from us because we designed the programs. We have purposes for which they are built and, relative to those purposes of ours, there are norms for how those programs are supposed to function. I don't mean to be making a general metaphysical claim to the effect that nothing that is an artifact, or that has the basic structure of a computer, could ever have norms of its own. Those possibilities seem to me to be open. But, I claim, no current artifact or computer (or anything very similar to them) has norms of its own. (The larger project, in a way, is to see why.)

But right now, our dilemma concerns people. Here, basic normativity cannot be understood as imposed from the outside by a designer, whose purposes determine what is correct and incorrect for us (not any more, anyway). Hence, to resort to mere dispositions, as a way of avoiding the presupposition of the inner dilemma (that all normative rules require interpretations) is simply to forfeit the distinction between correct and incorrect performance. Whatever happens is *ipso facto* "correct"; and, as Wittgenstein points out, if whatever happens counts as correct, then there is no point in speaking of correct and incorrect at all. The trouble is, if there's no difference between correct and incorrect, then we're not talking about *normative* rules. So, the outer dilemma is: either we accept the inner dilemma, or else we forfeit normativity – neither of which is tolerable.

The way to avoid the outer dilemma is to deny *its* presupposition: namely, that the only way to understand a normative rule – i.e. to know how to follow it correctly – is by means of an interpretation. So, Wittgenstein concludes:

> What this shows is that there is a way of grasping a rule which is *not* an interpretation, but which is exhibited in what we call "obeying the rule" and "going against it" in actual cases. (*Investigations* §201)

Unfortunately, although Wittgenstein makes this point quite clearly, he does not say much by way of explaining how it is possible. That is, how is it possible for there to be a difference between doing something correctly and doing it incorrectly that doesn't depend on an interpretation of some rule?

In the remainder of this paper, I will try to sketch an answer to this question, an answer in terms of which we can see why I place animals and (contemporary) machines on one side of a divide, and people on the other. My answer will actually be somewhat more specific than the question as it stands. That is, I will consider particular species of normativity, not the simplest case, but a case which is, I think, especially important. And that is the case in which what is correct or not depends centrally on how things stand with regard to some definite object or objects. Normative performances of this (loosely characterized) sort I will call *objective* performances. So my question is: how is objective normativity – objectivity – possible? I believe that this case was of great interest to Wittgenstein as well.

I will use *skill* as a general term for the capacity to follow rules in this sense, to perform correctly reliably. (This term is intended to evoke such Wittgenstinian terms as 'custom', 'technique', and 'practice' but without suggesting that my account is an interpretation of what he meant.) Linguistic skills are prominent among objective skills, but I will not emphasize them. Rather, I will concentrate on recognitive and manipulative skills, which, it seems to me, are often also objective, and are more basic than linguistic skills.

A skill is a special sort of acquired disposition – essentially different from other acquired dispositions, not to mention mere natural regularities, in that there is a difference between exercising it correctly and incorrectly. Skills are normative dispositions. Moreover, and crucially, this normative character belongs to the skill (or skillful agent) itself, in the sense that it is not merely imposed or attributed from the outside, relative to some external designer or interpreter, or somebody adopting a "stance". That's not where the normativity of the skill comes from. Skills, therefore, are (or might as well be) custom tailored to fill the gap between the horns of Wittgenstein's outer dilemma – by contravening its presupposition. My account of skills, however, is not going to sound much like Wittgenstein.

In order to develop the idea, I will need to introduce still more notions of 'rule' and 'following'. In particular, I want to draw upon a distinction proposed some years ago, in slightly different ways, by John Rawls and John Searle. Be warned, however, that I will not use this distinction in quite the same way that either of them does. The distinction is between those rules that merely guide or constrain some activity which is intelligible independently of the rules, and, on the other hand, those rules which establish or define a new kind of activity which could not be carried out apart from those rules.

Rawls calls his version of the former, the ones that merely guide or constrain, 'maxims' or 'rules of thumb'; whereas Searle calls his version 'regulative' rules, or 'regulations'. Now, maxims and regulations are not quite the same. This is where Rawls and Searle most differ. Rawls's maxims

are a kind of generalization garnered from experience, and they have the force of advice. That is, you've learned over time what is usually the best thing to do, and you formulate for yourself this rule of thumb, so you don't have to figure it out again each time – you just rely on the maxim. By contrast, regulations in Searle's sense are imperatives: they have the force of authority. Despite this difference, however, maxims and regulations are the same side of the main distinction, which Rawls and Searle both draw.

The other kind of rule, the other side of the distinction, Rawls calls the 'practice sense' of rules. Searle refers to them as 'constitutive' rules – and that's the term I will use. I think, on this side of the distinction, what Rawls and Searle say amount to pretty much the same thing.

Now both Rawls and Searle use games, such a chess, to explain the difference. Rawls also used baseball to explain the difference. (I think that the use of baseball as a philosophical example is a sign of great depth.) I too will begin with chess. But I think there are serious problems with using chess, or games more generally, as examples of constitution. Part of what I hope will emerge in what follows is some hint of those problems, and how different sorts of examples may not be subject to them.

But first: the important point about constitutive rules is that they define or establish the activity or the items to which they apply. For instance, the rules of chess define the game, including therefore all of its paraphernalia and all of its phenomena. Apart from the rules of chess, there is no such thing, for instance, as castling, no such thing as a knight fork, or capturing *en passant*. Indeed, there is no such thing as a knight, apart from the rules of chess – no such thing as a *chess* knight anyway, or a king or bishop. Chess pieces, positions, moves, combinations: none of that makes any sense at all outside of the context governed by the constitutive rules of chess. That's why those rules are *constitutive*.

By contrast, a maxim, in Rawls's sense – such as "Always lock the door when you leave the house" (good advice which one learns from sour experience) or "Never take candy from a stranger" – guides an activity that, as the activity itself, makes perfect sense apart from the maxim. Thus, what it is to lock the door when you leave the house is perfectly intelligible independently of this and related maxims; likewise for candy and strangers. The rules don't constitute what those things amount to. Similarly, for regulative rules, such as: "You must be in bed by 10 o'clock" or "Thou shalt not bear false witness". The activities in question make perfect sense apart from the rules.

Now, of course, maxims and regulations can also guide constituted activities. In fact I believe, in a deep sense, there is nothing else to guide. But the point is easy to see even in chess: "Don't bring your queen out too early" is a maxim in Rawls's sense; "Thou shalt move thy bishop only along unobstructed diagonals" is a chess regulation.

The latter, however, is the beginning of trouble. For instance, that regulation governing how players may move their bishops is very similar to

one of theconstitutive rules of chess, the one that defines what a bishop is: "A bishop is a piece that moves only along unobstructed diagonals." They're so close that it can be hard to tell them apart. (Searle, in fact, doesn't tell them apart: he says that constitutive rules are a special case of regulative rules, and would consider this to be an instance.) What is it to *follow* a constitutive rule? Indeed, who or what does follow those rules? The *players* follow (abide by) the regulations. Is it also the players that follow the constitutive rules? Well, if games like chess are the paradigm, then it is natural to think that following the constitutive rules just is playing by the rules, playing the game according to the rules. And then its almost unavoidable that the constitutive rules simply collapse into the regulative rules. The distinction disappears.

But there are two things wrong with this. First, it completely obscures what *constituting* means. What the constitutive rules do is establish what the relevant entities are, in terms of a way in which they can be understood; that's what it is to *constitute* a range of entities (as both Rawls and Searle appreciated). So for instance, the constitutive rules of chess establish what it is to be a bishop, by saying what bishops can and cannot do – how they move. But, it is only an accident of the fact that we are talking about a game that any move a *bishop* makes is, a the same time, a move that a *player* makes with his or her bishop. And it is this accident that underlies the (deceptive) appearance that constitutive rules are just regulative rules, slightly reformulated. It is an essential part of my position that there are (must be) constitutive rules for all intelligible domains – including domains of entities whose behavior extends well beyond manipulation by us. For these domains, there simply aren't regulative rules shadowing the constitutive rules; and that's why it's important not to confuse the two types of rule, even in the case of games.

Second, there is another kind of "rule" governing our own performance, that always does accompany constitutive rules, and so should not be confused with regulative rules of the above sort. These "rules" are the "order" or "regularity" that the pertinent *skills* exhibit. (I use all these scare quotes, because I don't want the terminology to imply that these "rules" are, or even can be, explicitly formulated; I don't think Wittgenstein's notion of a custom or practice implies formulability either.) For instance, in order to play chess, a player must be able to recognize a bishop when she or he sees one, be able to tell what square it's on, be able to move it (along an unobstructed diagonal), and so on. Such mundane, practical skills are essential to the possibility of chess as a constituted domain, but they are not at all the same as the constitutive rules (let alone the regulative rules printed on the box).

What's really fundamental here is the relation between constitutive rules and mundane, practical skills. Regulative rules enter the picture only as an artifact of the choice of games as examples; they are a distraction, and should basically be ignored.

We can work toward the genuine function of constitutive rules, while remaining within the familiar examples, by a simple alteration in point of view. Return to chess, but consider how you stand, your posture, vis-à-vis your *opponent*'s moves and pieces. Now, your opponent's bishop is still constituted by the rules for bishop moves, but *you* don't move it. The idea, here, is to pry apart the rules regulating our actions from those constituting the phenomena. So we're going to pretend that your "opponent" isn't another player; rather, those moves just happen on their own, as if they were natural phenomena. So, you interact with that opposing bishop only by observing it, and by making moves of your own that affect it, or are affected by it. It's quite clear, therefore, that your skillful behavior with regard to that (opposing) bishop is quite distinct from the behaviour of the bishop itself, by virtue of which it is constituted as a bishop.

But this does not mean that the rules constituting that bishop are a matter of indifference to you. On the contrary, your ability to recognize and interact with it *as* a bishop depends on its being a bishop; that is, you depend on it behaving according to the *constitutive* rules. You are not indifferent to that – it matters to you – that it behave properly. This means that, if what you took to be an opposing bishop were to seem to you to move in a way that bishops can't move (if it looked like an illegal move), you would have to see that as a problem. This is a problem for *you*: it threatens your ability to understand that piece as a bishop, hence your ability to understand what's going on at all. Your very ability to "play the game" is at risk and at stake.

This brings out an important point: it is essential to engaging in chess play that one *expect* opposing moves to be legal, in the strong sense of 'expect'. (I mean the sense that stern teachers use when they say: "I *expect* your work to be done on time.") This strong sense of 'expect' means, first, that you are counting on it, and second, that you are insisting on it – you demand it. You count on the opposing moves being legal, and you insist on it – on pain of giving up the game. Now, this strong expectation is not the same as following the constitutive rules; the pieces themselves follow those rules, not you. But it is, in fact, the way that you relate to the constitutive rules: you don't yourself follow them, but you *expect* the phenomena to follow them – in the strong sense of counting on it and insisting on it.

Our own skillful interactions with the constituted phenomena – recognizing and manipulating them, for instance – are not covered by this expectation, but they are dependent on it. How that dependence works is, I claim, the key to objectivity and correctness. When one learns to play chess, one certainly learns to recognize various pieces, squares, moves, and so on. Why is it that what one thus learns are recognitive (and manipulative) *skills*, rather than mere responsive dispositions? That is, why do they have a *normative* dimension? Recognition, of course, is essentially normative, since it is possible to *mis*recognize something – e.g., to mistake a bishop for something else, or mistake something else for a bishop. Inasmuch as there is a difference between correct and incorrect performance, the ability

is normative, a *skill*. Indeed, since correctness depends directly on whether what is "recognized" is a bishop or not, the skill is *objective*.

But, so far, that is just to announce that these skills are normative; no account has been offered of how this is possible. By virtue of what are the chess-players's recognitive skills governed by *norms*? Why would there be anything *wrong* if we "mistook" bishops for something else, or something else for bishops? And the answer is: *only* if we identify the pieces correctly can they fulfill our expectations – namely, our expectations that they themselves follow the constitutive rules. This is equivalent to saying that, only if we identify the phenomena correctly, can we actually play the game. So, the source of the "normative force" binding our mundane recognitive and manipulative skillful performances is, at one remove, our own expectation, our own *insistence* that the phenomena to be recognized accord with the corresponding constitutive rules.

What emerges, therefore, is a two-tiered structure. One tier – the "upper" tier, if you will – is the constitutive rules, in accord with which the phenomena are constituted (made sense of) as what they are. These rules determine legality in the game; or, to put it more generally, they determine what can and cannot happen. The other, or "lower", tier comprises our mundane practical skills for recognizing and otherwise dealing with those phenomena. Each tier obviously needs and depends on the other. If there were no constitutive rules, there would be no constituted, intelligible phenomena to recognize or deal with; and, if we were unable to recognize or deal with them in anyway, we could hardly make sense of them as according with the rules, let alone insist that they do.

The engine behind this structure, what makes it go, is signaled by the phrase: 'on pain of giving up the game'. Being *able* to play – to keep playing, and not have to give it up – is a kind of success condition. It presupposes both tiers: the "legality" of the phenomena themselves, and our reliable objective skills for dealing with them. More than that, it binds the tiers together. Undertaking to play the game involves undertaking a *commitment* to see that the phenomena in fact accord with the rules. This commitment is, by its nature, twofold. On the one hand, it includes the strong expectation that the phenomena do so accord; and on the other hand, it entails a responsibility to get the phenomena right – that is, to recognize and manipulate them correctly. Neither side makes sense without the other. Thus, the commitment embraces both tiers within itself.

Now, it must be admitted that these points strain the chess example almost beyond endurance – even if we confine our attention to the "opposing" moves. After all, chess is a simple human invention – merely a game – so the ability to keep playing is not all that impressive, nor is the fact that the phenomena are constituted by rules. There would be no chess phenomena at all if we hadn't invented them; and we can change them at will, by changing the rules. And, in the meantime, the device of considering the opposing moves without the opposing player – as if they happened by themselves –

is hard to take completely seriously. But that leaves the crucial distinction between constitutive rules on shaky ground, and puts the whole account at risk. This is why chess, and games in general, are a treacherous example.

If we're lucky, however – if I've succeeded in my expository strategy – we now have enough pieces in place that we can effect a kind of gestalt switch, keeping the basic structure intact, while leaving the examples behind. The switch will be from games to *science*. A slogan that expresses the leap might be: Science is playing chess with nature. How can we flesh this out? First the differences: science is not just a game; natural phenomena were not invented by us; and we cannot change them at will simply by changing the rules. On the other hand, two points in the account that were rather awkward in regard to games are not at all awkward when we turn instead to science. First, for scientists, the ability to keep "playing" – to perform mundane skills correctly while seeing that the phenomena accord with the rules – is anything but trivial and taken for granted. And, second, the idea of natural "moves" happening all by themselves is not at all hard to take seriously. In effect, nature makes all the moves in this game: scientists just recognize and cope with them.

The important claim, however, is that the overall structure is the same. The phenomena investigated by any science are constituted – made sense of – according to appropriate standards of scientific intelligibility. Paradigmatically, modern physical phenomena are constituted according to the standard of strict law-governedness. 'Strict' here means at least three things: the laws are mathematically precise (usually quantitative); collectively, they form a comprehensive system (they, plus physical boundary conditions, suffice to explain all physical phenomena that are explicable at all); and they form a closed system (nothing from outside the physical can interfere or contravene them). This, or something like it, is the standard of intelligibility of all modern physical science; it's what differentiates the modern era from the middle ages and the renaissance. (Readers of Davidson will recognize certain details in the formulation; but the general idea is widespread.)

What's more, scientists *expect*, in the same strong sense as before, that natural phenomena will accord with the relevant standards of scientific intelligibility – such as strict law-governedness. They count on it and insist on it, in the sense that, if something is found to be out of line, they are bound to regard that as a problem – a situation not to be tolerated. No scientist, *qua* scientist, can blithely accept a case that looks as if the laws of nature are being broken; nor can he or she simply turn away and ignore it. In such a case, something is clearly wrong; and it is incumbent on scientists to figure out what that is, and rectify it. This commitment is characteristic of science as such.

One way, by far the most common, to find out what went wrong is to find a mistake in the observations or procedures, such that it only seemed as if a law had been broken. *Mistake*, of course, is a normative concept. How comes it that scientific observations and procedures are subject to normative

evaluation? Just as for the mundane skills of observing and playing games, the binding "force" behind the norms derives ultimately from the alternative of having to "give up the game". In order to be able to continue, scientists must undertake to perform their scientific skills *correctly*. If there is no reliable distinction between correct and incorrect procedures, such that mistakes can reliably be detected and rooted out, there is no content to the insistence on constitutive standards. Therefore, scientific commitment is also essentially twofold, embracing both tiers within itself.

Unlike game players, however, scientists sometimes confront the question of what exactly the standards demand – what, after all, can and cannot happen. Sometimes, for instance, they will fail to uncover any mistake in their observations or procedures, yet still something is wrong. Then the very way in which the phenomena as such are constituted may fall into question – what the standards are or how they apply. Maybe it seems as if a law of nature is being broken, not because we're making a mundane mistake, but because we've misunderstood *how* nature is constituted as law-governed. Perhaps the laws aren't what we thought they were, either in some detail or altogether.

The prospect of "giving up the game" is much more serious and tangible in science, because nature just might not "play by the rules". And, in that case, there's nothing to do but revise (or abandon) the game. This is the element that makes science deeply empirical, not the fact that scientists perform experiments and make observations. (Game players experiment and observe.) Science, unlike ordinary game-playing, has, as part of its essential business, to figure out just what the "game" itself is. But the basic structure – a two-tiered commitment, by virtue of which, simultaneously, the phenomena are constituted as what they are, and our mundane skills for dealing with them are rendered normative – that structure remains the same.

That completes the promised sketch of an account of how normative and objective skills are possible. They are an integral factor in a larger account of how objective phenomena themselves are constituted; and their role in this account both requires and affords their objective normativity. That larger account, in turn, is itself grounded in the possibility of individuals undertaking a twofold commitment to seeing that the phenomena in fact accord with the constitutive rules or standards, on pain of giving up the game.

Now we can return to the original question: Can a machine follow a rule? If the sense of rule-following is what I have called a skill (and which I take to be at least related to what Wittgenstein calls a 'custom' or a 'practice', and says can be grasped without an interpretation), and if my account of the possibility of skills is on the right track, then this question reduces to another: Can a machine undertake the sort of commitment just described, on pain of giving up the game? Well, I don't want to make projections about hypothetical future machines, vaguely imagined. But as to contem-

porary systems, already built or clearly envisioned, in all the varieties of AI, I don't think any of them is capable of anything like such a commitment. Nothing is ever at stake or at risk for any of them. Nobody in AI is even thinking about machine commitment, or machines having a stake in things – in part, no doubt, because it's far from clear how to think about them. But that means there is an important sense in which a machine – so far, at least – *cannot* follow a rule. Indeed, in my opinion, they cannot even (really) play chess.

References

McDowell, J. 1984 "Wittgenstein on Following a Rule", *Synthese* 58, 325 – 363.

On the Relation of the Mental and the Physical

Johann Christian Marek

I

A very general formulation of the mind/body problem can be expressed by the question – what kind of relation is there between the mental and the physical? In order to find a more determinate and systematic formulation of this question, it is necessary to draw some ontological distinctions. Against the background of a developed ontology you can find a more systematic and precise approach, and, moreover, obtain tools for criticism of some suggested answers to the mind/body problem.

The ontology I have chosen is more or less Roderick M. Chisholm's theory of categories;[1] the position I would like to criticize by this ontology is the so called contingent identity theory.

II

For a brief sketch of Chisholm's ontology, I start with examples of sentences about occurrences like

> Anne feels sad
> Vesuvius smokes.

With these sentences we intend to refer to mental and physical occurrences (states, events)[2] respectively, namely to Anne's exemplifying the property of feeling sad and to Vesuvius' exemplifying the property of smoking.

Occurrences like these are concrete but dependent entities; they contain two further entities, the substrate and the content of the occurrence. The substrate is the bearer of the property in question, i.e., the entity which exemplifies the property. In our examples Anne and Vesuvius are the substrates. The content of the occurrence is the property being exemplified: feeling sad and smoking respectively. The dependency of occurrences on their substrates and contents can be expressed by the following:

> For every x and F, if there is an occurrence x-being-F, then it is impossible for this occurrence, x-being-F, that x would not be its substrate and being-F would not be its content.

[1] See Chisholm 1989 and 1992.

[2] At this stage of the discussion, the difference between states and events is not important. (By the way, Chisholm uses "states" in a general sense – as I use "occurrences" – and events are, therefore, special kinds of states.)

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 139-145

A person who is sad, for example, is not necessarily such that she is sad, but the occurrence, Anne-being-sad, is necessarily a state of Anne.[3] Therefore, I cannot share Anne's state of sadness. What we can have in common is the property exemplified by us, being sad.

Concerning the identity of occurrences, it is important that if an occurrence o_1 is identical with an occurrence o_2 then o_1 and o_2 have the same substrate and the same content. This condition suggests a rather fine grained theory of states and events. If their contents, the exemplified properties, are different, the states or events in question are different too.

Chisholm understands properties both intensionally and intentionally.[4] Properties are *intensional* in that they are not extensional, because properties that have the same instances may yet be different properties (for example: being an animal with kidneys vs. being an animal with a heart). And properties are *intentional* because they are defined intentionally as something which can be attributed to something. The criterion of identity for properties can be stated as following:

> A property P is *identical* with a property Q, if and only if P and Q are necessarily such that whoever attributes P to something also attributes Q to it, and whoever attributes Q to something also attributes P to it.[5]

Logical relations can be defined between properties, for example:[6]

> P *implies* $Q =_{\text{df}} P$ and Q are necessarily such that if anything has P, then something has Q.
> P *includes* $Q =_{\text{df}} P$ and Q are necessarily such that whatever has P has Q.

The property of reading a letter implies that of being a letter and that of reading, and the property of reading is also included in it. Similarly, the property of being blue includes (and also implies) that of being coloured. If P includes Q, then P also implies Q; but P may imply Q without including Q. Both cases, that of a property P not implying Q, and that of P not including Q, have to be distinguished from the two following kinds of exclusion:

> P *excludes* $Q =_{\text{df}} P$ and Q are necessarily such that whatever has P does not have Q.
> P *excludes strongly* $Q =_{\text{df}} P$ and Q are necessarily such that if anything has P then nothing has Q.

Examples of the exclusion in the weaker sense are easily found: *being a bachelor* versus *being female*. But it seems to be more difficult to find

[3] Chisholm 1989: 150, 164, and Chisholm 1992: 5.
[4] Chisholm 1991: 45.
[5] In 1991: 45f. Chisholm uses "x's believing has being-F as its content" instead of "x attributes F".
[6] Concerning property implication and inclusion, see Chisholm 1989: 101, 143, 153.

examples of the strong exclusion: *Thinking of everything* excludes strongly *not being thought by anything*, because if somebody has an idea of each thing, then there is nothing such that nobody has an idea of it.

III

What is said about properties in general can be applied to the mental and physical respectively. I take "having mental properties" ("having physical properties") as fundamental. This primary usage of "mental" ("physical") refers to a special kind of properties. In a derivative, secondary sense, you can also use it for individuals, and for states and events. For example, a *mental individual* is an individual which has mental properties, and a *physical individual* is an individual which has physical properties.[7] Concerning occurrences, you can say that an occurrence with mental properties as content is a mental one, that is, someone's having mental properties is something mental too, it is a mental state (event). Examples of mental properties are: feeling sad, desiring, judging; and examples of non psychological, physical properties are: moving, being extended, weighing 120 pounds, having a brain. "Mental" is also used in a broader sense which goes beyond the manifest experience. This broader usage concerns, for example, more or less complex dispositions such as being anxious, being intelligent, and also compositions of mental and physical properties such as perceiving the outside world. It even includes properties concerning the unconscious. I confine myself to the mental in the narrow sense, as I think that the occurrent, experiential mental properties can be used to explain the mental properties in the broader sense (and not the other way around).

The discussion about the mark of the mental can be seen as an attempt to find a criterion for mental properties which exhibits characteristics of mental properties and only of mental properties. Examples of those characteristics are intentionality, subjectivity, self-presentation, extensionlessness, and indivisibility.

Brentano's thesis of intentionality could be formulated by the following: Mental and only mental properties are intentional, and a property P is intentional just in case P includes the property of being intentionally directed to something.

A property P is subjective, on the other hand, just in case having P includes knowing what it is like to have P.[8]

Meinong's self-presentation can be interpreted as follows: A property P is self-presenting if and only if having P includes believing of oneself that one is having P, or in other words: P is impossibly such that x has P and x does not believe of himself that he is having P.

[7] According to Descartes, a mental entity is defined more strongly as something which has necessarily mental properties; something would not exist as a mental entity if it were not thinking.
[8] See Nagel 1974, 1980: 160.

With respect to the property of not being spatially extended (and similarly to indivisibility), there is at least a strong and a weak version. According to the stronger, the Cartesian version, having a mental property excludes the property of being spatially extended; and the weaker version claims only that having a mental property does not include the property of being spatially extended. In case you agree that physical properties include the property of being spatially extended, the weaker version is still a criterion demarcating mental properties from physical properties.

IV

At the beginning I said that an ontology can deliver tools for a systematic and precise overview of the mind/body problem. The question – what does the relation of the mental and the physical consist in? – becomes clearer and more determinate by asking what does the relation between mental and physical properties consists in, or more specifically:

Does *having a mental property* exclude *having a physical property*?[9]

This question presupposes that mental and physical properties can be meaningfully attributed to something.

Another, completely different strategy to "answer" the question would be to reject this presupposition, and therefore the whole question. You declare the question a pseudo-problem because you think the predicate "mental" ("physical") does not designate a specific property, i.e., because you think the predicate is senseless. In a similar way, you can reject the presupposition by finding good reasons for asserting that the properties in question lead to a contradiction.

Descartes certainly gave a positive answer. He would say that a mixed mental-physical individual, that is, an entity which has both mental and physical properties, cannot exist. Only purely mental and purely physical entities are possible. If you speak of a unit of a mental individual with a physical thing, you do not speak of a mixed (= psycho-physical) entity, but only of a mental individual which is interrelated (by causal relations for example) with a physical individual.

The negative answer, saying that mental and physical properties do not exclude each other, or alternatively, that mixed mental-physical individuals are possible, is advocated by several positions which differ from each other in other ways depending on their additional assertions.

A very strong position giving a negative answer is held by Logical Behaviourism. This strong kind of physicalism not only claims that mental properties do not exclude physical properties but also that the mental properties include the physical ones (whereas not all physical properties include mental properties). The reason for this is that mental properties *are* physical properties, that is, dispositions to behave, because mental predicates can

[9]For a more detailed survey of questions and answers concerning the mind-body problem see Marek 1989.

be defined by dispositional predicates.[10]According to Logical Behaviourism the further questions of whether the mental individual is identical with the physical individual in question, and of whether the mental state is identical with the physical one can be answered affirmatively.

Logical Behaviourism is not the only point of view from which you can answer negatively the question of whether *having a mental property* excludes *having a physical property*. This negative answer is a consequence of many other positions. Some theories of supervenience of the mental upon the physical, viz. of mental properties upon physical properties, imply the claim that *having a mental property* includes *having a physical property*.[11]

There are even weaker positions which also agree that *having a mental property* does not exclude *having a physical property*, but they do not go so far as to say that a mental property includes a physical property. In other words, psycho-physical individuals are possible, but purely mental as well as purely physical individuals are possible too. That is, a mental individual can also be a physical individual, just as, for example, something which is soft can be something round too. It is just a natural fact that mental individuals have physical properties. And the real mind-body problem emerges by asking, what kind of relation exists between mental and physical occurrences. Is it a special kind of correlation, a causal one, as a noncartesian dualist would say, or – as identity theorists suggest – is it an identity of the mental and the physical occurrence?

V

There has been a great deal of internal and external criticism of Logical Behaviourism. From an external point of view, you can criticize Logical Behaviourism by giving arguments for the claim that mental properties have characteristics physical ones do not have. If you accept the self-presenting, or intentional, or subjective character of the mental, or something else like that as a delimiting mark of the mental from the physical, you have already abandoned the logical behaviourist position.

An internal kind of criticism is given by arguing that a suitable definition of a mental property by dispositions to behave would always need further qualifications which employ again intentional, psychological concepts.[12]

A motivation to find another materialistic alternative to Behaviourism is mentioned by U. T. Place who defends the contingent identity theory. He thought that many psychological terms, especially cognitive ones, could be defined dispositionally, but that for a lot of other terms there were doubts about this solution. For the notions of consciousness, experience, sensation, and mental imagery, for example, something would be left out by the

[10]See Carnap 1932/33: 131. He suggests only a sketch of a definition for 'being excited': "Die Person X ist aufgeregt" ("Person X is excited") can be defined by "Wenn jetzt Reize von der und der Art ausgeübt werden, so reagiert X darauf in der und der Weise" ("If X is exposed to stimuli of such and such kind, then X reacts in such and such manner").

[11]See Kim's concept of strong supervenience in Kim 1984/5: 163-171.

[12]See Chisholm 1955/56 and 1989.

dispositional account. Those mental occurrences could not be adequately understood as dispositional states – something, an inner process, is going on when, for example, you are feeling pain.[13]

Moreover, although a mental property is not a physical one, there are other identities of the mental and the physical – as contingent identity theorists would claim: The mental individual is a physical individual too, and the mental event (being in pain) is also identical with the corresponding physical event (having stimulated C-fibres). But the reason for those identities is not based on the identity of the mental and the physical property.

Let us be more precise. Actually we can at least distinguish two versions of the identity theory, a stronger and a weaker one. According to the stronger version, a mental event is a physical event, because mental event types are identical with physical event types. Whereas the weaker variant only affirms the identity of the mental event token with the physical one – without referring to type identity of events.

Both kinds of contingent identity theories claim – and therefore they are called "contingent" – that the identity is not necessary; it is a contingent, empirical one. In order to illustrate their intuitions, the identity theorists rely on analogies. On the level of physical things, for example, there is a contingent identity between the morning star and the evening star, and, on the level of physical occurrences, there is a contingent identity between having a specific temperature and having a specific amount of mean kinetic energy. The sciences discover those identities by empirical investigation and not by conceptual analysis. And in a similar way we can discover the identity of mental and physical occurrences.

VI

Like Kripke's semantic-ontological objection to the contingent identity theory, you can present an *a priori* rejection of this theory when you accept a Chisholmian ontology such as I presented in sections II and III above. According to such an ontology, a mental property, M, is a necessary constituent of a mental occurrence, o_M. The mental occurrence o_M would not exist if the property in question, M, were replaced by a property which is not identical with M. Thus o_M would not exist if its determining property were not mental.

Mental occurrences could only be physical occurrences, if the underlying mental properties involved were physical properties, and that would lead us back to Logical Behaviourism.

Concerning the type-type identity theory, a further problem arises. How can types of occurrences (states, events, processes) be interpreted, if they are not the underlying properties? The ontology I presented does not require types of occurrences in addition to the properties as a further ontological category. Either the occurrence types are the underlying properties,

[13] Place 1969: 22.

that is, the content of the occurrences, and that would be Logical Behaviourism again, or they are of a different ontological kind. But what kind of entities are they? According to this ontology, occurrence types are worse than superfluous, there is no room in it for them.[14]

References

Carnap, R. 1932/33 "Psychologie in physikalischer Sprache", *Erkenntnis* 3, 107–142.

Chisholm, R.M. 1955/56 "Sentences about Believing", *Proceedings of the Aristotelian Society* LVI, 125–148.

Chisholm, R.M. 1985 "Preface" to *Philosophy of Mind – Philosophy of Psychology* Proceedings of the 9th International Wittgenstein Symposium, Wien: Hölder-Pichler-Tempsky.

Chisholm, R.M. 1989 *On Metaphysics*, Minneapolis: Univ. of Minnesota Press.

Chisholm, R.M. 1992 "The Basic Ontological Categories", in K. Mulligan (ed.), *Language, Truth, and Ontology*, Dordrecht: Kluwer, pp. 1–13.

Chisholm, R.M. 1991 "An Intentional Explication of Universals", in *Conceptus* XXV no. 66, 45–48.

Kim, J. 1979 "Causality, Identity, and Supervenience in the Mind-Body-Problem, in P.A. French *et al.* (eds.), *Midwest Studies in Philosophy* Vol. IV: *Studies in Metaphysics*, Minneapolis: University of Minnesota Press, pp. 31–49.

Kim, J. 1984/85 "Concepts of Supervenience", *Philosophy and Phenomenological research* 45, 153–176.

Marek. J.C. 1989 "Ein Zugang zum psychophysischen Problem", in W.L. Gombocz, H. Rutte and W. Sauer (eds.), *Traditionen und Perspektiven der analytischen Philosophie*, Wien: Hölder-Pichler-Tempsky, pp. 297-321.

Nagel, T. 1974 "What Is It like to Be a Bat", *Philosophical Review* 83, 435–450. (Quotations from the reprint in N. Block (ed.), *Readings in the Philosophy of Psychology*, Vol. 1, Cambridge, MA: Harvard University Press, 1980, pp. 159-168)

Place, U.T. 1969 "Is Consciousness a Brain Process", *The British Journal of Psychology* XLVII, pp. 44-50. (Quotations from the reprint in J. O'Connor, *Modern Materialism: Readings on Mind-Body Identity*, New York: Harcourt, Brace & World, pp. 21-31)

[14] I would like to thank Keith Lehrer for linguistic and philosophical advice with this article.

Modes of Perceptual Representation

Fred Dretske

Representational theories of the mind are at their best when dealing with thought-like states and processes, those that take (or, like reasoning and inference, operate on states that take) propositions as their object. They are at their best here because thoughts are identified and distinguished from one another in the same way representations are – by their content, by what they say about the matters whereof they speak. It seems natural, therefore, to take thought to be a form of internal representation.

Representational theories are at their weakest when applied to sensory states–those that comprise the way things look and feel, sound and taste. They are weakest here because sensory states are identified and distinguished from one another, in part at least, by intrinsic qualities (qualia) that have little or nothing to do with what (if anything) the sensation is a sensation of, with what (if anything) the sensation represents. Seeing someone play a piano differs markedly from hearing them play a piano even though these experiences are, in a sense, *of* the same thing: a person playing a piano.

Perceptual experiences not only differ from one another in striking and subjectively obvious ways, they also differ from the perceptual beliefs to which they normally give rise. In seeing (or hearing) a piano being played, one normally comes to believe that a piano is being played. Compared to the experiences, though, the belief that a piano is being played is a colorless affair indeed. One can believe a piano is being played without seeing or hearing it being played; and, conversely, one can see or hear a piano being played without believing that one is being played. Compared to an experiencer of played pianos, there isn't much it is like to be a believer in played pianos. Believing that a piano is being played is, in this sense, a *pure* representational state. Experiencing it being played, on the other hand, if it is representational at all, is *impure*. It is mixed with the modality-specific way the piano playing is represented. A representational theory of the mind, if it aspires to completeness, must be able to capture these facts about sense experience.

My purpose here is to say something about how this might be done, how representational ideas might be applied to both cognitive and sensory phenomena, to both perceptual experience and perceptual belief. My intent, furthermore, is to do this with a naturalistic theory of representation, a theory that conceives of representation (and, thus, intentionality) as an aspect of the physical world. I begin, therefore, with a sketch of a naturalized,

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 147-157

an information-based, account of representation.[1]

1 Information and Representation

Events carry information about other conditions in the world to the extent that they depend on the existence and character of these other conditions. For terminological convenience, I will say that one event carries information about another if the dependency is such that the first indicates something about the second. Thus, a pointer on an instrument carries information about a quantity if and only if positions of this pointer depend on the quantity being measured in such a way as to indicate the (approximate) values of this quantity.

I will not try to say what degree of dependency is needed for indication – and, hence, for an event to carry information about another. I shall merely assume that if one type of event depends on another in such a way that, in circumstances C, tokens of the first would not occur unless tokens of the second occur, then, when such circumstances obtain, tokens of the first carry information about the second, the information that the second occurred. This kind of dependency *suffices* for indication.[2]

Since this subjunctive conditional can be true without anyone knowing it to be true, this assumption is equivalent to assuming that the flow of information, the relation of indication, does not depend on our knowledge or understanding. Information, the power of one state of affairs to indicate something about another, is an objective affair, a fact that in no way depends on our interpretive or cognitive efforts. We may *use* information, but the information we use does not depend on our recognition or use of it for its existence.

Indicating *to* (or *for*) someone is, of course, an interpretive phenomenon. We are here concerned, however, not with what things mean or indicate to someone, but what they mean or indicate full stop. In this natural sense of meaning or indication, we discover what things mean, we do not stipulate or create this meaning.

For the most part we live in a lawful world, a world in which everything depends, to some degree, on a great many other things. Information, therefore, is all around us. It exists wherever there is the kind of dependency that enables one state of affairs to indicate something about another. Information, in this sense, is omnipresent.

Representation, as I shall use the word, is a much rarer commodity. Not every event that carries information about object o, the information (say) that o is F, represents o as being F. For r to represent o as being F it is not enough that r carry information that o is F (carrying this information

[1]The details of this theory of representation are in Dretske 1981, 1986, 1988 (Chapter 3).

[2]In Dretske 1981 I argued that such dependency was also *necessary* for one event to carry information that a second event occurred. I still think this is necessary, but I will not press the point here. It isn't essential to this paper.

is not even necessary). For r to represent o as being F r must either be an object, or be the state of an object, whose *function* it is to carry information about the F-ness of objects. It can be a thing's function to carry this kind of information without its actually carrying it, and it can carry this information without its being its function to carry it. An ordinary speedometer represents (by the position of a pointer) the speed of the car because it is the job, the function, the purpose of this device to indicate (by means of the pointer's position) how fast the car is going. When things go badly, when the device fails to provide the information it is its function to supply, then the position of the pointer misrepresents the speed of the car. It "says" something false. Speedometers in other cars also fail to supply information about the speed of my car, but, unlike my (broken) speedometer, they do not *misrepresent* my car's speed. They do not misrepresent the speed of my car because it is not their job to supply this kind of information. Other speedometers fail to represent, while my speedometer, when malfunctioning, misrepresents the speed of my car. Since none of them carry the information, the only difference is in their information-carrying function.

Representation, then, combines information-theoretic with functional or teleological ideas. If the concept of representation is to do its job in cognitive and semantic studies, if we hope to use this notion to clarify the nature of thought and meaning, then the idea must be rich enough to allow for misrepresentation; it must include the power to get things *wrong*, the resources for saying P when, in fact, P is false. This is what the teleology, the idea of something having an information-carrying *function*, is doing in the present conception. Since a thing can retain its function even when it fails to perform it (think of the heart or kidneys), an event can retain its information-carrying function–hence, continue to represent–even when things go badly, even when it fails to carry the information it is its function to carry (thus misrepresenting). It is the possibility of misrepresentation, this power to say or mean something that isn't so, that functions confer on information-processing systems.

Letting R be some system (organ, device, mechanism), we can express the idea of R representing some property or magnitude as follows:

(R1) R represents the property $\phi \equiv R$'s function (when suitably deployed) is to indicate the ϕ-ness of things.

If the object o is the thing whose ϕ-ness R represents at time t, then we can say that, at time t, R represents o. It may be deployed differently at another time – thus, at different times, representing different objects. Hence,

(R2) R represents o at $t \equiv$ For some ϕ, R represents (possibly misrepresents) the ϕ-ness of o at t.

If r_i are the assorted states which R gets itself into in order to indicate the ϕ-ness of o, then the r_i are representations. When R is working in the way it is supposed to, the way it does when it is doing its job, the

r_i carry information about o, the information that it is ϕ_i where ϕ_i are determinate forms of the determinable ϕ. So, for example, the speedometer has a mobile pointer the positions (r_i) of which indicate different car speeds (ϕ_i). The speedometer has the function of indicating car speed (ϕ). The pointer positions (r_i) are representations of determinate car speeds (40 mph, 55 mph, etc.). Given the function of the device, that is the information they are supposed to carry.

The speedometer in my car represents the speed (ϕ) of my car. The property or magnitude (ϕ) it represents is speed; the object it represents (o) is the object – in this case, my car – whose speed it represents. Install the same device in your car and it represents the same property (speed) but a different object – your car, not my car. Changing the deployment of a representational device changes what *object* it represents but not necessarily what properties it represents that object as having.

There may be many conditions in the causal chain between the object, o, being represented and the representation itself and r may carry information about all these conditions, but the condition r represents is the one it has the function of carrying information about. Photographs carry information, not only about the objects they are pictures of, but also about the representational device itself. Just as we can tell from the photograph that the subject was sitting or wearing a hat, we can tell that the camera was out of focus and the lens wide-angle. Nonetheless, the photograph has the subject and not the camera or lens as its object, as what it is a picture (representation) of because, given the conventional use (and thereby function) of photographic equipment, the resulting pictures are supposed to provide information about the object at which the camera is pointed – in this case, some person – not the mechanisms responsible for delivering this information. For the same reason a photograph is not a picture of the light reflected from the subject even though it carries as much or more information about the light than it does about the person.

2 Natural and Conventional Representations

In the case of gauges and instruments the functions on which the representational powers of a device depends are conventional functions in the sense that they depend for their existence on our – their designers and users – purposes, attitudes, and beliefs. We change an object's function merely by changing our collective mind about it or our collective use of it. We pour mercury in a glass tube, seal it up, paint a few numbers on its side, and call it a thermometer – thereby giving it the job of supplying information about temperature. Other pieces of metal, the volumes of which are likewise proportional to (thus carrying the same information about) temperature, do not, like thermometers, represent the temperature. They do not represent temperature, though they supply information about it, because it isn't their job to supply information about temperature. Most of them don't have jobs,

and those that do (e.g., paper clips, ball bearings, thumb tacks) have a different job. Hence, their information-carrying failures, when they occur, are not misrepresentations.

Unlike the functions underlying the representational efforts of artifacts, the functions of bodily organs are not conventional. The function of the kidneys in eliminating waste, the respiratory function of the lungs, and the circulatory function of the heart unlike the informational and directive functions of traffic signs, are a fact about these objects that in no way depends on their being conscious beings in the world to appreciate, understand or underwrite their functions. It is, in this sense, an objective fact about the heart that this is what it is for, what it is supposed to do. Harvey did not *create* a function for the heart. He *discovered* it.

Philosophers disagree about what functions are and how things acquire them. Some (e.g., Searle 1992: 238; Dennett 1987: 287ff) are skeptics about natural functions. I have nothing useful to say about this issue that can be said within the compass of two or three pages. The matter deserves careful attention, the kind of attention I cannot give it in this context. I, therefore, merely assume, as part of a representational approach to the mind, that there are natural functions, functions whose existence is not dependent on the purposes and intentions of conscious beings. No particular analysis of functions is essential to the argument. All that is essential to this project of naturalizing the mental is that there *be* natural functions, functions whose existence does not depend on the very things (experience, thought, intention, purpose) we are using the concept of representation (and, hence, function) to understand and analyze.

Descartes thought that the purpose of the perceptual systems was to "inform the mind" of what is beneficial and harmful. A good many contemporary philosophers and scientists would agree. If we read "function" for "purpose", and accept R1 and R2, then this is tantamount to endorsing a representational theory of perceptual experience and belief. If the senses do, indeed, have the job of informing the mind, of indicating how things stand in the world, then, like a good measuring instrument, they must, in doing their job, produce representations. If, in addition, the senses have this as their natural function – as, let us suppose, their biological function – then the representations they produce will exhibit what has been called *original intentionality*. The representations will be *about* the world under an aspect. The representations will say – perhaps correctly, but possibly incorrectly – how the world is in a certain respect. And, unlike human artifacts – speedometers and the like – they will say this without any help from us.[3]

[3]In saying that the representations will be about the world under an aspect I mean that in representing o as ϕ, a representation does not (necessarily) represent o as T even if o (or everything) always is T when it is ϕ. Hence, o's ϕ-aspect is represented without its T-aspect being represented. Representations have "aspectual shape" (Searle 1992) that they derive from their indicator functions. A device can have the function of carrying information about the ϕ-ness of things without having the function of carrying

This, indeed, is what makes perceptual experience and belief unique, something *more* than the mere (internal) effects of objects acting on the body, even those effects that carry information about their cause. There are plants – poison ivy, for instance – that can cause a certain neurophysiological condition (and subsequent skin rash) that carries information about the plant. Medical experts could describe the plant (three leaves, etc.) from diagnosing its effects on one. Yet, though this is a causal transaction in which information about the cause is conveyed to a person, there is (or need be) no conscious awareness, no perception, *of* the plant that poisons one. One does not, or need not, *experience* the plant or hold beliefs about it in order for it to effect one in this way.

Perception of the plant is different. When we perceive the plant, it has an effect on us that not only carries information about the cause (the plant itself), but is such as to make us aware of the cause. We experience (see or feel) the plant itself. This difference between information carrying causal transactions – the difference between seeing the plant and being poisoned by it – lies in the nature of the effect that the plant has on us. The sensory processes, those that culminate in an experience of the plant, and possibly also a belief about the plant, are conducted by mechanisms that have, as their natural function, the pick-up and delivery of information about the external cause. Hence, the result of such processes is a representation *in* the nervous system, not just an effect *on* the nervous system, of the external cause – the plant. The same is not true of allergic reactions to the plant. Whatever may be the function of the neurophysiological processes responsible for the body's reaction to poison ivy, it is surely not their *function* to provide information about this plant.

Because perceptual systems have an information-providing function, the job of indicating the state or condition of their causal source, it follows that sensory processes initiated by o, unlike allergic reactions to o, culminate in *representations* of o, internal states that have o as their object, as what they are of. In a way, then, an experience of o, the terminus of a perceptual processes initiated by o, is like an internal photograph of o. It is like a photograph of o, *not* because experiences are pictorial in character, but because, like pictures, they are representations, not merely (like allergic reactions) effects, of events that cause them. It is the representational nature of experiences, the fact that they are products of processes having an information-providing function, that gives them their power to be *about*, and not just (like allergic reactions) *caused by*, the objects and events they carry information about.

Up to this point, I have ignored – in fact, I have deliberately suppressed – the distinction with which I began, the distinction between a perceptual belief and a perceptual experience. I have suggested that if the senses do,

information about the T-ness of things even if it turns out that it cannot carry information about the one without carrying it about the other.

indeed, have the function of supplying an organism with information about its environment, then both our perceptual experience of the world and our perceptual beliefs about it can be conceived of as products of such mechanisms and, therefore, as representations of the objects in the world. But this leaves us with our original question: what is the *difference* between perceptual experience and perceptual belief? What, in representational terms, is the difference between seeing a piano being played and believing that a piano is being played? If both are natural representations of the piano playing, what makes them so different? Why do the experiences come laden with qualia while the beliefs are, by comparison, phenomenally barren?

I think we can make a beginning (but only a beginning) at answering these questions by looking at the different kinds of functions underlying natural representations. This will not give us a satisfying answer to all the questions philosophers have about the difference between belief and experience, between thought and sensation, but it will, I think, set us on the right track.

3 Ontogenetic and Phylogenetic Representations

Functions, both natural and conventional, can usefully be classified into those that attach to individual objects, devices, and mechanisms, in virtue of their membership in a certain kind or family – call these phylogenetic – and those functions that attach to individuals independently of such familial affiliations – call these ontogenetic. The biological function of bodily organs – the heart, the kidneys, the eyes – is, presumably, phylogenetic. Individual organs – your heart, for instance – gets its blood-pumping function from what Millikan (1984) calls the reproductive family of which it is a member. According to an evolutionary (i.e., natural selection) account of biological functions, the functions of individual bodily organs depend on the history of the species of which an individual organism is a member. Ontogenetic functions, on the other hand, attach to individuals in virtue of the way those particular objects perform (or, in the case of conventional functions, are regarded or intended to be used).

If we label a representation whose constitutive function is phylogenetic a P-representation and one whose constitutive function is ontogenetic an O-representation, we have taxonomy shown in Figure 1.[4]

Ontogenetic functions are sometimes superimposed on (and thus coexist with) phylogenetic functions. By declaring an appropriate intention, I can (in order to describe a famous battle) make an ash tray stand for an army division. Thanks to my declared intention, the ash tray's (ontogenetic) function is now to indicate, by its position and movements, the position and

[4] Although not shown in the figure, there can be conventional P-representations and O-representations. Token symbols and signs (e.g., a stop sign) presumably get their indicator functions from the family (type) of which they are members (tokens), but we could, in a game or by way of devising a code, give individual tokens of these symbols or signs a special function.

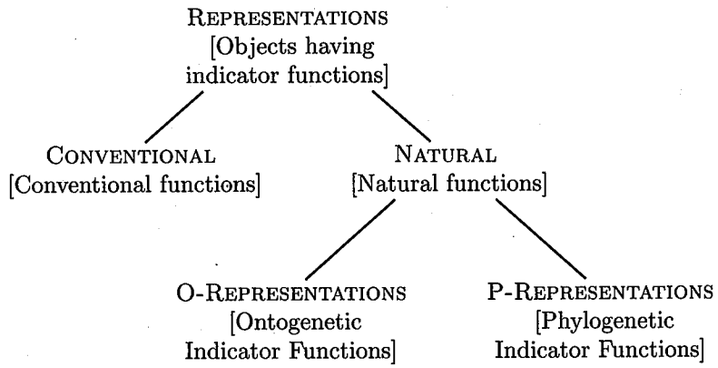


Figure 1: Taxonomy of Representations

movements of an army division. It nonetheless remains an ashtray – an object with an entirely different (phylogenetic) function.

To take a slightly more relevant example, suppose we take an ordinary pressure gauge and convert it to an altimeter. Since air pressure is inversely proportional to altitude, we re-calibrate the face of the gauge in feet-above-sea-level, add a few adjustments to compensate for variations in surface level pressure differences, and use the instrument as an altimeter in an aircraft. Though now being used as an altimeter, the instrument still does what it was designed to do, what it has the phylogenetic function of doing. It continues to indicate air pressure. It now, however, has (or has been given) a different or an additional (superimposed) indicator function. In virtue of the newly assigned function, the instrument now represents altitude. Of the many things an indicator indicates, the one it has the job of indicating is the one it represents. By changing the instrument's job description, by giving it the job of informing us about altitude rather than pressure, we have made it into an altimeter. The instrument, of course, still registers differences in air pressure. That, after all, is the *way* it detects changes in altitude. Now, however, it has a different job, the job of *using* information about air pressure in order to say something about (i.e., represent) altitude. A device whose function was (and, phylogenetically speaking, remains) to register air pressure has been given a new function, the ontogenetic function of indicating altitude. *P*-representations of pressure have been converted into *O*-representations of altitude.

I shall return to this kind of process, the process whereby a representation of one magnitude or property is transformed into a representation of another, in a moment. The process is important. It suggests a way of understanding how developmental processes, those that take place during the life of an individual organism, can superimpose conceptual representations

(i.e., beliefs) onto general purpose sensory representations (i.e., experiences) of the same objects.

4 Perceptual Experience and Perceptual Belief

Perceiving objects (persons, pianos) or events (e.g., a person playing a piano) is experiencing objects and events in a modality-specific way. Seeing, hearing, and feeling *o* is to have a visual, auditory, and tactile experience of *o*, the sort of experience that, when caused in the right way by *o*, makes one perceptually aware of *o*. All perceptual experiences have this *intentional* aspect. They are of something – whatever it is we are in perception consciously aware of.

Even hallucinations – which I take to be genuine experiences, just not experiences of a real object or event – *purport* to have something they are of or about. Though they lack an objective reference, they may nonetheless be indistinguishable from an experience that is of something real. This is a case where the representation (experience) represents a constellation of sensory properties (ϕs) but lacks an object, *o*, which has these properties. We still have representation R1, just not representation of any object (R2).

Experiences of *o* may (but may not) give rise to identificatory beliefs about *o*. Sometimes, when I see it at close range and in decent light, my perception of a golf ball gives rise to a belief that it (what I see) *is* a golf ball. I not only see (have visual experiences of) the golf ball, I see, recognize, identify it *as* a golf ball (see *that* it is a golf ball). At other times, when the lighting is bad or the ball is far away (but still visible), my perception of golf balls does not give rise to golf ball beliefs. When I was younger, in fact, before I learned what golf balls were, my perception of them – even the nearby ones – *never* gave rise to golf ball beliefs.

If, as I am now assuming, both my experience of the golf ball and my perceptual belief that it is a golf ball are representations of the golf ball, it is obvious that the experience is a mode of representation that, unlike the belief, does not await the kind of developmental (learning) processes necessary for knowing what golf balls are. We may learn to identify trees and people, golf balls and poison ivy, learn to recognize them for what they are, but we do not, at least not in the same way, learn to see and feel them. If the senses have an information providing function, then it seems safe to infer that this function, like that of the heart and kidney, is a phylogenetic function. We are, in the case of sense experience, talking about *P*-representations of objects.[5] It seems equally clear that beliefs are *O*-representations. To believe

[5] I do not claim that all *P*-representations are experiences, only that all experiences are *P*-representations and that it is this fact that explains the intentional character of experience. There are other aspects of experience – most prominently, their qualitative aspects – for which the representational story is not sufficient. There are doubtless many physiological processes – those used to carry out digestive and homeostatic chores, for instance, or those used by the immune system to detect and identify foreign cells – that have (or at least it is plausible to conjecture that they have) an information-carrying

that o is F one must have the concept F, know what Fs are, and for most (perhaps all) values of F, this requires individual learning.

To better appreciate this difference, think about my earlier example of perceiving a golf ball. I was five years old before I knew what a golf ball was. I had seen them before (I lived near a golf course), but I didn't know what they were made of, what they were used for, or what they were doing in the weeds where I often found them. Being of normal eyesight, I was, in seeing these golf balls, picking up and processing information about them. They looked (and felt) like small, hard, white spheres with a uniformly dimpled texture. My experience of these objects represented them as having this cluster of sensory properties. That is *still* the way my experience represents these objects, *still* the way they look to me in what philosophers choose to call the phenomenal sense of 'look'. Over the years, however, I have acquired a new way of representing golf balls. I now, knowing what golf balls are and how they look, represent them as (i.e., take them to be) golf balls. They still look much the same to me as they did when I was five years old,[6] but, as a result of learning, a conceptual mode of representation has supplemented, has been superimposed on, a pre-existing sensory mode of representation. It is a bit like the pressure gauge that was converted into an altimeter. In my case a golf ball detector – something capable of representing a small white dimpled sphere as a golf ball – was superimposed on a representation that represented it, the golf ball, as a small, white, dimpled sphere. Perceptual mechanisms that have the biological function of delivering information about, and thereby representing, the size, color, shape, and texture of things like golf balls are – through a developmental processes – supplemented with structures having the (ontogenetic) function of indicating, and thereby representing, the kind of object (golf ball? tennis ball? basketball?) that is being experientially represented by means of its shape, color, size, and texture.

Golf balls, of course, are artifacts that have not been around very long. If there is anything in us that represents them as golf balls, then, it is most implausible to suppose that the capacity to represent them this way is of phylogenetic origin. Nothing in us has the *biological* function of indicating that an object is a golf ball. So if golf ball beliefs are representations, and we accept the view that representations are structures having an

function. These are, therefore, P-representations, but they do not possess qualitative character. There is nothing it is like to have them.

This is simply to say that if perceptual experiences are P-representations, they are a very special kind of P-representation. I agree with this. I also agree that I have not said what it is about them that makes them so special. My present task, however, is more modest. It is not to describe the special kind of P-representation that experience is, but merely to distinguish perceptual beliefs from perceptual experiences and to locate, in the idea of P-representation and O-representation, one source of this difference.

[6] I do not take 'looking like a golf ball' to be part of the phenomenal appearance of golf balls. Nor do I wish to deny that learning what something is may influence the way it looks.

indicator function, then golf ball beliefs, unlike golf ball experiences, must be O-representations of golf balls. Unlike one's capacity to experience (see and feel) golf balls, one's capacity to represent them as (believe them to be) golf balls is a representational capacity that develops during the life of the individual believer. The indicator functions of belief are ontogenetic.

The difference between a golf ball belief and a golf ball experience, then, is that (under normal conditions) the experience indicates the color, shape, size and texture of the ball. That is its job. It does not represent the ball as a golf ball. It is not the function of a visual experience to indicate the kind of sport in which a ball is used. That is, however, the job of a golf ball belief. The belief represents the ball as a *golf* ball and, unlike the experience, is totally silent about its color, shape, size and texture. This is not to say, of course, that the golf ball belief cannot be accompanied by beliefs about its color, shape, size and texture. It is only to say that these are separate beliefs. One can believe that o is a golf ball while being ignorant of its shape, size, color or texture (i.e., the properties that are represented in an experience of a golf ball).

Perceptual experiences, then, are representations that provide an organism with the kind of all-purpose information (about color, size, shape, texture, movement, etc) required to identify prey and predators, shelter and mates. Exactly what one would expect of a mechanism whose information-carrying duties had a phylogenetic origin. Perceptual beliefs, on the other hand, are dedicated representations of the same objects, representations that exploit the information in an experience in order to classify perceptual objects into relevant natural or artifactual kinds. With organisms capable of learning, the relevant kinds are normally determined by ontogenetic factors. Thus the distinction between the phylogenetic functions of experience and the ontogenetic functions of belief.

References

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

Dretske, F. 1986 "Misrepresentation". In Radu Bogdan (ed.) *Belief*, Oxford: Oxford University Press.

Dretske, F. 1988 *Explaining Behavior.* Cambridge, MA: MIT Press.

Millikan, R. G. 1984 *Language, Thought, and Other Biological Categories: New Foundations for Realism*, Cambridge, MA: MIT Press.

Dennett, D. 1987 *The Intentional Stance*, Cambridge, MA: MIT Press.

Searle, J. 1992 *The Rediscovery of Mind*, Cambridge, MA: MIT Press.

Objects of Thought

Ernest Sosa

"What makes my idea of him an idea of *him*?" So queried Wittgenstein, and his query is part of a family. For example, you can refer to someone through an idea but also through a thought. So: "What makes my thought about him a thought about *him*?" I defend a liberal answer to such questions, an answer so liberal as to be latitudinarian:

(L) S thinks x to have property ϕ if there is an individuating concept α satisfied (uniquely) by x in a context C such that, in C, S thinks *de dicto* a proposition that predicates ϕ with respect to α.[1]

The proposition that the pen in my hand is black predicates being black with respect to the individuating concept ⟨the pen in my hand⟩. Individuating concepts are expressed by singular terms, such as the definite description 'the pen in my hand'. We may think here of the relevant context C as constituted by a subject S and a time t. When does a pen x satisfy (uniquely) the individuating concept ⟨the pen (now) in my hand⟩ in such a context C constituted by S and t? When x is uniquely a pen in S's hand at t. Finally, if S thinks *de dicto* (the proposition) that the F is G, then not only the property of being G but also the individuating concept ⟨the F⟩ both enter into his conception of how things stand.

We arrive thus at an answer to Wittgenstein's question: My thought about him is about *him* in virtue of having for its content a proposition that predicates something with respect to some individuating concept which, in the context, is satisfied (uniquely) by *him*.

A

Here, first of all, is an argument for the sufficiency of such *de dicto* thought for its *de re* counterpart. It has three premisses, in the form of three corresponding principles, (P1)–(P3), and three conclusions: the full and explicit (C1) and (C2), and the briefer (C3), which abbreviates the earlier two. (I focus on the case of belief, but the argument can be extended easily to cover also other attitudes and forms of thought.)

[1] A more linguistic version of latitudinarianism is defended in my "Propositional Attitudes *De Dicto* and *De Re*" and "Rejoinder to Hintikka". Note, moreover, that (L) permits individuating concepts to be "perspectival" or "indexical", since they need be satisfied not absolutely but in a context C, one that can supply the indices or parameters required for such a concept to be satisfied.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 159-167

(P1) If there is such a thing as the F, then the proposition that the F is G is about that thing (the F), and attributes to it the property of being G.

(P2) $\forall x$(If S believes a proposition P and P is about x, and attributes to x the property of being G, then, in believing P, S has a belief that is about x and attributes being G to x).

(P3) $\forall x$(If, in believing a proposition P, S has a belief that is about x and attributes being G to x, then in that belief S refers to x and attributes being G to x).

(C1) If S believes (the proposition) that the F is G, and there is such a thing as the F, then, in believing that proposition, S has a belief that is about that thing (the F) and attributes being G to it.
By (P1), (P2).

(C2) If S believes (the proposition) that the F is G, and there is such a thing as the F, then, in believing that proposition, S refers to that thing (the F) and attributes being G to it. By (C1), (P3).

(C3) If S believes that the F is G, and there is such a thing as the F, then the F is believed by S to be G (and S believes the F to be G).
By (C1) or (C2).

Unless one is prepared to reject at least one of the three principles (P1)–(P3), one is therefore committed to accepting the sufficiency of a *de dicto* belief about an entity for its corresponding *de re* belief about that entity. (To avoid this conclusion, moreover, one must reject *both* (P2) and (P3), so long as one accepts P1.) In contrast with such latitudinarianism, according to a narrower doctrine of *de re* attitudes you do not attain genuine reference to a particular object merely by having in your thought some concept that picks out that object. To think about, to refer in thought to, the tallest spy, it is not enough simply to have the thought that the tallest spy is bound to be a spy, even supposing there is such a thing as *the* tallest spy. According to this narrower doctrine (N) genuine reference requires some special relation binding the thinker with the object of reference, probably some causal psychological relation like perception or memory.

B

Consider now a case where there is such a thing as the F, and there is such a thing as the G. And suppose that though these are actually distinct things, a subject S believes *falsely* that the F = the G. In that case, if we just assume S to be a nimble and powerful enough logician, then the F and the G will be *indistinguishable* from the perspective of S's mind. No propositional attitude of S's will have a content constituted somehow by the individuating concept ⟨the F⟩ unless it is matched by a corresponding propositional attitude whose content is constituted analogously by the individuating concept ⟨the G⟩. S's

mind will contain no belief, desire, hope, fear, or any other attitude about the F without containing a matching attitude about the G. How then can we possibly understand selective action by S that discriminates in favor of one (e.g., the F) over the other (the G)?

The moral we are meant to draw is that for selective action by S on the F in preference to the G, we must postulate some stronger *de re* attitude by S about the F itself, one that will discriminate the F from the G and enable selective action on it by S. Such more substantial *de re* attitudes require some real causal relation, perhaps a relation of perception, that connects S selectively with the F, and discriminates the F from the G, enabling S's selective action upon it, despite the fact that the F and the G are indistinguishable within the perspective of S's *de dicto* attitudes.

Return, for example, to the case of the black pen in my hand. And suppose I had recently received as a gift a pen of that make, model, and color. Indeed, in that situation I believe that the pen in my hand = my recent gift. However, as it happens, my wife also has a pen of that make, model, and color, and the two have just been switched by accident so that actually the pen in my hand = my wife's pen (and my recent gift = the pen on my wife's desk). Suppose now my son walks in as I stand by my wife's desk, and asks me to show him my recent gift. How can I possibly respond rationally and reasonably to his request, given the latitudinarian account of *de re* attitudes? How can I do so when I believe or desire about the pen in my hand all and only what I believe or desire about the pen before me on my wife's desk? The two pens are absolutely indistinguishable from my own mind's perspective. If I believe about the pen in my hand that it is a recent gift to me, I must believe the same of the pen on my wife's desk; if I believe of either that it is black, I must believe the same of the other; and, finally, if I desire to show either one to my son now, I must desire the very same, and to the same degree, with regard to the other. How then can I manage to act selectively and discriminatingly on one rather than the other?

The moral we are meant to draw is, again, that I must bear some special, more direct and causal relation to one of the pens, to the one in my hand, perhaps: some special relation that does not connect me equally with the pen before me on my wife's desk. It is this difference that is meant to explain my showing the pen in my hand, rather than the pen on my wife's desk, in answer to my son's request. And doubtless there *are* differences in the causal and perceptual relations that connect me with the two pens, but we should not jump to the conclusion that my *de re* references to the pen in my hand must be *constituted* by some such further relations that I bear to it (and do not bear to the pen on my wife's desk). For we can explain my differential action, as I show the pen in my hand to my son, without departing from latitudinarianism, as follows.

While it is true for latitudinarianism that I have the same *de re* beliefs about the two pens, some such beliefs are especially pertinent to which pen ends up being shown to my son in answer to his request. And there is a

salient difference here as we compare these beliefs about the pen in my hand with the corresponding beliefs about the pen on my wife's desk: namely, that the beliefs in question are in fact true of the pen in my hand while false of the other pen. For example, I believe both of the pen in my hand and of the other pen that it is in my hand. But this *de re* belief is of course true only of the former. More generally, an agent S will act on an intended object of action O only if S has *true* beliefs that relate O appropriately to actions under S's direct control at the time. Thus in our example I might believe: "If I raise my hand, I will display the pen in my hand (i.e., my recent gift)", but this would be true only of the pen in my hand (I will display it), and not of my recent gift (*it* will not be displayed, but will remain on my wife's desk). It is this difference that explains in latitudinarian terms why I succeed in showing my son the pen in my hand without showing him my recent gift (though the two are indistinguishable from the perspective of my mind). Such an explanation is unaffected by the fact that the two pens are thus indistinguishable, which means that the two are indistinguishable in respect of my *de re* beliefs. What matters is that some important such beliefs of mine are true of the pen in my hand and false of the other pen (my gift, the pen on my wife's desk), which determines, for example, which pen is at hand and under my immediate control.[2]

More generally, for S to ϕ entity x intentionally at time t, there must be a property F of x which both *locates* x for S to ϕ by performing B at t and *motivates* S to ϕ x by perforing B at t, where B is *basic* relative to ϕ-ing, S, and t. We need some account for each of the three expressions in italics.

- An action B is *basic* relative to an action A (e.g., ϕ-ing x), an agent S, and a time t, iff (i) S at t intentionally performs-A-by-performing-B, and (ii) there is no action C distinct from B such that, at t, S intentionally performs-B-by-performing-C.
- Property F *locates* x for S to ϕ by performing action B at t iff x is the thing with F and, at t, S knows *de dicto* (or at least is correct in thinking, *de dicto*) that by performing B in the circumstances, he would then ϕ the thing with F.
- Property F *motivates* S to ϕ x by performing B at t iff x is the thing with F and at t S wants *de dicto* (sufficiently, given all relevant considerations)[3]to then ϕ the thing with F by performing B.

In our example, we may explain why I act selectively with regard to one of the two perspective-indistinguishable objects as follows. There is a single property that both *locates* one of the two objects and *motivates* me to show it by performing some action of mine that is *basic* with regard to my

[2]But this is not to suggest that only one's true *de re* beliefs about an object can help explain one's actions on that object.

[3]Thus we are dealing here with "all-things-considered" motivation.

doing so. But there is no such property that can play such a role relative to myself and the other object, given the example as presented.

The pen in my hand, x, has the property H (of being the pen in my hand) such that H locates x for me to show x by performing the action E (of extending my hand in a certain way) such that E is basic relative to myself and the action of showing x (and the time in question), and, moreover, that property H also motivates me to show x by performing E (at that time).

However, there is no such property of the pen on my wife's desk, no property that plays the dual role of locating it and motivating me to show it by performing some action of mine that can be basic relative to my showing it (at that time). That pen *is* my recent gift, and this property of it (G) *does* motivate me to show it by extending my hand in a certain way. However, property G does not locate that pen for me to show it by performing any basic action of mine in the circumstances. Nor is there any other property that plays such a dual role with regard to myself and the pen on my wife's desk, in which respect that pen is distinct from the pen in my hand: this pen does have a property, property H, that plays the dual role of locating it and motivating me.

In our example, the following practical syllogism is available to me:

1 My recent gift = the pen in my hand.

2 I will show my recent gift.

3 I will show the pen in my hand.

It is this syllogism that explains my wanting to show the pen in my hand, and it can be extended further as follows:

4 If I extend my hand in a certain way, I will thereby show the pen in my hand.

5 I will show the pen in my hand by extending my hand in that way.

Of course, the move from 4 to 5 is fraught, and succeeds only on pertinent assumptions, which we may assume to be given in the circumstances.

C

Compare a case from mathematics. S believes that $\sqrt{1089} = 43$ (although in fact $\sqrt{1089} = 33$). Of course, S has many other related beliefs as well: e.g., that $43^2 = 1089$, and that $\sqrt{1089} = \sqrt{1089}$. And, being a nimble and powerful logician, since S *believes* that $\sqrt{1089} = 43$, and since $\sqrt{1089} = 33$, there is no belief that S has about 33 (*de re*) not matched by a corresponding belief about 43. (Not if we accept the implications of latitudinarianism.) In particular, both with regard to 43 *and* with regard to 33, he believes that it is identical to $\sqrt{1089}$, and with regard to each (and with regard to $\sqrt{1089}$ as well), he believes that its standard name is '43' and also that its standard name is '33'! So how can we explain his acting selectively with regard to

43, when asked to give the standard name of $\sqrt{1089}$, by giving the standard name of 43, namely '43', rather than the standard name of 33, i.e., '33'? Why does S act thus even though he believes that '43' is the standard name *both* of 43 *and* of 33 (as he does given latitudinarianism), and believes as well that '33' is the standard name of each? Moreover, although S believes 43 to be $\sqrt{1089}$, true enough, he also believes 33 to be $\sqrt{1089}$ as well (or so he does according to latitudinarianism). So, again, why does he answer '43' when asked for the standard name of $\sqrt{1089}$? In order to avoid getting side-tracked from our specific questions, let's suppose that the request put to S takes the following specific form: What is the standard name of the number picked out by '$\sqrt{1089}$'?

In order for S to give the standard name of 43 intentionally at t there must be a property F of 43 which both locates 43 for S to give its standard name by performing B at t and motivates S to give its standard name by performing B at t, where B is basic relative to giving the standard name of 43, S, and t.

Saying '43' is basic relative to giving the standard name of 43, S, and t, iff (i) S at t intentionally gives the standard name of 43 by saying '43', and (ii) there is no action C distinct from saying '43' such that at t S intentionally says '43' by performing C.

Being the number 43 locates 43 for S to give its standard name by saying '43' at t iff 43 is the number with the property of being the number 43 and at t S knows *de dicto* (or at least is correct in thinking *de dicto*) that by saying '43', in the circumstances, he would then give the standard name of the number 43.

Being the number 43 motivates S to give the standard name of 43 by saying '43' at t iff 43 is the number with the property of being the number 43 and at t S wants *de dicto* (sufficiently, given all the relevant considerations) to then give the standard name of the number 43.

However, there is no corresponding property of the number 33 that will play the role *vis a vis* it that the property of being the number 43 pays above *vis a vis* the number 43. There is no such property of 33 that both locates it for S to give its standard name and motivates S to give its standard name by performing some basic action, such as saying '33'. In particular, S does not want (*de dicto*) to give the standard name of the number 33.

The following practical syllogism is available to S:

1 $43 = \sqrt{1089}$

2 The standard name of $\sqrt{1089}$ = the standard name of 43.

3 I will give the standard name of $\sqrt{1089}$.

4 I will give the standard name of 43.

That syllogism explains why S wants (*de dicto*) to give the standard name of 43 (by saying '43'). But S has no corresponding syllogism for 33 (and '33'). A practical syllogism is within the scope of *de dicto* psychological

attitudes, so that substitution of '33' for '$\sqrt{1089}$' at step 3 of the syllogism above is unavailable. Moreover, S does not believe that $33 = \sqrt{1089}$, so we do not even attain the corresponding first premiss.

D

Latitudinarianism is meant to cover propositional attitudes in general and intention in particular. Thus:

(LI) S intends to ϕ x iff there is an individuating concept α satisfied uniquely by x in a context C, such that, in C, S intends *de dicto* a proposition that predicates ϕing with respect to his first-person concept of himself (his concept of himself as himself) and α (in that order).

If S intends *de dicto* a proposition that predicates ϕing with respect to himself (as himself) and ⟨the F⟩, in that order, then not only the property of ϕing, but also the individuating concepts ⟨Myself⟩ and ⟨the F⟩ both enter into his conception of how, according to his intention, things are meant to turn out. Thus if I intend *de dicto* (the proposition) that I feed the cat before me, then not only the property of feeding but also the individuating concepts ⟨Myself⟩ and ⟨the cat before me⟩ both enter into my conception of how things are to turn out.

However, according to a further objection, such latitudinarianism is refuted by the "shell game" problem:

> Suppose there was earlier a pepper mill (x) to the left of S, and that S saw it and (mis)took it to be a full salt-shaker. Later the pepper mill was removed and replaced with a full salt-shaker (y), without S's realizing it and without S's ever perceiving or knowing of y. S now reaches to the left in order to pick up the object he believes to be there (the pepper mill, presumably, the one that he erroneously believed and believes to be a full salt-shaker). S does pick up the object that is now there: as it happens, fortuitously, a full salt-shaker, but not the object S was reaching for, the object S believed and believes to be there.

It has been objected that in this example the conditions laid down by latitudinarianism (LI) for S to intend to ϕ y (the salt-shaker now to S's left) are all satisfied.[4] For, we are told, ⟨the salt-shaker to the left⟩ is an individuating concept α such that, in the context of that example (C):

- α is uniquely satisfied by the salt-shaker y in that context C; and
- S, in C, intends *de dicto* a proposition that predicates reaching with respect to himself and α in that order.

Yet, according to the objection, the agent S really intends to reach the pepper mill rather than the salt-shaker, since it is the pepper mill that he saw

[4] This problem is drawn from Thomas McKay, "Actions and *De Re* Beliefs".

and remembers and believes to be still there (even though he misperceived it and erroneously believed and believes it to be a full salt-shaker).

Here again the example is supposed to cast doubt on any latitudinarian conception of reference in thought, and to promote a narrower conception of the sort indicated by N above. It is supposed that through a combination of memory and perception the agent S is genuinely related in thought to the pepper mill and not to the "imposter" salt-shaker, which merely happens to fulfill by accident the abstract conceptual content of S's intention.

For a solution to the puzzle which preserves the spirit of our latitudinarian approach we need three concepts as follows:

(C1) α is an *individuating concept* (or individuator) of x for S at t iff α is a concept that only x satisfies relative to the perspective of S at t (e.g., being to his left).

(C2) Individuator β is *epistemically derivative* from individuator α for S at t iff S at t knows (or believes) that something satisfies β on the basis of knowing (or believing) that something satisfies α and that whatever satisfies α satisfies β, but not *vice versa.*

(C3) If β is epistemically derivative from α for S at t, if y satisfies β relative to S and t, and x satisfies a relative to S and t, and not $(x{=}y)$, then concerning S's thought $[\phi, \beta]$ at t, x is a *deeper* and *more primary* object of that thought by S at t than is y (though y is also an object, a relatively superficial, secondary, object of that thought by S at t).

There *is* then a sense in which the subject is "really" reaching for the pepper mill, but a sense acceptable to latitudinarian intuitions. For, relative to that subject and time, and to the situation described by the example, the pepper mill is a deeper and more primary object of his thought, since the individual concept *the salt-shaker now to my left* is epistemically derivative for S and t from *the container I saw to my left a while back*, or perhaps on *the object with such and such a shape and color that I saw to my left a while back.*[5]

References

Donnellan, K. 1966 "Reference and Definite Descriptions", *The Philosophical Review* LXXV, 281–404

[5]The proposed solution is due materially to David Sosa. Other puzzles seem also amenable to this approach: for example, the martini-drinker case and other cases in Keith Donnellan's "Reference and Definite Descriptions". It is a solution in a latitudinarian spirit since it does not explicate reference of any sort through any brute causal relation, but insists that reference, even deep and primary reference, is always under a description or through an individuating concept or sense. For this "broad-gauge latitudinarianism", then, belief under a description or through an individuating concept (any such description or concept) is both necessary and sufficient for superficial or secondary reference, and belief under a description or through an individuating concept is necessary even for deep or primary reference.

McKay, T. 1984 "Actions and *De Re* Beliefs", *Canadian Journal of Philosophy* 14, 631–635

Sosa, E . 1970 "Propositional Attitudes *De Dicto* and *De Re*", *Journal of Philosophy* 67, 883–96.

Sosa, E. 1971 "Rejoinder to Hintikka", *Journal of Philosophy* 68, 498-501.

Do Pains Have Representational Content?

Michael Tye

Philosophers have usually supposed that pains are not like images or memories or visual percepts: they have no representational content. So, it is widely held that pain provides a counterexample to Brentano's Thesis, that intentionality is the mark of the mental. This, I now believe, is much too hasty. What I shall argue in this paper is that pains *are* representational.

My discussion begins with an old objection to the token identity theory in connection with after-images, and a modern response to it which has become widely accepted. This response, I shall show, is unsatisfactory, as it stands. But, with one key revision, it is, I believe, defensible, and it has ramifications for our understanding of pain. In particular, it points to the conclusion that pains have representational content, as does at least one other facet of our everyday conception of pain. In Section II, I consider the question of what sorts of representations pains are most plausibly taken to be. Are they sentences in an inner language like beliefs and desires, on the usual computational conception of the latter states? Or are they representations of a different sort? I suggest that a purely sentential approach is difficult to reconcile with some of the neuropsychological data on pain, and I make an alternative hybrid proposal.

I

In the 1950's J.J.C. Smart raised the following objection to the identity theory for sensations: After-images are sometimes yellowy-orange; brain processes cannot be yellowy-orange.[1] So after-images are not brain processes. The reply that Smart made to this objection was to deny that after-images exist, there really being, in Smart's view, only experiences of *having* after-images, which are not themselves yellowy-orange.

This response requires further elaboration. In particular, an account is needed of experiences of having after-images which does not itself presuppose that after-images exist. Such an account is, of course, available. It is commonly known as the adverbial theory.[2] On this view, having an after-image is a matter of sensing in a certain way rather than experiencing a certain object, an image.

I would like to thank Sydney Shoemaker and Ned Block for helpful comments.

[1] See Smart 1959.

[2] See Tye 1984.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 169-177

Another less radical response is available on behalf of the identity theory. It is on this response which I wish to focus here. Why not say that in predicating color words of images we are not attributing to them the very same properties which we attribute to external objects via our use of color language? So, after-images are not literally green or blue in the way that grass or the sky have one or the other of these features. Now it is no longer obviously true that brain processes cannot be yellowy-orange in the relevant sense.

The obvious problem which this response faces is that of explaining how it is that color vocabulary is applied at all to after-images, given that they do not really have the appropriate colors. One solution proposed by Ned Block is to say that color words are used elliptically for expressions like 'real-blue-representing', 'real-green-representing', and so on, in connection with images generally.[3] In my view, this solution has a number of important virtues.[4] To begin with, brain processes can certainly represent colors. So, the identity theory is no longer threatened. Secondly, as Block has noted,[5] terms like 'loud' and 'high-pitched' are standardly applied directly to oscilloscope readings used in connection with the graphical representations of sounds. In this context, these terms evidently do not name real sounds made by the readings. One possibility, then, is that they pick out representational properties such as loud-representing and high-pitched-representing. If this is so, then there already exists an established usage of terms which conforms to the one alleged to obtain in the case of color terms and after-images. Finally, it seems to me plausible to suppose that when we say that an after-image, or indeed any image, is blue, say, we are adverting to its phenomenal character. But there are, I believe, independent reasons for thinking that the phenomenal character of visual experiences is intentionally based.[6] So we have here further support for the proposed representational reading of color terms as applied to images.

There is a serious difficulty, however. Mental images are not literally square any more than they are literally blue. So, extending the above proposal to shape, we get that a blue, square after-image is an image that is square-representing and also blue-representing. From this it follows that an image that represents one thing as blue and another *distinct* one as square (for example, an image of a blue circle next to a red square) itself counts unequivocally as a blue, square image.[7] This is intuitively wrong-headed. Surely a blue, square image cannot represent different things as blue *and* square. 'Blue', then, in application to images, cannot mean 'blue-representing'. Likewise 'square'.

This difficulty is not peculiar to images. Precisely the same problem

[3]See Block 1983, especially p. 518

[4]These virtues led me to accept the proposal until very recently. See Tye 1991.

[5]See Block 1983, 516–517.

[6]See here Tye 1992b and 1994.

[7]This is a version of Frank Jackson's Many Property Problem. See Jackson 1977.

can be raised in connection with oscilloscope readings. The way out, I suggest, is to appreciate that there is nothing elliptical about the meanings of terms like 'blue' or 'loud' in the above contexts. Instead it is the contexts themselves that need further examination. Let me explain.

The contexts 'hopes for an F' and 'hallucinates a G' are typically intensional. Thus, I can hope for eternal life and hallucinate a pink elephant, even though there are no such things. Similarly, I can hope for eternal life without hoping for eternal boredom, even if in reality the two are the same. It seems evident that the terms substituting for 'F' and 'G' in these contexts retain their usual meanings. The above peculiarities are due to the fact that hoping and hallucinating are representational states, and to the special character of representation itself.

Now precisely the same peculiarities are present in the case of the context 'an F image', where 'F' is a color or shape term. Thus, in a world in which nothing is really triangular, I can still have a triangular image. Also, if I have a red image, intuitively it does not follow that I have an image the color of most fire-engines, even given that most fire-engines are red. The explanation, I suggest, is straightforward: An F image is an image which represents that *something* is F.

Likewise, an F, G image is an image which represents that something is *both* F and G.[8] My suggestion, then, is that there is nothing elliptical or peculiar about the meanings of the terms 'F' and 'G' in the context 'an F, G image'. Rather the context itself is an intensional one, having a logical structure which reflects the representational character of images generally.

It may still be wondered why we *say* that the image itself is F and G, for example blue and square. This is, I suggest, part of a much broader usage. Frequently when we talk of representations, both mental and non-mental, within science and in ordinary life, we save breath by speaking as if the representations themselves have the properties of the things they represent. In such cases, in saying of a representation that it is F, what we mean is that it represents that something is F. So, when it is said of some given oscilloscope reading that it is loud and high-pitched, what is being claimed is that loud and high pitched are features that the reading represents some sound as having. 'Loud' and 'high-pitched' mean what they normally do here. The context itself is intensional.

The above proposal solves the problem of the blue, square image. Here the image represents that something is both blue and square, not merely that something is blue *and* that something is square.[9] We are now ready to turn

[8]In my view, an F, G image need not represent that some *material object* has the properties, F-ness and G-ness. After-images typically represent that regions of space have the appropriate properties. See here Tye 1991, p. 121.

[9]I might add that, in my view, a necessary condition of any image representing that something is both F and G is that it represent that something is F. So, if I have an F, G image, I must have an F image. The argument for the premise here is straightforward: in having a blue, square image, I experience blue as a feature of some object or region of space, a feature co-instantiated with square. *What* I experience, in part, is *that* something

to the case of pain.

It is often supposed that terms applied to pain which also apply to physical objects do not have their ordinary meanings. Ned Block, who takes this view, says the following:

> There is some reason to think that there is a systematic difference in meaning between certain predicates applied to physical objects and the same predicates applied to mental particulars. Consider a nonimagery example: the predicate '— in —'. This predicate appears in each premise and the conclusion of this argument:
>
> The pain is in my fingertip.
>
> The fingertip is in my mouth.
>
> Therefore, the pain is in my mouth.
>
> This argument is valid for the 'in' of spatial enclosure..., since 'in' in this sense is transitive. But suppose that the two premises are true in their *ordinary* meanings... The conclusion obviously does not follow, so we must conclude that 'in' is not used in the spatial enclosure sense in all three statements. It certainly seems plausible that 'in' as applied in locating pains differs in meaning systematically from the standard spatial enclosure sense. (Block 1983, p. 517).

This seems to me quite wrong at least for the case Block cites. There is no more reason to adopt the strange position that 'in' does not mean spatial enclosure in connection with pain than there is to say that 'orange' in connection with images has a special meaning. With the collapse of the latter view, the former becomes unstable. And the inference Block cites does *not* establish his claim. To see this, consider the following inference:

I believe that I am in Buffalo.

Buffalo is in the USA.

Therefore, I believe that I am in the USA.

The term 'in' has the same meaning in both premises and the conclusion. But the argument is invalid: I might conceivably believe myself to be in Buffalo whilst still believing that I am in Canada. The same is true, I suggest, in the case of Block's example, and the explanation is the same. In both the first premise and the conclusion, the term 'in' appears in an intensional context. Just as when we say that an image is blue, we are saying that it represents that something is blue, so when we say that a pain is in my fingertip, we are saying that it represents that something is in my fingertip.

is blue. So, my image, in part, represents that something is blue.

It is perhaps tempting to suppose that the image itself is the object of the experience here, so that if what I experience is that something is blue, then the image must be blue. But this conflates two senses of 'object of experience'. In one sense, an object of experience is a (non-abstract) item which enters into the content of the experience. In another sense, an object of experience is a (non-abstract) item to which the subject of the experience is related, an item which bears the content. Images (including after-images) are objects of experience in the latter sense, but not in the former.

That there is a hidden intensionality in statements of pain location is confirmed by our talk of pains in phantom limbs. We allow it to be true on occasion that people are subject to pains in limbs that no longer exist. How can this be? Answer: You can have a pain in your left leg even though you have no left leg, just as you can search for the Fountain of Youth. Again the context is intensional: specifically, you have a pain that represents that something is in your left leg.

Of course, there is some temptation to say that if you don't have a left leg, then you can't really have a pain in it. But that is no problem for my proposal. For there is a *de re* reading of the context, namely that to have a pain in your left leg is for your left leg to be such that you have a pain in *it*. Now a left leg *is* required.

But doesn't a pain in the leg represent more than just that something is in the leg? To answer this question, it is necessary to make some more general remarks about pain. To have a pain is to feel a pain, and to feel a pain is to experience pain. Thus, if I have a pain, I undergo a token experience of a certain sort. This token experience is, I suggest, the particular pain I have.

Now *what* I experience or feel, in having a pain in a leg, is that something in a leg is painful or hurts. So, a pain in the leg is a token experience which represents that something in the leg is painful or hurts.[10]

This proposal may seem to encounter an immediate difficulty. Token experiences, including pains, are themselves located in the brain. So, there really are no pains inside legs. So, the above experience, in representing that something in the leg is painful, must be *mis*representing what is going on. And this is highly counter-intuitive. Surely, a person who feels a pain in the leg, in normal circumstances, is not subject to an *illusion*.

This objection contains a *non sequitur*. From the fact that there are no pains in legs it does not follow that a pain which represents that something in a leg is painful is a misrepresentation. When it is said that a cut in a finger or a burn or a deep bruise is painful or hurts, what is meant is simply that it is *causing* the person's experience of pain, that the token pain he or she is undergoing is caused by it. So, a pain in the leg is a pain which represents, under the appropriate mode of representation, that something in the leg is causing *that very pain*. Since pains in legs are normally caused by

[10] According to pain researchers, people who have been given prefrontal lobotomies, or certain other treatments, often report that they feel pain but that they do not mind it or that it does not really bother them. See here Melzack 1961. These reports, even if taken at face value, do not threaten the proposal in the text. For what they entail, on a literal reading, is that pain can occur without the desire to be rid of it. So, as long as it is granted that pains can represent that parts of the body are painful or hurt without eliciting the desire that they cease, there is no difficulty in accomodating the reports. This position, I might add, is compatible with holding that pain is aversive in normal circumstances (and perhaps essentially so), even though it is not essentially aversive *simpliciter*.

disturbances of one sort or another inside legs, such pains do not normally misrepresent.

This account seems to me intuitively very plausible. A man who reports to his doctor that he has a pain in his left arm is not taken to have lied, if it is discovered that the real cause of his pain lies in his heart. Such a man has a pain in his left arm, but in this case he *is* under a kind of illusion: there really is nothing in his *left arm* which is hurting him.

There is one further objection worth mentioning here. Perhaps it will be said that a person who experiences a pain in a certain bodily part does not experience that something inside the relevant part is *causing* the experience. Such a proposal is too complicated to fit the phenomenology of experiences of this sort. If, for example, I have a pain in my left thumb, I surely do not actually feel that something in my thumb is causing the very pain experience I am undergoing.

This objection is not compelling. It is true that when I have a pain in my left thumb, what I experience is simply that something in my left thumb hurts. But this is certainly compatible with holding that the experience is veridical, that what it represents is the case, if and only if something inside my thumb is causing my token pain experience. For the context 'experiences that' is highly intensional: not even analytically equivalent expressions may be safely substituted *salva veritate*.

I am inclined to suppose that an intentionalist treatment can be given of the use of a number of other terms for pain that are also used to describe physical objects. Consider, for example, a stinging pain in the leg. Here, it seems phenomenologically undeniable that stinging is experienced *as* a feature tokened within the leg, and not as an intrinsic feature of the experience itself. What is experienced as stinging is something *inside* the leg. A stinging pain in the leg, then, is a pain which represents that something in the leg is stinging. Likewise, a burning pain in the leg is a pain which represents that something in the leg is burning.

It may be objected that if stinging pains represent what bees and wasps do and burning pains represent the production of heat by fires, then the proposal seems very counter-intuitive. It also has the implausible consequence that either creatures who live in worlds without bees, wasps, and fires cannot have stinging and burning sensations or that in these worlds creatures undergoing such sensations are misrepresenting what is going on in them.[11]

This is not my proposal, however. Bees and wasps sting. But so do whips, chemicals, and slaps in the face. The physical transactions occurring in these cases differ. What is common to all of them is that they cause pains with a distinctive felt character. This causal sense of the term 'sting' is a perfectly ordinary one (to be found in any dictionary). And, in my view, 'sting', as it is used in connection with pains, has a causal sense too. The primary difference between the application of 'stinging' to pains and the

[11]I owe these objections to Sydney Shoemaker.

application of 'stinging' to non-mental physical objects is that in the former case the context is intensional. A stinging pain in the leg is a pain which represents, under the appropriate mode of representation, that something in the leg is causing that very pain. This mode of representation is what is responsible, in my view, for the special, shared phenomenal character of stinging pains, and what distinguishes them phenomenally from other pains.[12]

A similar story can be told in the case of burning pains. Fires burn. But so do pulled muscles, and certain chemicals when they come into contact with the skin. In one standard sense of the term 'burn', physical particulars burn by causing the experience as of being burned (which is phenomenally distinctive). 'Burning', as it used in connection with pains, has a causal sense too, I claim, but again the context in which the term is applied is intensional. A burning pain is one which represents, under the relevant mode of representation, that some bodily region is causing that pain, where again the representational mode is what is responsible for the special phenomenal character shared by experiences as of being burned, but not by other phenomenally different experiences.

Pains, I conclude, like images, have representational content. Unlike images, however, they have bodily locations (in the representational sense I have elucidated).[13] So, although pains, in my view, are really constituted by physical processes in the head, it is also true to say that they can occur anywhere in the body.[14]

II

The "language of thought" hypothesis is an empirical hypothesis about how the representational contents of mental states are, in fact, encoded in the head. It is not an *a priori* philosophical analysis. So it is not intended to cover the contents of mental states of all actual and possible creatures. In its most general form, it concerns the coding of *all* actual mental contents. The basic thesis, stemming from the computer model of mind, is that such contents are encoded in symbol-structures in an inner language.

In the case of the so-called "propositional attitudes" – that is, those mental states like belief and desire whose contents are standardly expressed

[12]This proposal is developed further in Tye 1995, as is the account presented below of burning pains.

[13]I am not claiming here that pains always have precise locations or that it is metaphysically impossible for a creature to have a pain without a bodily location. But I deny that so-called "psychological pains", for example, pains of regret or embarrassment, lack any bodily locations. I think it plausible to hold that such states are labelled 'pains' because, like pains, people are aversive to them. But this usage of 'pain' is, I suggest, metaphorical or analogical. This is not to deny, of course, that real pains may have psychological causes. Embarrassment may certainly *cause* burning facial pain. See here Stephens and Graham 1987: 413.

[14]The constitution relation is weaker than the relation of identity. A can be constituted by B even though A and B differ in some of their modal properties. See here Tye 1992a.

in 'that' clauses – it has typically been supposed that the relevant symbol-structures are sentences.[15] In the case of pain, however, there are certain pieces of evidence which count against a purely sententialist view.

We know that, in visual perception, the retinal image is reconstructed in the visual cortex so that, in a quite literal sense, adjacent parts of the cortex represent adjacent parts of the retinal image. There is, then, an orderly topographic projection of the retinal image onto the brain. This has been established from experiments in which a recording electrode is placed inside the visual cortex. Greater neural activity is picked up by the electrode when light is shone onto a particular spot on the retina. Moving the electrode a little results in the continued registration of greater activity only if light is directed onto an adjacent part of the retina.

Topographic organization of this sort is also found in the somatosensory cortex. There is, for example, an orderly topographic representation of the surface of the human body which is dedicated to touch. Here adjacent regions of the body surface are projected onto adjacent regions of the cortex. Enhanced activity in one of the relevant cortical regions represents that the region of body surface projected onto it is being touched. Some relatively small portions of the body, e.g., the hands and face, provide input to more neurons than do some relatively large portions, e.g., the trunk. This is why when people are asked to say whether there are two separate points that are both being touched on their faces or just one, the smallest distance between the points at which both can be felt is much less than the smallest distance when the points are located on the trunk.

There are further representations of the human body in the somatosensory cortex which are similarly structured. In one of these, enhanced activity represents that the tissue in the associated body region is being damaged or harmed. It has been established that the experience of pain is associated with activity in the somatosensory cortex.[16]

The fact that the somatosensory cortex is topograpically organized and that it is the primary locus of pain raises doubts about the sentential view of pain. For sentences do not have the requisite map-like representational structure.[17]

Still, there is also some reason to suppose that pains *are* sententially structured. In particular, there is the fact that pain experiences are fine-grained: as I remarked earlier, not even analytically equivalent expressions can be safely substituted salva veritate within the context 'experiences that —'. In this respect, pain is like the propositional attitudes generally.

The way to reconcile these apparently conflicting strands of thought

[15] See here Fodor 1975.

[16] This is not to deny that other neural regions also play a role in some pain experiences. In particular there are pain pathways which terminate in both the posterior parietal cortex and the superior frontal cortex.

[17] For a discussion of the representational differences between sentences, pictures, and maps, see Tye 1991.

is, I suggest, to hold that pains have a *complex* representational structure, one component of which is sentential and another map-like. In my view, pains are patterns of active cells occurring in topographically structured 3-D arrays to which sentences are attached. Space limitations prevent me from unpacking this proposal properly in the present paper.[18] The basic idea is that pains represent in something like the way that maps represent which contain additional descriptive information for salient items ('treasure buried here', 'highest mountain on island'). In this respect, they are very like mental images, as I conceive them.[19]

This suggestion needs much further development, of course. But what it gives us is an alternative way of thinking about pains as representations, one which seems to me more promising than a purely sentential approach.

There is another very important feature of pain on which much more needs to be said: its phenomenal character. In what exactly does the phenomenal character of pain consist? What exactly is the relationship of phenomenal character to representational content? These are pressing questions. But for my present purposes, it does not matter what the answers are. For my concern, in this paper, is simply to make plausible the view that pains have representational content and to give some preliminary sketch of the kind of representations pains are. [20]

References

Block, N. 1983 "Mental Pictures and Cognitive Science", *Philosophical Review* 93, 499–542.

Fodor, J. 1975 *The Language of Thought*, New York: Thomas Crowell.

Jackson, F. 1977 *Perception*, Cambridge: Cambridge University Press.

Melzack, R. 1961 "The Perception of Pain",*Scientific American* 204.

Smart, J. 1959 "Sensations and Brain Processes", *Philosophical Review* 68, 141–156.

Stephens, L. and Graham, G. 1987 "Minding Your *P*'s and *Q*'s: Pain and Sensible Qualities", *Nous* 21, 395–406.

Tye, M. 1984 "The Adverbial Approach to Visual Experience", *Philosophical Review* 93, 195–225.

Tye, M. 1991 *The Imagery Debate*, Cambridge, MA: MIT Press.

Tye, M. 1992a "Naturalism and the Mental", *Mind* 101, 421–441.

Tye, M. 1992b "Visual Qualia and Visual Content", in T. Crane (ed.), *The Contents of Experience* Cambridge: Cambridge University Press, 158–176.

Tye, M. 1994 "Qualia, Content, and the Inverted Spectrum", *Nous* (forthcoming).

Tye, M. forthcoming 1995 "A Representational Theory of Pains and their Phenomenal Character" in J. Tomberlin (ed.), *Philosophical Perspectives* 9: *AI, Connectionism and Philosophical Psychology*, Atascadero, CA: Ridgeview.

[18] A full account is given in Tye 1995.

[19] See Tye 1991.

[20] For an extended discussion of the relationship between the phenomenal character of pains and their representational content, see Tye 1995.

Toward a New Theory of Content

George Bealer

Frege's puzzle has proven to be a highly recalcitrant puzzle about content: if two sentences arise from one another by substitution of co-referential proper names, how can the two sentences express different propositions?

Many people advocate a pragmatic solution according to which such sentences must have the same literal meaning, attributing apparent differences in meaning to pragmatic confusions. I will assume that this sort of response is unacceptable. A correlative puzzle is how co-referential proper names can fail to be intersubstitutable *salva veritate* in propositional-attitude contexts. These two puzzles may be thought of as instances of an underlying puzzle about the reference of 'that'-clauses: how can ⌜that $A(a)$⌝ and ⌜that $A(b)$⌝ refer to different propositions when the names ⌜a⌝ and ⌜b⌝ are co-referential?

Frege's solution, which is based on his theoretical distinction between *Sinn* and *Bedeutung*, has been undermined by the arguments of Donnellan and Kripke.[1] They argue that proper names do not have descriptive senses. But if names do not have descriptive senses, what could the sense of a name be? How could co-referential names have different senses? How could we have epistemic access to such senses? No satisfactory answer to these questions appears to be forthcoming.

To solve Frege's puzzle and related puzzles about content, one must have the right sort of background theory of intensional entities (properties, relations, and propositions). There are four main theories: the possible-worlds theory, the propositional-function theory, the propositional-complex theory, and the algebraic theory.

Elsewhere I have argued that only the last of these is satisfactory.[2] At the heart of many of the problems confronting the other three is the fact that they are reductionistic: each attempts to reduce intensional entities of one kind or another to extensional entities – either sets or extensional functions.[3] My view is that this extensional reductionism has hampered the solution to the indicated family of puzzles and that what is needed is a theory which treats intensional entities as irreducibly intensional. This is what the algebraic theory offers.

The purpose of this paper is to lay out the algebraic theory and then

[1] Donnellan 1970 and Kripke 1980.

[2] For a defense of this assumption, see Bealer and Mönnich 1989.

[3] Functions f and g are extensional if $\forall x\ (f(x) = g(x)) \rightarrow f = g$.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 179-192

to show how it can be implemented in new solutions to a variety of these puzzles about content.[4]

1 The Algebraic Approach

On the algebraic approach, no attempt is made to reduce properties, relations, and propositions. Intuitively obvious truths like the following are accepted at face value requiring no reductionistic explanation. The proposition that $A\&B$ is the conjunction of the proposition that A and the proposition that B. The proposition that not A is the negation of the proposition that A. The proposition that Fx is the result of predicating the property F-ness of x. The proposition that there exists an F is the result of existentially generalizing on the property F-ness. And so forth. Such examples serve to impart a firm intuitive grasp of the indicated logical operations – conjunction, negation, singular predication, existential generalization, and so forth. The aim of the algebraic approach is to systematize the behavior of properties, relations, and propositions (conceived as irreducible entities) with respect to these logical operations.

There is a direct line of development in algebraic logic from Boolean algebras, to transformation algebras, to polyadic and cylindric algebras, and finally to intensional algebras. A Boolean algebra is a structure $\langle D, \mathsf{disj}, \mathsf{conj}, \mathsf{neg}, F, T\rangle$.[5] D is a domain of entities which may be thought of as primitive and irreducible; disj and conj are binary operations which may be thought of as the logical operations of disjunction and conjunction, respectively. The operation neg is a unary operation which may be thought of as the logical operation of negation. F and T are distinguished elements of the domain which may be thought of as falsity and truth, respectively. The operations in a Boolean algebra must satisfy certain standard rules which may be thought of as codifying our intuitive understanding of the operations of disjunction, conjunction, and negation, respectively. Boolean algebras are extensional models of sentential logic: in the simplest case, D would be just the set of truth values $\{F, T\}$ and disj, conj, and neg would be the standard truth functions. Boolean algebras are also extensional models of certain artificial fragments of first-order predicate logic. Consider, for example, a fragment of the monadic predicate calculus in which every atomic formula contains the same variable (and in which there are no quantifiers or individual constants). The following Boolean algebra would be a standard model for this fragment: D would be the power set of some given non-empty set of objects; disj would be the set-theoretical operation of union; conj would be intersection; neg would be complementation; F would be the null set; and T would be D itself. (One usually thinks of Venn diagrams as pictorial representations of this sort of Boolean algebra.) Or consider a fragment of the n-adic predicate calculus in which every atomic formula consists of an n-ary predicate letter followed by n distinct variables always occurring in the same order (and in which there are no quantifiers and no individual constants). For example, when $n = 3$ we have molecular formulas like '$((Fuvw \vee Guvw)\&\neg Huvw)$'. The following Boolean algebra would be a standard model for this fragment: D would be the power set of the n^{th} Cartesian product of some antecedently given non-empty set of objects; disj would be the union operation; conj would be intersection; neg would be complementation; F would be the null set; T would be D.

To obtain an extensional model of first-order predicate calculus (without quantifiers and without individual constants) in which the indicated restriction on the variables is dropped, one considers algebras $\langle D, \mathsf{disj}, \mathsf{conj}, \mathsf{neg}, \tau, F, T\rangle$ which resemble Boolean algebras. The main difference is that there is a new element τ, and D has more structure.[6] In particular, for some antecedently given non-empty set d of entities, D is the union of the truth values $\{T, F\}$ and the set of n-ary relations-in-extension over d (for all $n \geq 1$). (That is, $D = \{T, F\} \cup \bigcup_{n\geq 1} \mathcal{P}(d^n)$.) And τ is a set of auxiliary logical operations intended to be semantical counterparts of syntactical operations such as repeating the same variable one or more times within a given formula and of changing around the order of the variables within a given formula. For example, τ would contain an operation conv which maps the relation-in-extension $\{xy : x \text{ loves } y\}$ to its converse $\{yx : x \text{ loves } y\}$; and τ would contain the operation reflex which maps the relation-in-extension $\{xy : x \text{ loves } y\}$ to its reflexivization $\{x : x \text{ loves } x\}$. To obtain an extensional model of the predicate calculus with quantifiers (but without individual constants), one considers structures $\langle D, \mathsf{disj}, \mathsf{conj}, \mathsf{neg}, \mathsf{exist}, \tau, F, T\rangle$ that are like the previous structures except that they contain an additional operation, exist.[7] This operation is to be thought of as the logical operation of existential generalization. For example, it takes a binary relation-in-extension (e.g., $\{xy : x \text{ loves } y\}$) to an appropriate unary relation-in-extension (e.g., $\{x : (\exists y)x \text{ loves } y\}$). All the above algebraic ideas are standard nowadays.

To obtain an *intensional* model for the predicate calculus (without individual constants), one considers closely related algebraic structures $\langle D, K, \mathsf{disj}, \mathsf{conj}, \mathsf{neg}, \mathsf{exist}, \tau, F, T\rangle$. Here the domain D is the union of denumerably many disjoint subdomains $D_{-1}, D_0, D_1, D_2, D_n, \ldots$. The subdomain D_{-1} is to be thought of as being made up of particulars; D_0, propositions; D_1, properties; D_2, binary relations-in-intension; D_n, n-ary relations-in-intension. The elements of D are to be thought of as primitive, irreducible items. The new element K is a set of *possible extensionalization functions.*

[4] For a more detailed exposition of this theory and for more thorough bibliographical references, see Bealer 1993.

[5] It is more common to write: $\langle D, +, ., -, 0, 1\rangle$. The notation in the text will be more perspicuous for present purposes.

[6] These structures $\langle D, \mathsf{disj}, \mathsf{conj}, \mathsf{neg}, \tau, F, T\rangle$ are closely related to Halmos's transformation algebras (Halmos 1962: 27f.). For related ideas, see Quine 1960.

[7] These structures are closely related to cylindric algebras (see Henkin *et al.* 1971) and polyadic algebras (see Halmos 1962). For similar approaches to algebraic models for the predicate calculus, see Quine 1960 and William Craig 1974.

Each extensionalization function $H \in K$ assigns to the elements of D an appropriate extension as follows: for each proposition x (i.e., for each $x \in D_0$), $H(x) = T$ or $H(x) = F$; for each property x (i.e., for each $x \in D_1$), $H(x)$ is a subset of D; for each n-ary relation-in-intension x (i.e., for each $x \in D_n$), $H(x)$ is a subset of the n^{th} Cartesian product of D; in the case of particulars x (i.e., $x \in D_{-1}$), let $H(x) = x$. Among the possible extensionalization functions in K there is a distinguished function G which is to be thought of as the *actual* extensionalization function; it tells us the actual extension of the elements of D. The operations conj, neg, and so forth in an intensional algebra behave in the expected way with respect to each extensionalization function $H \in K$. For example, for all x and y in D_0, $H(\mathsf{conj}(x, y)) = T$ iff $H(x) = T$ and $H(y) = T$. For all x in D_0, $H(\mathsf{neg}(x)) = T$ iff $H(x) = F$. And so forth. For ease of presentation I will hereafter write simply $\langle D, K, \tau \rangle$ with the understanding that D and K are as indicated and τ is an ordered set of operations including, in order, disj, conj, neg, exist, and those in τ. No harm is done if τ contains further operations in addition to those indicated; so this will be permitted. Finally, for convenience, F will be identified with the null set and T with the domain D. With these details in place one can say what it takes for one of these algebras $M = \langle D, K, \tau \rangle$ to be *intensional*: there are elements in some $D_i \subset D$, $i \geq 0$, which can have the same possible extension and nevertheless be distinct. That is, M is intensional iff, for some x and y in $D_i \subset D$, $i \geq 0$, and for some $H \in K$, $H(x) = H(y)$ and $x \neq y$. For example, if x and y are in D_0, perhaps $G(x) = G(y) = T$ but $x \neq y$.

These intensional algebras yield intensional models of the predicate calculus (without individual constants). An intensional interpretation is a function I that maps i-ary predicate letters to i-ary relations-in-intension. Relative to an intensional interpretation I and an intensional algebra M, it is easy to define an intensional valuation function V_{IM} which maps sentences of the predicate calculus (without individual constants) to relevant propositions in D. For example, $V_{IM}(\text{`}\neg(\exists x)Fx\text{'}) = \mathsf{neg}(\mathsf{exist}(I(\text{`}F\text{'})))$. A sentence $\ulcorner A \urcorner$ is true relative to I and M iff its actual extension $= T$. That is, $\mathrm{Tr}(\ulcorner A \urcorner)$ iff $G(V_{IM}(\ulcorner A \urcorner))$ is the truth value T.

So far, however, I have not indicated how intensional algebras can model the predicate calculus *with* individual constants. By 'individual constant' I mean variables with fixed assignments, Millian (or Russellian) proper names,[8] and intensional abstracts. Suppose that the notion of interpretation is extended so that I assigns to each variable a value in M's domain D and to each Millian (or Russellian) proper name a nominatum in D. Then, it would be desirable to be able to assign some proposition in D as the intensional value of open sentences $\ulcorner Fx \urcorner$. Similarly, suppose that $\ulcorner a \urcorner$ is a Millian (or Russellian) proper name. It would be desirable to be able to assign a propositional meaning to the sentence $\ulcorner Fa \urcorner$. Finally, suppose that the language

[8] By 'Millian (or Russellian) proper name' I mean a syntactically simple singular term that is not a variable and that has a rigid denotation and no connotation or sense.

is fitted-out with intensional abstracts.[9] For example, let the 'that'-clause $\ulcorner$that $(\exists x)Gx \urcorner$ be represented by the singular term $\ulcorner[(\exists x)Gx]\urcorner$. It would be desirable to be able to assign a proposition in D as the intensional value of sentences with forms like '$B[(\exists x)Gx]$' (the symbolic counterpart of, say, 'It is believed that something is green'). This threefold problem is solved by restricting ourselves to intensional algebras $M = \langle D, K, \tau \rangle$ in which τ contains an additional logical operation, namely, *singular predication* – pred_s, for short. The operation of singular predication behaves exactly as one would expect. For example, when singular predication is applied to a property and an item, the proposition that results is true iff the item is in the extension of the property. That is, for all $x \in D_1$ and $y \in D$, and for all extensionalisation functions $H \in K$, $H(\mathsf{pred}_s(x, y)) = T$ iff $y \in H(x)$. Using singular predication, one can then assign appropriate intensional values to the three cases: $V_{IM}(\text{`}Fx\text{'}) = \mathsf{pred}_s(I(\text{`}F\text{'}), I(\text{`}x\text{'}))$; $V_{IM}(\text{`}Fa\text{'}) = \mathsf{pred}_s(I(\text{`}F\text{'}), I(\text{`}a\text{'}))$, and $V_{IM}(\text{`}F[(\exists x)Gx]\text{'}) = \mathsf{pred}_s(I(\text{`}F\text{'}), \mathsf{exist}(I(\text{`}G\text{'})))$. Because intensional abstracts may be evaluated in this way, intensional algebras provide models of first-order intensional logic.

My solution to our family puzzles about content will depend on two further developments. The first concerns the kind of predication involved in certain descriptive propositions. The second concerns the distinction between Platonic and non-Platonic modes of presentation.

2 Descriptions

There are four leading theories of definite descriptions: Frege's, Russell's, Evans's, and Prior's.

1. Frege. On this theory $\ulcorner$the $F\urcorner$ is an ordinary singular term having a sense and often a reference. The term $\ulcorner$the $F\urcorner$ has the form $\ulcorner(\iota x)(Fx)\urcorner$, where $\ulcorner(\iota x)\urcorner$ is a unary operator which combines with a formula to yield a singular term. If there is a unique item satisfying the predicate $\ulcorner F\urcorner$, the singular term $\ulcorner$the $F\urcorner$ refers to it; otherwise, $\ulcorner$the $F\urcorner$ has no reference. Truth conditions are as follows:

   (a) if $\ulcorner$the $F\urcorner$ has a reference, $\ulcorner$The F Gs$\urcorner$ is true (false) iff $\ulcorner(\forall x)(Fx \rightarrow Gx)\urcorner$ is true (false);

   (b) otherwise, $\ulcorner$The F Gs$\urcorner$ is neither true nor false.

[9] An intensional abstract is a 'that'-clause or a gerundive (or infinitive) phrase. That is, a proposition abstract, a property abstract, or a relation abstract. Because λ–abstracts $\ulcorner(\lambda v)(\text{that}A)\urcorner$ denote propositional functions and because properties are not propositional functions, use of λ–abstracts to denote properties invites confusion. A better notation is $\ulcorner[v_1 \ldots v_n : A]\urcorner$ where $n \geq 0$. Thus, whereas $\ulcorner\{v_1 : A\}\urcorner$ denotes the set of things v_1 such that A, $\ulcorner[v_1 : A]\urcorner$ denotes the property of being a v_1 such that A. Whereas $\ulcorner\{v_1 \ldots v_n : A\}\urcorner$ denotes the relation-in-extension holding among $v_1 \ldots v_n$ such that A, $\ulcorner[v_1 \ldots v_n : A]\urcorner$ denotes the relation-in-intension holding among $v_1 \ldots v_n$ such that A. In the limiting case where $n = 0$, $\ulcorner[A]\urcorner$ denotes the proposition that A. For more on this sort of notation see Bealer 1979 and 1982.

Truth-value gaps are not essential to Frege's theory; to eliminate them, one need only revise clause (ii) as follows: if ⌜the F⌝ has no reference, ⌜The $F\,G$s⌝ is false. In my subsequent remarks I will adopt this revised theory for simplicity of exposition.

2. Russell. On this theory ⌜the F⌝ is an incomplete symbol, meaningful only in the context of a complete sentence. Sentences containing definite descriptions are mere abbreviations for (or transformations from) sentences containing no descriptions. For example, ⌜The $F\,G$s⌝ is an abbreviation for (transformation from)

$$\ulcorner(\exists x)Fx\,\&\,(\forall x)(\forall y)((Fx\&Fy)\rightarrow x=y)\,\&\,(\forall x)(Fx\rightarrow Gx)\urcorner.$$

3. Evans.[10] On this theory 'the x' is treated as a binary quantifier which combines with a pair of formulas to yield a new formula. For example, ⌜The $F\,G$s⌝ has the form ⌜[the x]$(Fx : Gx)$⌝. The truth conditions are Russellian.

4. Prior *et al.*[11] On analogy with ⌜some F⌝ and ⌜every F⌝, ⌜the F⌝ is treated as a restricted quantifier ⌜[the $x : Fx$]⌝ which combines with a formula to yield a new formula. For example, ⌜The $F\,G$s⌝ has the form ⌜[the $x : Fx$](Gx)⌝. The truth conditions are again Russellian.

Each of these four theories can easily be incorporated into the algebraic approach. I will illustrate how to do this in the case of Frege's theory. Consider intensional algebras in which the set τ contains a unary operator the (akin to the Frege-Church operator ι) which takes properties to properties thus: for all properties $u \in D_1$, all $H \in K$, and all items $w \in D$, $w \in H(\mathsf{the}\,(u))$ iff $H(u) = \{w\}$. The values of the are properties that may be thought of as "individual concepts". For example, the (F) may be thought of as the individual concept of being the F. Starting with the property of being G and the individual concept of being the F, how does one form the proposition that the $F\,G$s? This proposition is not the result of a singular predication. When the operation of singular predication is applied to the property of being G and the property of being the F – i.e., $\mathsf{pred}_s(G, \mathsf{the}(F))$ – the result is the proposition that the property of being the FGs. A very different proposition! The relation holding between the property of being G, the property of being the$F$, and the proposition that the $F\,G$s is therefore not singular predication but rather a quite distinct kind of predication, which may be called *descriptive predication* – pred_d, for short. This relation of descriptive predication is implicit in Frege's informal theory of senses: it is the relation holding between the sense of a predicate ⌜G⌝, the sense of a definite description ⌜the F⌝, and the sense of a sentence ⌜The $F\,G$s⌝.[12] To represent Frege's theory of definite descriptions algebraically, one merely need to restrict oneself to intensional algebras in which the set τ contains both the and pred_d, where pred_d behaves thus: for all $u, v \in D_1$, and all $H \in K$, $H(\mathsf{pred}_d(u, v)) = T$ iff $\emptyset \neq H(v) \subseteq H(u)$. So, for example, the proposition that the $F\,G$s $= \mathsf{pred}_d(G, \mathsf{the}(F))$. This proposition is true relative to $H \in K$ iff $\emptyset \neq H(\mathsf{the}(F)) \subseteq H(G)$. That is, relative to H, the proposition that the $F\,G$s is true iff there exists something that is the unique element in the extension of the property of being F and the extension of the property of being F is included in the extension of the property of being G.

[10] Evans 1977a and 1977b.

[11] Prior 1963. Paul Grice, Richard Sharvy, and Richard Montague also advocated versions of this theory.

The operation of descriptive predication is also used to form other sorts of descriptive propositions within a Fregean setting. For example, consider one of Stephen Neale's number-neutral descriptive propositions: the proposition that whoever shot Kennedy is crazy. Within a Fregean setting this proposition may be represented thus: $\mathsf{pred}_d(C, \mathsf{whe}(S))$, where whe is Neale's number-neutral description operation.[13] This operator takes the property of shooting Kennedy (i.e., S) as argument and gives as value the number-neutral descriptive property being whoever shot Kennedy (i.e., $\mathsf{whe}(S)$). Relative to a possible extensionalization function H, the proposition $\mathsf{pred}_d(C, \mathsf{whe}(S))$ is true iff the extension of $\mathsf{whe}(S)$ is a non-empty subset of the extension of C.

The point is that, in addition to various description operators – the, who, etc. – there is an operation of descriptive predication which combines predicative intensions and descriptive subject intensions to form descriptive propositions. In what follows, I will make use of this aspect of Frege's theory; more specifically, I will make use of intensional algebras in which the set τ contains the operation pred_d. In doing so, I do not wish to commit myself to Frege's theory of definite descriptions. I could instead adopt something more in the spirit of Russell, Evans, or of Prior. I pursue the Fregean option because it is so natural (and because it is of much historical interest).

3 Non-Platonic Modes of Presentation

I have noted that the domain D in an intensional algebra partitions into subdomains D_{-1}, D_0, D_1, D_2, ... We have been thinking of D_1 as consisting of properties. But we could instead think of it as consisting of modes of access or modes of presentation (*Arten des Gegebenseins*). Properties, which are

[12] If, instead, one were to formalize Frege's informal theory by identifying the sense of a predicate with a function whose arguments are individual concepts and whose values are propositions, the relation of descriptive predication would collapse into a special case of the relation of application of function to argument. This approach, however, exposes the informal theory of senses to the various flaws of the propositional-function theory. When the propositional-function thesis is divorced from Frege's informal theory, one gets the picture presented in the text.

[13] See Neale 1990a and 1990b. Neale's elegant treatment provides only truth conditions; it does not identify the propositions expressed by such sentences. This remaining task is what is accomplished by the technique being described in the text.

purely Platonic entities, are just one kind of mode of presentation. There are also certain "constructed" entities that present objects to us. For example, pictures do. Certain socially constructed entities also function as modes of presentation. Prominent among these are linguistic entities. Indeed, linguistic entities provide the only access most of us have to various historical figures – for example, Cicero. These entities have the important feature of being *public* entities shared by whole communities.

Historical naming trees (or causal naming chains) are one kind of linguistic entity which fulfill this role. For example, the 'Cicero'-historical naming tree provides us with access to Cicero. A closely related mode of access is our very practice of using 'Cicero' to name Cicero. Another is the name 'Cicero' itself. (Of course, names here must be understood not as mere phonological or orthographic types but as fine-grained entities individuated by the associated practices. E.g, just as our practice of using 'Cicero' to name the Illinois town differs from our practice of using 'Cicero' to name the orator, so the town's name, which is comparatively new, differs from the orator's name, which is much older.) Insofar as these linguistic entities (the tree, the practice, the name) provide us with access to Cicero, they count as modes of presentation of Cicero.[14]

I will now indicate how these three kinds of non-Platonic modes of presentation can lead to candidate solutions to our puzzles. (I should emphasize that these are not the only candidate solutions feasible within the present general framework.) Note that there is a natural one-one map from historical naming trees onto conventional naming practices (the tree may be thought of as the practice "spread out in history"), and there is a natural one-one map from conventional naming practices onto the associated names. Because there exist these natural correspondences, it will make little difference which kind is best – historical naming trees, conventional naming practices, or names themselves. For illustrative purposes, I will fill out the idea with naming practices playing the key role. It will be easy to see how the idea would go if one were to let names or naming trees play that role.

On the Kripke picture, a conventional naming practice typically consists of an initial act of baptism, with or without a baptized object actually present, together with an ongoing convention for using the name with the intention of referring to whatever it was that was referred to by previous uses of the name. Let $P_{\text{'Cicero'}}$ be our practice of using 'Cicero' to refer to Cicero, and let $P_{\text{'Tully'}}$ be our practice of using 'Tully' to refer to Tully. The acts of baptism which initiated $P_{\text{'Cicero'}}$ and $P_{\text{'Tully'}}$ were baptisms of one and the same object. Accordingly, these two practices provide us with two presentations of one and the same object. Insofar as $P_{\text{'Cicero'}}$ and $P_{\text{'Tully'}}$ present an object to us, there are intensional algebras in which they are elements of the subdomain of modes of presentation.

[14]In virtue of what do these modes of presentation present objects? There are variety of plausible answers, e.g., causal, historical, intentional. I need take no stand on which is best.

Because $P_{\text{'Cicero'}}$ and $P_{\text{'Tully'}}$ both present Cicero (= Tully), the extensionalization functions H in such intensional algebras behave accordingly: $H(P_{\text{'Cicero'}}) = \{\text{Cicero}\} = \{\text{Tully}\} = H(P_{\text{'Tully'}})$. In these intensional algebras, relevant logical operations would be defined for all modes of presentation – non-Platonic as well as Platonic. So, for example, the operation of descriptive predication pred_d may take as arguments, say, the property of being a person and $P_{\text{'Cicero'}}$. The result $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ would be a proposition. Likewise, for $\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})$. Note that these non-Platonic modes of presentation (as opposed to descriptive properties formed from them by means of the, whe, or some other description operator) are themselves the arguments in these descriptive predications.

Let us examine the features which these two propositions would have. Given that $P_{\text{'Cicero'}}$ and $P_{\text{'Tully'}}$ are distinct, $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ and $\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})$ would be distinct. Next, let us agree with essentialists like Kripke that every person is necessarily a person.[15] Given that $H(P_{\text{'Cicero'}}) = \{\text{Cicero}\} = \{\text{Tully}\} = H(P_{\text{'Tully'}})$, it follows that $H(\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})) = T$ and $H(\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})) = T$. Since this holds for all possible extensionalization functions H,[16] our two propositions $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ and $\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})$ would be necessarily true. That is, these two propositions would have the modal value that Kripke *et al.* would like to attribute to the proposition that Cicero is a person and the proposition that Tully is a person.[17]

Furthermore, our two propositions – $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ and $\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})$ – are distinct from all propositions expressible with the use of definite descriptions (with or without actuality operators). For example, $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ is distinct from each of the following: the proposition that the thing presented by our conventional naming practice $P_{\text{'Cicero'}}$ is a person; the proposition that the thing presented by this conventional naming practice is a person; the proposition that the thing actually named 'Cicero' is a person; and so forth. Finally, these propositions – $\mathsf{pred}_d(\text{being a person}, P_{\text{'Cicero'}})$ and $\mathsf{pred}_d(\text{being a person}, P_{\text{'Tully'}})$ – are not metalinguistic in the standard senses.[18] First, these propositions are distinct from all propositions expressible by sentences containing metalin-

[15]I.e., for all $x \in D$, if $x \in G(\text{being a person})$, then, for all extensionalization functions $H \in K, x \in H(\text{being a person})$, where G is the actual extensionalization function. "Serious actualists" deny that each person is necessarily a person; instead, they hold that each person is such that, necessarily, if he exists, he is a person. Accordingly, serious actualists would require: if $x \in G(\text{being a person})$, then, for all $H \in K$, if $x \in H(\text{existence})$, $x \in H(\text{being a person})$.

[16]I am taking it for granted that conventional naming practices are "rigid": for example, if there were a practice of using 'Cicero' to refer to someone other than Cicero, it would not be *our* practice (i.e., *this* very practice of using 'Cicero' to refer to *him*).

[17]Likewise, $\mathsf{pred}_d(\mathsf{pred}_d(\text{identity}, P_{\text{'Tully'}}), P_{\text{'Cicero'}})$ has the same modal value (i.e., necessity) that Kripke *et al.* attribute to the proposition that Cicero = Tully.

[18]This requirement is insisted upon in Burge 1978: 127 ff., Burge 1979: 97 and Schiffer 1987: 67 ff.

guistic vocabulary. Second, when someone (e.g., a child or an ill-educated adult) is thinking one of these propositions, there is no evident need for the person to be employing any relevant concepts from linguistic theory, e.g., the concept of a conventional naming practice. The two propositions are seamless; only in their logical analysis do metalinguistic modes of presentation appear.[19]

Let me sum up. We have been seeking a theory of propositions in which, for example, the proposition that Cicero is a person and the proposition that Tully is a person should have the following features. They should be distinct from each other. They should be necessarily true. They should not be the sort of proposition expressible by sentences containing definite descriptions. Finally, they should not be metalinguistic in the standard senses. Propositions such as pred_d(being a person, $P_{\text{`Cicero'}}$) and $pred_d$(being a person, $P_{\text{`Tully'}}$) have all these features. Thus, they are promising candidates for the sort of propositions which have been eluding us.

In a wholly analogous way fine-grained names and historical naming trees could be incorporated into intensional algebras as non-Platonic modes of presentation; doing so would yield other candidate propositions with the desired characteristics. Besides these three proposals – practices, names, naming trees – there are others based on other candidate types of non-Platonic modes of presentation. It would be premature to declare any one of these proposals to be best; rather, one should canvass the full range of proposals and let the data determine the best. Nevertheless, because this general approach provides such a rich array of finely discriminated propositions, my conjecture is that at least one of these proposals provides a formally adequate solution to our family of puzzles. For the remainder of the paper I will assume that this conjecture is correct.[20]

Notational convention On each proposal I have considered, there is a regular connection between expressions and associated non-Platonic modes of presentation – for example, between 'Cicero' and our conventional linguistic practice $P_{\text{`Cicero'}}$. Suppose that on the proposal that validates my conjecture (just stated) – one of the above three proposals or some further proposal – there is a regular connection like this. In this case, the following notational convention may be introduced: if e is an expression and m is the non-Platonic mode of presentation to which e bears the indicated regular connection, then m will be denoted by the expression that results from enclosing e in double quotation marks. So, for example, "Cicero" would be our

[19]Thus, although no metalinguistic *sentences* express these propositions, there are metalinguistic *descriptions* – i.e., 'pred_d(being a person, $P_{\text{`Cicero'}}$)' and '$\mathsf{pred}_d$(being a person, $P_{\text{`Tully'}}$)' – which provide correct logical analyses of them.

[20]Note that this conjecture does not take a stand on how to formulate the semantics for the sentences in our problem area. My goal has simply been to show how to provide a rich enough array of propositions to underwrite a formal semantical treatment of our puzzles.

conventional linguistic practice $P_{\text{`Cicero'}}$, or some other non-Platonic mode of presentation, depending on which candidate proposal is correct.

I have been discussing non-Platonic modes of presentation that have regular connections with names. But there are also non-Platonic modes of presentation that have regular connections with predicates (e.g., our conventional linguistic practices of using a given predicate to express a relevant property or relation; intentional predicating trees; etc.) The above notational convention is also intended to apply to predicates. So, for example, '"chew"' and '"masticate"' are to denote relevant non-Platonic modes of presentation.

4 Some Applications

These ideas put us in a position to suggest candidate solutions to a variety of further puzzles about content.

1. Kripke's puzzle about Pierre's beliefs.[21] Upon seeing a picture of a pretty-looking city labeled 'Londres', Pierre states 'Londres est jolie'. Later, after living in an unattractive section of London, he states 'London is not pretty'. But it does not seem that Pierre believes a contradiction. Why not? The solution is that on the first occasion the proposition he asserts and believes on the first occasion is pred_d(being pretty, "Londres") whereas the proposition he asserts and believes on the second occasion is $\mathsf{neg}(\mathsf{pred}_d$(being pretty, "London")).[22] These two propositions are not in contradiction, for pred_d(being pretty, "London") $\neq$ pred_d(being pretty, "Londres"). This is so because "London" $\neq$ "Londres".[23]

2. The traditional problem of negative existentials: how can a sentence like 'Pegasus does not exist' express a true proposition given that 'Pegasus' lacks both a reference and a descriptive sense? The proposed solution is that the sentence expresses (something like) the true proposition $\mathsf{neg}(\mathsf{pred}_d$(existing, "Pegasus")).

3. An analogue of Frege's puzzle involving predicates rather than names. The problem is to explain why, e.g., 'There exists something that chews

[21]Kripke 1979.

[22]Or he might mean – and believe – $\mathsf{neg}(\mathsf{pred}_s$(being pretty, London)). This proposition does not contradict the one he stated and believed originally. After all, "London" $\neq$ "Londres"; moreover, singular predications and descriptive predications are always distinct.

[23]Kripke poses a second puzzle. Peter makes a certain pair of apparently contradictory assertions about a musician Polish Prime Minister named 'Paderewski'. I am inclined to the view that Peter's assertions and beliefs really are contradictory and that what the example shows is that a person's rationality is determined, not by *all* of the person's beliefs, but only by a certain privileged subset of them. People who disagree with this assessment seem to me to be focusing on *auxiliary* beliefs that Peter must have had rather than on the two beliefs that Peter actually articulated when he sincerely asserted the relevant sentences with the intention of speaking literally. Suppose, however, that I am mistaken and that Peter's two beliefs are *not* contradictory. In this case, the framework in the text could be extended in obvious ways to provide the relevant propositions.

and does not masticate' and 'There exists something that masticates and does not chew' intuitively do not mean the same thing even though chewing is the same property as masticating. A candidate solution is to invoke distinct non-Platonic modes of presentation of this property (e.g., "chew" and "masticate") to explain the indicated difference in meaning. For example, perhaps 'There exists something that chews and does not masticate' means exist(conj("chew", neg("masticate"))) whereas 'There exists something that masticates and does not chew' means exist(conj("masticate", neg("chew"))).[24]

4. Consider an English speaker who is familiar with the name 'Phosphorus' but not 'Hesperus'. Suppose that by pure chance the person makes the stipulation that 'Hesperus' is hereafter to be another name for Phosphorus. By an adaptation of Kripke's meter-stick example, Kripke would be committed to holding that the person would know something *a priori*. But what? Would the person know *a priori* that Hesperus = Phosphorus? That is, would the person know *a priori* the oft discussed necessity? If so, Kripke's famous doctrine that this necessity is essentially *a posteriori* would collapse. But we have on hand tools for solving this problem. The familiar *a posteriori* necessity is a descriptive prediction formed from one of our *standing* non-Platonic modes of presentation. By contrast, the necessity which the person knows *a priori* is a descriptive prediction formed instead from a *new* non-Platonic mode of presentation associated with the person's stipulation. Because these non-Platonic modes are distinct, so are the two propositions. So goes the solution. I believe that something like this is required to solve the problem and, more generally, to reconcile Kripke's scientific essentialism with the sort of *a priori* knowledge associated with stipulative definitions.

5. The foregoing ideas might also provide raw materials for treating demonstratives. Suppose that I see an object x directly in front of me and simultaneously see the same object x (without realizing that it is the same) through a complicated lens set-up on my left. Suppose that, while glancing straight ahead, I sincerely assert 'This is a pencil' with an intention of speaking literally. Intuitively, I would mean – and believe – something different from what I would mean – and believe – if, while glancing to the left, I sincerely assert 'That is a pencil'. What is the difference? The above theory provides a range of promising answers. The simplest is this. When I assert 'This is a pencil', the proposition I mean and believe is pred_d(being a pencil, "this"), and when I assert 'That is a pencil', the proposition I mean and believe is pred_d(being a pencil, "that"). The idea is that "this" and "that" are

[24] Analogously, perhaps the non-Platonic mode of presentation "arthritis" is responsible for the oblique use of 'arthritis' discussed in Burge 1979.

limiting cases of the sorts of non-Platonic modes of presentation discussed above: for example, perhaps "this" = my act of referring to x by uttering 'this' on the indicated occasion, and perhaps "that" = my act of referring to x by uttering 'that' on the indicated occasion. In this case "this" $\neq$ "that", and therefore, pred_d(being a pencil, "this") $\neq$ pred_d(being a pencil, "that"). Perhaps this is the intensional distinction we are seeking. Now although this idea cannot be the whole story (e.g., it does not deal with phenomena such as pronoun anaphora descending from an initial use of a demonstrative), it might be a first step toward a successful treatment of demonstratives.

5 Conclusion

The foregoing is really only the outline of a theory. No doubt there are problems, and the theory will need to be modified in various ways. But I hope these ideas make it plausible that, despite recent doubts, a theory of properties, relations, and propositions can provide a promising general framework for the theory of content.[25]

[25] I wish to extend my warm thanks to Paul Hovda for his expert help in readying this paper for publication.

References

Bealer, G. 1979 "Theories of Properties, Relations, and Propositions", *The Journal of Philosophy* 76, 643–648.

Bealer, G. 1982 *Quality and Concept*, Oxford: Oxford University Press.

Bealer G. and Mönnich U. 1989 "Property Theories", *Handbook of Philosophical Logic* volume 4, Dordrecht: Kluwer, pp. 133–251.

Bealer, G. 1993 "A Solution to Frege's Puzzle", *Philosophical Perspectives* 7, J. Tomberlin (ed.), Atascadero, CA: Ridgeview Press, pp. 17–61.

Burge, T. 1978 "Belief and Synonymy", *The Journal of Philosophy* 75, 119–39.

Burge, T. 1979 "Individualism and the Mental", *Midwest Studies in Philosophy* 4, 73–122.

Craig, W. 1974 *Logic In Algebraic Form*, Amsterdam: North-Holland.

Donnellan, K. 1970 "Proper Names and Identifying Descriptions", *Synthese* 21, 335-338.

Evans, G. 1977a "Pronouns, Quantifiers, and Relative Clauses (I)", *Canadian Journal of Philosophy* 7, 467–536.

Evans, G. 1977b "Pronouns, Quantifiers, and Relative Clauses (II)", *Canadian Journal of Philosophy* 7, 777–97

Halmos, P. 1962, *Algebraic Logic*, New York: Chelsea.

Henkin, L., D. Monk, and A. Tarski 1971 *Cylindric Algebras* Part I, Amsterdam: North Holland.

Kripke, S. 1979 "A Puzzle About Belief", in *Meaning and Use, Papers Presented at the Second Jerusalem Philosophical Encounter*, ed. A. Margalit, Dordrecht: Reidel, pp. 239–83.

Kripke, S. 1980 *Naming and Necessity*, Cambridge, MA: Harvard University Press.

Neale, S. 1990a *Descriptions*, Cambridge, MA: MIT Press.

Neale, S. 1990b "Descriptive Pronouns and Donkey Anaphora", *The Journal of Philosophy* 87, 113–50.

Prior, A. 1963 "Is the Concept of Referential Opacity Really Necessary?" *Acta Philosophica Fennica* 16, pp. 189-98.

Quine, W. V. O. 1960 "Variables Explained Away", *Proceedings of the American Philosophical Society* 104, pp. 343-47.

Schiffer, S. 1987 *Remnants of Meaning*, Cambridge: MIT Press, 1987.

On Nonsense on Reference

Herbert Hochberg

1 Searle's Unsatisfied Intentions: Or How Terminology Replaces Ontology

In 1910-11, G. E. Moore took a belief to be essentially connected to a fact, whose existence provided its truth ground. He suggestively called attention to the transparency of the connection by holding that if the blanks in

(M) 'the belief that... refers to the fact that...'

are filled by tokens of the same sentence the result is an obvious or necessary truth. In the 1950s, Gustav Bergmann, separating particular acts of belief from their contents, which he took to be properties exemplified by (particulars in) acts, claimed that his version of (M), 'the thought that... *means* the fact that...' was an *analytic* pattern. In recent years, Searle has repeated the theme by claiming that a belief is intrinsically or internally a representation of its conditions of satisfaction.

Searle is rightly concerned, as was Russell long ago, to avoid a regress resulting from introducing a further mental act or state or agent to make the connection. This is one reason he speaks of the connection being intrinsic or internal. But he provides no resolution of basic philosophical problems involved in talk of an intrinsic connection, for he provides no analysis or ontological assay of the fact that intentional states and contents intrinsically represent conditions.

One problem concerns the implicit appeal to propositional entities. Suppose John, Peter and Saul believe that-p. The belief that-p is common to them. But, what is the belief and what is it for it to be common to various individuals or intentional states? Moore, seeking to avoid propositions, implicitly treated the belief as a universal property of individual acts of belief. Searle talks of representational contents and propositional contents as if he recognizes propositional type entities or content properties. But, he also speaks of intentional states being "realized in the neurophysiology of the brain" and of propositions expressing contents. His symbolic representation of an intentional state as 'Believe (It is raining)' raises questions about the representational roles of both the parenthetical sentence and the juxtaposition of it and the term 'believe'.

Bergmann took an intentional state or mental act to be a basic particular exemplifying two properties. One property, which he called a "thought",

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 193-205

supplied the content; the other, a generic property, played the role of what Searle calls a "mode", determining the state to be one of belief. Bergmann's mental acts are thus conjunctive facts consisting of a particular exemplifying two properties. It is then clear in what sense a content is intrinsic to an act. It is also clear that the content is intrinsically connected to a fact in a different sense of 'intrinsic'. This latter sense is based on the content property standing in a basic relation or nexus to such a fact *and* the supposed analytic or logical nature of the connection. On Searle's account nothing is clear or spelled out, and he even hops from talking of 'intrinsic' in the sense of a content's being intrinsic to a state to speaking of 'intrinsic' in the sense of a condition's connection to a content.

Bergmann's explicit analysis is helpful for understanding Searle's implicit view. When Searle says that a state is composed of a content and a mode, he is really talking of a *kind* of state, and not of John's particular state, as Moore spoke of the belief that lions exist. The complex expression 'Believe(it is raining)' thus represents a complex property, and intentional states, as kinds, are common properties. When Searle speaks of such a state being "realized" he is taking it to be a property of something, as one may say a particular realizes a property by being an instance of it. The particulars, for Searle, are brains, and intentional states, as properties, are realized in individual brains, John's, Peter's, etc. Thus, it becomes clear what it means for a content to be intrinsic to an intentional state. But it is also clear that Searle's ontology unconsciously embraces content properties or propositions as entities.

Searle believes he resolves the problem about propositional entities by distinguishing the ontological issues from the logical issues involved and invoking a simple form of materialism, reminiscent of Feigl's mid-century physicalism, to handle the former. But taking intentional states to be macro-brain states fails to resolve the basic problem. For diverse brain-states, John's and Peter's, will be the same *kind* of intentional state, and it is in virtue of being of such a kind that an individual state has the content it has. Searle does not escape the recognition, implicit as it may be, of propositional entities taken as content properties. Nor does he escape questions about how such properties, as realized in a brain, are intrinsically connected to conditions. Searle reproduces the basic features of the Moore-Bergmann analysis, with brain-states replacing mental acts. His materialism solves nothing; it merely introduces further problems.

Matters are not helped by Searle's acknowledging propositions as being "only common features" of intentional states. The problem remains of connecting such common features or properties to conditions. Searle's use of the term "only" suggests a summary dismissal of such properties as entities. However, they can be so dismissed only if the intentional states they characterize are not connected to conditions in virtue of being so characterized, but are so characterized in virtue of being so connected. Yet, for Searle, it is because a propositional content is intrinsic to an intentional state that such a

state is, in turn, intrinsically connected to its conditions. Thus propositions are not "only" common features – they link intentional states to conditions via a purportedly intrinsic relation between propositional entities (contents) and conditions, even when such conditions do not exist.

Moore spoke of a belief being true when the fact it referred to "existed" and false when such a fact had "no being" or "was not". He self-consciously took this to be an unavoidable manner of speaking, not a commitment to non-existent facts. Searle follows Moore by speaking of "representing truth conditions", of "features of the... state of affairs represented" and of "truth conditions" that do not "hold". Thus, if his intentional relation – x has as its conditions of satisfaction y – intrinsically connects a belief to its conditions, he must either account for x's having y as conditions, when the latter do *not* exist, or, if x is not connected to a non-existent y, explain why the conditions are *its* conditions.

Searle writes that he "can be in the Intentional state without the object or state of affairs that the Intentional state is 'directed at' even existing"[1] and indicates the kind of solution he does not want:

> The fact that our statements may fail to be true because of reference failure no longer inclines us to suppose that we must erect a Meinongian entity for such statements to be about... An Intentional state has a representative content, but it is not about or directed at its representative content. Part of the difficulty here derives from "about", which has both an extensional and an intensional-with-an-s-reading. In one sense (the intensional-with-an-s), the statement or belief that the King of France is bald is about the King of France, but in that sense it does not follow that there is some object which they are about. In another sense (the extensional) there is no object which they are about because there is no King of France. On my account it is crucial to distinguish between the content of a belief (i. e., a proposition) and the objects of a belief (i.e. the ordinary objects).[2]

What has gone wrong is clear. Searle thinks one who recognizes "Meinongian entities" does so by mistakenly taking the representative content of an intentional state to be what the state is about. And, of course, such a state is not *about* its content. Rather, it is about what it is about in virtue of the content it has. So, it can be about something, in the sense that it has a content, *and* not about something, if there is no intended object. The solution is facile, but not viable; "Meinongian" questions remain about the content's representation of its conditions.

Searle's particular states represent what they do because they have contents that represent. It is verbal juggling to get us to concentrate on the state and tell us that the state represents in that it has an intrinsic content, and not in that something is represented. Searle must add, as he

[1] Searle 1983: 4.
[2] Searle 1983: 17.

does at places, that the content intrinsically represents. But, then, whatever Meinong's view involves, his ghost haunts Searle's non-solution. In that Searle's relation – x has as conditions of satisfaction y – intrinsically connects contents (as well as states) to conditions, he faces Moore's problem: How can such a relation obtain between a content and a non-existent condition?

The problem is not dispelled by Searle's invoking a further dichotomy, that of product-process, and taking the "ambiguity" it involves (regarding the requirement and the thing required) to mislead those who raise the question. Ironically, back in 1910, Moore sought to reassure his listener by an anology between 'refers to' and 'imagines'. As one imagines what is not and as a process requires what is not yet, so one unproblematically refers to what does not exist. Searle's analogy, no more than Moore's, exorcises Meinong's ghost.

Nor does it help to recite that intentionality-with-a-t differs from intensionality-with-an-s and contents from intentional objects, their *extensions*. To repeat that an intentional state can represent conditions of satisfaction, since it can have a content, without representing anything, because that is what it is to be an intentional-with-a-t state, is to restate the problem, not solve it. Searle implicitly uses a representation function that, for contents as arguments, yields conditions as values. This forces the recognition of non-obtaining facts. To deny that, as he does, is to implicitly do what Reinhardt Grossmann explicitly, but incoherently, does: take intentional connections to be relations that need not relate–to be one-term relations in certain cases. Intentionality-with-a-t allows an intentional state or content to represent, even though it fails to relate.

Searle's two-fold use of 'intrinsic', as characterizing a content's connection to an intentional state and the connection that both particular intentional states and contents have to their conditions of satisfaction, highlights his failure to resolve a third problem. Bergmann, like Moore, recognized a basic connection between the belief that-p (as well as particular acts of belief) and the fact that-p. He also took this intentional connection between a content property and its truth ground to be analytic or necessary and the truth ground to be an actual or a possible fact. To substantiate his claim he undertook an extensive reexamination of the concept of analyticity and the nature of logical truth. Moore simply claimed he was acquainted with reference, as a fundamental relation connecting beliefs to facts. Searle raises a question as to what "exactly" is the relation "between Intentional states and the objects and states of affairs that they are in some sense about or directed at?"[3] But he does not answer it. As Searle avoids the ontological implications of taking intentional states to be intrinsically "directed at" states of affairs, by talking of macro-brain states, extensionality and intentionality-with-a-t, he resolves the problem about the relation itself by labelling it 'intrinsic'.

[3]Searle 1983: 4.

2 Geach on Russell and Mental Acts: Or How to Twiddle with Concepts

Geach introduces what he calls a "twiddle" operator, §(·). He literally speaks of a linguistic device for forming predicates from other predicates. What he actually does is introduce two functions. One takes predicates as arguments and yields twiddle predicates as values, and its use implicitly assumes three things: (1) that each twiddle predicate stands for a relation or property; (2) that the relation or property represented is a constituent of mental acts that relate (or characterize) what Geach calls "Ideas"; and (3) that such a relation or property, in turn, represents a relation or property. He thus has a second function that yields twiddle relations or properties as values for relations or properties as arguments.

Early in his book Geach tells us that "an Idea" is "the exercise of a concept in judgment"[4] and that Ideas correspond to concepts. So there is James' Idea of a cat as well as a concept – the concept *cat*. Allow Geach the concepts *cat* and *mat* and corresponding Ideas so that we may consider what he calls his "theory". Representing the two Ideas by the signs '**a**' and '**b**', let a subject S be said to judge that a cat is on a mat when S has an act of judgment consisting of S's Idea **a** standing in the relation §(**is on**) to S's Idea **b**. In addition to the postulates involved in the interpretation of twiddle predicates, Geach must also postulate that a fact, S's mental act consisting of **a** standing in §(**is on**) to **b**, represents the state of affairs that a cat is on a mat. Geach must not only link Ideas to concepts and twiddle relations to their correlates but judgment facts, consisting of Ideas in twiddle relations, to their truth grounds.

The need to link acts of judgment to truth grounds raises the same problem that Searle faces, and Geach avoids it by ignoring it. It is easy for him to do so, since he also ignores the familiar problem of universals. Taking a traditional term for properties and relations, i.e. 'concepts', Geach crosses the obvious verbal bridge provided and concerns himself with subjects *having* and *exercising* concepts, while speaking only in passing of objects having properties and standing in relations. But though we are told that concepts are exercised and had, it is never clear just what it is that is had or exercised or how all this is connected to objects having properties and standing in relations. Geach goes through the obligatory ordinary language exercises concerning when S is said to have or exercise a concept, but such explorations of linguistic use do no philosophical work. He speaks of things and ideas being related, yet he never asks what grounds the truth of a judgment nor specifies an answer.

Geach is not wholly oblivious to the problems involved, and he later seeks to explain his two basic notions, that of an Idea and that of the twiddle operator. He does so by developing an idea that he finds "rather obscurely"

[4]Geach 1971: 58.

in the early Wittgenstein.[5] Geach does not note that Russell also found the idea to be there, "shortly" rather than "obscurely". Russell, in his typically generous way, not only acknowledged his debt but proceeded, in the early 1920s, to develop Wittgenstein's idea into a new analysis of judgment facts, an analysis Geach essentially repeats.[6] What Geach does is "define" what it is for an Idea to stand to another Idea in a twiddle relation. It is to have a mental act that consists of concatenated utterances of words spoken "in the heart" – mental utterances.[7] To prepare the reader for the expression "in the heart", he sets out his version of Wittgenstein's Tractarian distinction between a sign and a symbol and makes use of Wittgenstein's "psychical constituents", from a letter to Russell. Geach avoids the ontological problems posed by such moves by taking a concept, like the concept cat, to be what is expressed by an expression, such as 'a cat'. To know what is expressed is to know how to use the expression. So disappear the classical problems, about universals, concepts and facts, into the web of ordinary language. Were Geach to face the issues, he would have to acknowledge mental acts, as judgment facts, and the facts such judgments purportedly represent, along with their consituents.

In the early chapters of his book Geach criticises Russell's Multiple Relation Theory of judgment and offers a first version of his own view as a revision of Russell's theory. One criticism has to do with the problem of the order of terms. Geach mentions that Russell noted the problem but didn't do much about it. Ironically, one of the main themes of the 1913 *Theory of Knowledge* manuscript that Russell abandoned was a detailed and systematic attempt to analyze relational order in facts, not only in judgment facts. A second criticism has to do with Russell's taking relations as terms in judgment facts. How can a relation, like *loves*, connect its own terms, Cassio and Desdemona, while functioning as a term for another relation, *believes*, in a single fact? This was already raised by Wittgenstein prior to 1914 and is discussed as an unresolved problem by Russell in the Logical Atomism lectures of 1918. It was a reason he eventually abandoned the theory by 1919, but Geach does not mention the Logical Atomism lectures in his discussion. A third criticism Geach makes stems from his arbitrary postulate that to judge that a stands in R to b is to judge that b stands in the converse of R to a. Again it is ironic that in the 1913 manuscript Russell holds that there is only one relation involved: there are not two facts, one having R as a constituent, the other the converse of R. Geach thinks a problem of ordering is introduced by converse relations, but this can only arise if there are two relations. And, even if Russell allowed for converses of relations, his 1913 analysis of order allows for an easy solution to such simple problems of order. That aside, Geach presents no argument for the view that a judgment that a has R to b is also a judgment that b has the converse of R to a, as he presents no argument for the view that acknowledging R involves acknowledging the converse of R. Perhaps this is due to his thinking in terms of predicates, rather than in terms of relations and properties. In a similar way, Geach proclaims, without argument, that to have a concept is to have a correlated negative concept. Such a claim raises a number of questions that we cannot and need not take up here, but we can note that what Geach means is that to be able to "exercise" the one concept is to have the capacity to exercise the other.

[5] Geach 1971: 101.
[6] Russell 1922: xix-xx.
[7] Geach 1971: 99-100.

Geach's theory is a verbose variation of Russell's 1925 adaptation of Wittgenstein's Tractarian views. Russell takes there to occur inner items of thought that stand in a certain relation. The inner items represent objects, properties, and relations, while the relation they stand in, "predication", represents exemplification. When the latter relation holds among the objects (particulars, properties and relations) represented ("meant"), the belief is true.[8] Wittgenstein sought to avoid the outright commitment to the possible facts that such inner complexes could be taken to represent, by packing the possibilities into the relevant "logical forms" of the represented and representing items. Russell sees no problem, since the representative role of the complex, the thought, is determined by the representative roles of its constituents. Geach, by transforming acts of judgment into concatenations of utterances made "in the heart", repeats Russell's pattern of analysis and follows Russell in overlooking the need to coordinate judgment facts to truth grounds, in addition to coordinating their respective constituents.

3 Davidson's Satisfiers: Or How to Trivialize Tarski, Truth and Meaning

Davidson seeks to dispense with meanings by appealing to truth conditions. Russell and Wittgenstein also sought to avoid classical propositions or content properties, as intermediaries between judgments or, for simplicity, atomic sentences and facts, as truth conditions. Thus, an atomic judgment or sentence was taken to be connected to a fact via their respective consitituents. Davidson echoes the move of the logical atomists by construing 'S knows the meaning of "Fa"' in terms of

[8] Whitehead and Russell 1950: 662. The use of concatenations of tokens, as Geach does, also follows Russell: Whitehead and Russell 1950: 661. It is also worth recalling that in the 1918 lectures Russell notes that one need not employ relational predicates but could take a relation among the subject terms to represent a relation among the objects the terms represent. Unlike Wilfrid Sellars, who took this to be significant in itself as well a key for the reading of the *Tractatus*, Russell saw no philosophical significance in doing so.

[8] For my purposes here we need not distinguish between talking about truth grounds for intentional states and for atomic sentences. Similarly we need not dwell on the sense in which any sentence, say a contradiction, can be said to be given a "truth condition", as opposed to a truth ground, by a T-sentence (or a theory to contain a truth predicate, as opposed to providing a theory of truth). For related discussions of truth predicates, truth grounds, truth conditions and truth theories, as well as of questions about complex and atomic thoughts and sentences see Hochberg 1992 and 1978.

(D) S knows that 'Fa' is true iff Fa.

This presents more than one problem, but the basic problem is Davidson's repetition of the pattern of the logical atomists without any philosophical substance.

To take the meaning to be given by specifying truth grounds is to take 'Fa' to be true if and only if there is a fact consisting of the object represented by the name exemplifying the property represented by the predicate. But, in so taking the ascription of a truth predicate, one recognizes the ontological implications of the use of the sentence 'Fa' in the formula 'S knows that "Fa" is true iff Fa' and thus accounts for 'Fa' having a truth value. A truth ground is not provided by echoing or transcribing the sentence to which a truth predicate is ascribed.

Consider '"Fa" is true iff Fa'. To avoid taking 'F' to represent an attribute, and the sentence a fact, Davidson talks in terms of 'F' ('Fx') being "satisfied by" or "true of" a. But to use these notions as he does is to do one of two pointless things. He either simply repeats or transcribes 'Fa', in place of providing an ontological assay of the relevant truth maker; or he treats satisfaction as a relation between a thing (sequence) and a predicate (open sentence) and takes the truth maker as a dyadic fact that has a linguistic item, a predicate or open sentence, as a constituent. Thus we arrive at the most absurd form of idealism, or "anti-realism" as it is now called, linguistic idealism. Aside from linguistic items, Davidson only recognizes the object a, in the above case. Yet the object cannot be the ground of truth, though it is a "satisfier" for Davidson.[9] For it can satisfy the predicates 'F' and 'G', but 'Fa' and 'Ga' cannot reasonably be said to have the same ground of truth, a, unless all such attributes of a are essential attributes.

Davidson's failure to deal with the basic problems faced by providing a theory of truth differs significantly from Searle's failure. Searle's problems result from his taking the connection between contents and conditions to be intrinsic, as Moore and Bergmann do, and he neither faces nor explicates what is involved in his use of 'intrinsic'. That is one failure. Implicitly using a pattern like (M), he requires the intrinsic connection to hold when the conditions do not exist. Thus he unknowingly acknowledges non-existent facts as well as propositions. That is his second failure. Davidson attempts to avoid facts altogether. Doing so he fails to provide a ground of truth for true atomic sentences. His failure does not stem from adopting an implicit connection between atomic sentences and non-existent truth grounds. One might say that Davidson, like Russell earlier, attempts to use an extrinsic connection to specify truth conditions, by using 'Fa', rather than '"Fa" represents Fa and Fa', as the right side of the relevant biconditional.

Russell, aware of the issues, attempted to avoid both non-obtaining facts and Fregean thoughts (propositions), while connecting atomic sentences

[9]Davidson 1969: 758.

(judgments) to truth grounds. To do that he introduced a part-whole relation between facts and their constituents and described the truth ground for an atomic sentence like 'Fa' as "the fact consisting of F and a". Not taking atomic sentences (or judgments) to refer to facts, or express propositions, he employed:

(T1) 'Fa' is true iff $\exists!(\imath p)(p$ consists of F and of $a)$.

By not invoking a reference relation between atomic sentences (beliefs) and facts, he avoided Moore's implicit commitment to non-existent facts. This commitment becomes transparent when (M) is replaced by:

(T2) 'Fa' is true iff 'Fa' refers to Fa & Fa (obtains).

This pattern is resurrected by Searle and Geach, if we provide them with an explicit ontology. By contrast, Russell's (T1) is Tarski-like in that the semantic term 'refers to' does not occur to the right of the biconditional. Rather, the predicate 'consists of' that occurs there represents a relation between a fact and its constituents. But (T1) is a far cry from Davidson's trivialization of Tarski, truth, and Russell that is embodied in (D).

When Russell adopted his Wittgensteinian account of judgment, he followed Wittgenstein in claiming that the meanings of the constituents of an atomic judgment (sentence) determined the meaning of the sentence. Wittgenstein sought to avoid possible facts by taking logical forms or internal properties of objects and signs to connect atomic sentences to atomic facts. Intrinsic natures or formal properties of objects, properties and relations replaced possible facts. The possibility that R connects a and b, in the appropriate order in an atomic fact, was replaced by a, b and R having formal or essential properties. Russell abandoned the earlier pattern of (T1) to take the terms of an atomic sentence (judgment) to determine its truth ground, without recognizing either Wittgenstein's essential attributes or that a sentence, as a complex, required interpretation. Thus, in the 1920s he overlooked a problem he had earlier solved.

4 Kripke's Intention to Refer: Or How One Issue Replaced Another

To get at the philosophical issues involved in Russell's distinction between reference and description, we need not bother with Madagascar as a part of Africa or the unknown Ramses VIII. A simple case will do. Draw a white patch on a blackboard, then label it or baptize it 'β' and consider three situations:

(1) The use of a token of 'β' in saying or thinking to oneself, or saying "in one's heart", 'Call this "β"!'

(2) A subsequent thought expressed by 'This is β', which implicitly involves remembering and using a token of the name.

(3) Looking at some other object and asserting or thinking or saying to oneself 'This is not β'.

The three situations pose several problems. First, what occurs – what is the assay of the facts that obtain – when β is labelled and referred to by a token of a demonstrative, as in (1)? Second, does one refer to the patch in the same sense, by using the label 'β', when one sees it again, as in (2)? Third, what does the relational predicate 'refers to' itself represent, as just used? Fourth, does one refer to the patch in the same sense in (3), when the object is not observed? If so, why? If not, why not? These are philosophical problems posed by reference and intentions to refer. Such problems do not concern those who search for features common to cases where we would normally say a reference was made, in order to distill conditions (necessary or sufficient or necessary and sufficient) for saying a reference was made and to what. As in the familiar philosophical problems of perception, one set of questions concerns whether we really do what we ordinarilly say we do. Do we, as some say we do, refer to the magic sword Excalibur, the legendary King Arthur, and Aristotle?

Return to the case of the patch on the board. The patch is the focus of attention when a token of 'β' or of 'this' is uttered. Is this what it is for such a token to refer to or represent the object? Or does a unique intentional relation come into play? It does not help to speak of the intentional state being *intrinsically* representational, unless that is spelled out. Introspecting does not help either, while ceremonial incantations, like 'Call this "β"!' or 'I baptize thee "β"', add nothing. They only serve to indicate, to oneself or listeners, later uses of tokens of 'β'. But an audience (or a community) is irrelevant to the philosophical issues.

There is more than attending and uttering. If we try to attend to one object and refer to another, by uttering a demonstrative, without an accompanying description, we find that we cannot do so. But we do not simply discover that we cannot presently focus and not focus on the same object. We also find we can focus on an object and utter a demonstrative without referring to it. Such simple, phenomenologically grounded facts point to our awareness of a basic reference relation that holds between a token and its referent. Initially attending to the patch, I could simply have called it 'β', without invoking any rule or baptismal pronouncement. That aside, assume, having referred to the patch, I attend to it again, saying 'This is β' or 'β is white'. Do such tokens of 'β' represent the object? If so, why?

Focusing on the object in the absence of any thought contents to clutter up the situation, such as memories of a dog named 'β', generally suffices for the token to refer. When it does, I know that a reference is made and to what, as I know that a sound occurs or a twinge of pain. To be sure, memory is involved as, in normal circumstances, a complex causal history of language learning is involved. Noting that does not explicate anything, it only indicates causal conditions for what takes place taking place. The real

problem is found in our third case. I turn away, see Kripke, and say 'That is not β'. The token of 'that' does not pose a problem; the token of 'β' does. Suppose I were asked what I referred to. To reply, 'The patch I just drew on the board', would satisfy Kripke's ordinary criteria for taking it to be what I intended to refer to.[10] But what is it to intend to refer?

Whatever takes place physiologically, nothing relevant takes place phenomenologically, besides the occurrence of the token. This often happens when something or some event is remembered. One remembers a face or a scene through the occurrence of a mental image, without Russell's purported awareness of pastness, or of any phenomenological indication of a connection to the past. The memory image simply occurs. Our third case of reference is a case of remembering, whether or not memory is involved in the second case. The token represents the object as a memory image represents what is remembered. I remember the patch, but neither Aristotle nor Bismarck, for neither was ever a focus of attention. So enters one aspect of Russell's Principle of Acquaintance and the consequence that whatever causal chains may reach from ancient Greece to utterances in this room, none of the latter refer to Aristotle. It is irrelevant that, by ordinary standards, I referred to The Philosopher several times.

We no more deal with the philosophical problems of reference and intentionality, by distilling the causal conditions for saying that I referred to Aristotle, than we resolve the classical problems of perception by specifying the causal network involved in perceiving β. A primitive reference relation is acknowledged, since it is required to fit the phenomenological facts. But, does such a relation obtain between purported representations and objects that are not presented?

To hold that we can only refer to what is presently given in experience is to virtually deny that thinking takes place; for, if such a principle is applied to tokens of predicates, thought contents will be limited to properties presently presented.[11] To insist that properties differ from particulars in that, though not perceived when a predicate token occurs, properties are conceived of is to argue that properties, but not particulars, can be simply remembered, while particulars are remembered by means of properties. This repeats one familiar theme; properties are recognizable as such, particulars are not: and recalls a second; particulars, as such, are unintelligible substrata, i.e. noncognizable, except as the bearers of properties. Hence, they are only known and recognized by description. But remembering is not recognizing. While recognition of a particular might involve recognizing or remembering properties, remembering a particular need not.

[10] Kripke 1980: 96-97.

[10] This point lies behind much of what Searle has to say and his correct rejection of homunculi. See Searle 1983: 21, and Hochberg 1978: 194.

[11] Unless the occurrence of a predicate token is assumed to indicate that a property is represented.

One cannot hold that the recognition or remembering of properties involves recognizing or remembering properties of such properties. This leads to a point about the misnamed "descriptive theory of proper names", which in spite of current labels is not Russell's theory. It purports to be a general theory of reference, but it can only be sensibly held by a nominalist. For, if primitive predicates refer to properties, one cannot hold to such an account of reference without falling victim to an obvious and vicious regress. In so far as nominalism fails, so does the descriptive account. Moreover, descriptive accounts are irrelevant to the real issues, for, like causal accounts, they merely furnish criteria for deciding when we "would say" that someone referred to a described object. Neither account tells us what it is to so refer. If, in the style of verificationist theories, one claims that what 'reference' means is explicated by the criteria of the account, then we have a stipulated definition that purportedly picks out those cases we would ordinarilly call cases of reference. But the real issues concern the analysis of initial cases of reference – 'This is white', 'Call this 'β'!' – and whether the reference relation involved in such cases is present in cases where the referent is not present.

Kripke might argue that, just as 7 is a number, since 0 is a number and 7 belongs to the posterity of 0, so too, given an initial reference to Aristotle and the present token being linked in a causal chain to that initial reference, I referred to Aristotle (assuming I intended to refer to him). But, one cannot take the class of references to Aristotle as the "posterity" of an initial reference, as one specifies the class of natural numbers as the posterity of 0. Reference in an initial case involves a basic reference relation. One can then stipulate a definition for 'refers to', a new disjunctive predicate, that will apply to both initial cases and "descendants" of such. This does not show that the basic reference relation that obtains in an initial case holds between a descendant token and the object. Kripke avoids the issue when he takes the "notion of intending to use the same reference as given"[12] and offers neither an ontological assay of nor an argument for Aristotle being the referent of the preceding token.

References

Davidson, D. 1969 "True to the Facts", *Journal of Philosophy* 66, 748–764..

Geach, P. 1971 *Mental Acts*, New York: Humanities Press.

Hochberg, H. 1978 *Thought, Fact and Reference: The Origins and Ontology of Logical Atomism*, Minneapolis: University of Minnesota Press.

Hochberg, H. 1992 "Truth Makers, Truth Predicates, and Truth Types," in *Logic, Truth and Ontology*, ed. K. Mulligan, Amsterdam: Kluwer, pp. 87–117.

Kripke, S. 1980 *Naming and Necessity*, Cambridge: Cambridge University Press.

Russell, B. A. W. 1922 "Introduction," in Wittgenstein 1950, ix-xxii.

Searle, J. 1983 *Intentionality*, Cambridge: Cambridge University Press.

[12]Kripke 1980: 97.

Whitehead, A. N. and Russell, B. A. W. 1950 *Principia Mathematica*, v. 1, Cambridge: Cambridge University Press.

Wittgenstein, L. 1950 Tractatus Logico- Philosophicus, trans. D. F. Pears and B. F. McGuinness.

Can there be a Language of Thought?

Ansgar Beckermann

1. Cognitive sciences in a broad sense are simply all those sciences which concern themselves with the analysis and explanation of cognitive capacities and achievements. If one speaks of *cognitive science* in the singular, however, usually something more is meant. Cognitive science is not only characterized by a specific object of research, but also through a particular kind of explanatory paradigm, i.e. the information processing paradigm. Stillings *et al.*, for example begin their book *Cognitive Science* as follows:

> Cognitive scientists view the human mind as a complex system that receives, stores, retrieves, transforms, and transmits information. (Stillings 1987: 1)

The information processing paradigm however, leads directly to the paradigm of symbol processing, because a system can, as it seems, only receive, store and process information if it has at its disposal a system of internal representations or *symbols*, i.e. an internal language in which this information is encoded. At least this appears to be an idea which suggests itself and which Peter Hacker expresses as follows:

> ... if information is received, encoded, decoded, interpreted and provides grounds for making plans, then there must be a language or system of representation in which this is all done. (Hacker 1987: 486f.)

And indeed the assumption that in cognitive systems there must be something like a system of internal representations, or a language of thought,[1] lies at the heart of many new works in the fields of cognitive psychology and cognitive neurobiology. For these sciences this assumption has the status of an empirical hypothesis, that is to say, for them, internal representations or symbols are theoretical constructs which are postulated because they allow

This is a revised version of a German paper which I read at the "Jahrestagung des Instituts für deutsche Sprache" on March 16, 1993, in Mannheim. I would like to thank Antonia Barke for translating the paper into English.

[1] The expression 'language of thought' ('*lingua mentis*' – '*Sprache des Geistes*') which was – as far as I know – first used in this context by Harman (1973) is seriously misleading because the expressions of the language of thought are – and Fodor e.g. agrees with this – internal physical states of the respective system, as for example certain neuronal firing patterns or bit patterns in the memory of a computer. Hence expressions like 'language of the brain' or 'language of the computer' would be more precise.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 207-219

us to explain cognitive achievements in a well corroborated and systematically particularly satisfying way. On the other hand, there are philosophical approaches that support the assumption through very general considerations concerning the nature of mental states.[2]

These and other related approaches have been criticized by a variety of authors in many different ways. Especially in Oxford however, criticisms have been formulated which are based on the late Wittgenstein and which radically question the symbol processing paradigm in general.[3] Peter Hacker, for example, in his article "Languages, Minds and Brains" asks the rhetorical question:

> Is this [*sc.* the idea that there is a language of the brain] just a picturesque metaphor or helpful analogy? Or is it a symptom of widespread confusion in the presentation, description and explanation of experimental data...? (Hacker 1987: 487)

And his answer indeed states that the idea of a system of symbols in the brain is founded upon a fundamental confusion of concepts and therefore is literally *nonsensical.* What are Hacker's reasons for this devastating assessment?

His argumentation begins with a characterization of the idea he then wishes to attack:

> The general conception at work involves the supposition that the brain has a *language* of its own, which consists of *symbols* that *represent* things. It uses the *vocabulary* of this language to *encode information* and it produces *descriptions* of what is seen... (Hacker 1987: 488)
>
> A 'symbolic description' is presumably an array of symbols which are so combined as to yield a true (or false) characterization of a certain aspect of the world. It must be cast in a certain language which has a vocabulary and grammar. (Hacker 1987: 488)

We, thus, have to ask what it *could* mean for the brain to possess a language with its own vocabulary and its own grammar. Before trying to answer this question however, we should first get clear about what it *does* mean in general to say that someone possesses a language.

> Someone who *has* a language has mastered a technique, acquired or possesses a skill of using symbols in accord with rules for their correct use, or – if you prefer – in accord with their meaning. (Hacker 1987: 491f.)

Someone's having a language thus consists in his possessing certain abilities. He understands utterances made in the language; he knows the

[2]The main figure in this field is Jerry Fodor, who developed his Representational Theory of Mind over many years before casting it into its canonical form in *Psychosemantics.* See Fodor 1975; 1978; 1981; 1987.

[3]See Hacker 1987, and also the new collection of essays Hyman 1991.

meaning of the words of this language and is able to use them in order to carry out a broad variety of speech acts: He can call a taxi, ask for the way to the rail station, tell stories or make jokes, order wine with a meal, introduce a friend, describe a landscape, and so on. Over and above all that he can - should it happen that he is not understood - explain what the words he has used mean and what he wanted to say by uttering them.

> If [someone] understands a language he can respond in various ways to others' uses of words and sentences, as well as correcting others' errors, querying their unclarities and equivocations. (Hacker 1987: 492)

From this fact alone – that mastering a language implies all these skills – it follows, according to Hacker, that it is literally nonsensical to say that the brain possesses a language.

> Only of a creature that can perform acts of speech does it make sense to say that it has, understands, uses, a language. But it is literally unintelligible to suggest that a brain, let alone *a part of a brain*, might ask a question, have or express an intention, make a decision, describe a sunset, undertake an obligation, explain what it means, insist, assert, instruct, demand, opine, classify, and so forth. (Hacker 1987: 492)

In order to be capable of possessing a language one must be able to carry out certain actions – actions which belong to a different level than those from which one can meaningfully say that they are being done by a brain, let alone parts of a brain. Brains or parts of brains are not therefore, for conceptual reasons alone, possible language users.

But there are more reasons which, in Hacker's view, show that the idea of a language of the brain becomes increasingly absurd the more the implications of this idea become clear to us. The expressions of a language, he continues, have a use governed by convention and someone who masters a language must know the correct use of these expressions, i.e. he must be able to distinguish correct uses from incorrect ones.

A rule-guided use of language which refers to standards of correctness can, however, only be founded on a social practice.

> For only where there is a practice of employing a sign can there also be an activity of matching the application of the sign against a standard of correctness. Since signs have a meaning, a use, only insofar as there is a convention, a standard of correctness for their application, there must be a *possibility* of correcting misuses by reference to the standard of correctness for the use of the expression which is embodied in an explanation of meaning. The use of language is essentially a normative activity. (Hacker 1987: 496)

This is another reason why according to Hacker it is altogether impossible that brains or brain cells employ a language: One cannot meaningfully say that brains or brain cells follow conventions, because conventions

can only be followed if they exist at all. They can only exist however, where they are used within a social community in order to teach and learn, to correct mistakes and explain and justify actions.

> Only of a creature who has the *ability* to make a mistake, who can *recognize* his mistake by reference to a standard, who can *correct* his action for the *reason* that it was erroneous, only of such a creature can one say that it follows and uses conventions. (Hacker 1987: 496)

It is for the very same reason that Hacker believes that even the talk of cerebral maps is nonsensical: because maps are only maps of something if appropriate conventions exist. There simply is no such thing as representing a territory on a map without employing specific sets of conventions of representation including specific methods of projection (e.g. the Mercator projection).

> So there are no representing maps without conventions of representation. There are no conventions of representation without a *use*, by intelligent, symbol-employing creatures, of the representation. And to *use* a representation correctly one must *know* the conventions of representation, understand them, be able to explain them, recognize mistakes and correct or acknowledge them when they are pointed out. Whether a certain array of lines is or is not a map is not an *intrinsic* feature of the lines, nor even a *relational* feature (that is, the *possibility* of a 1:1 mapping), but a *conventional* one (that is, the *actual* employment, by a person, of a convention of mapping). (Hacker 1987: 497f.)

Thus, one is forced to accept the conclusion that the idea of a language of the brain is literally nonsensical. There can be no meaningful symbols in the brain, because meaning presupposes the existence of conventions and conventions in turn imply the existence of a corresponding social practice. A "social practice"' of the kind required however, is *conceptually* impossible with respect to brain cells. The assumption that the brain employs a language or uses a system of symbols is therefore literally "inconceivable".

2. At first sight this argumentation appears to be extremely plausible. And it indeed forms the core of a Wittgenstinian theory of meaning, which is shared by many. A closer look, however, will reveal that this argumentation is not quite as cogent. This is so because even the reference to a social practice cannot – at least if one follows Kripke's reasoning concerning this point[4] – provide grounds for the *normative* character of meaning. This is at least the way in which Paul Boghossian reads Kripke.[5] Boghossian asks what the normative character of meaning consists in and answers:

[4]Kripke 1982.
[5]Boghossian 1989.

> Suppose the expression 'green' means *green*. It follows immediately that the expression 'green' applies *correctly* only to *these* things (the green ones) and not to *those* (the non-greens). The fact that the expression means something implies, that is, a whole set of *normative* truths about my behaviour with that expression: namely, that my use is correct in application to certain objects and not in application to others... meaningful expressions possess conditions of *correct* use. (Boghossian 1989: 513)

From this follows the sceptical problem for all theories of meaning:

> Having a meaning is essentially a matter of possessing a correctness condition. And the sceptical challenge is to explain how anything could possess *that*. (Boghossian 1989: 515)

Correspondingly, Kripke's main argument against all theories which attempt to reduce meaning to natural properties of individual persons, and especially against the dispositional analysis of meaning, runs like this: None of the natural properties presented by these theories can account for the fact that expressions have conditions of correctness, and it is precisely because of this that all these theories, *as* theories of meaning, are doomed to failure.

At this point the Wittgenstinian brings into play rules which are grounded upon social practices and argue: Everything said so far is right, but what it shows is simply that meaning is not constituted through properties of *isolated* individual persons. The meaning of a linguistic expression only springs from the rules on which the use of the expression in question is founded. And these rules, in turn, result from a common social practice. But does this answer suffice? Can rules and can especially a social practice give better grounds for the conditions of correctness of a linguistic expression than the properties of individual persons?

Following Hart (1961: 54ff.) we can explain the fact that in a community there exists a rule R as follows:[6]

(1) The members of the community rarely deviate from R,

(2) If a member of the community deviates from R, then (s)he is exposed to sanctions from the other members of the community,

(3) These sanctions are – generally – accepted.

If this is so, than the fact that there exists a rule within a community consists only in the *dispositions* of the members of that community. And this in turn leads to the question: 'In which way can the dispositions of a number of people provide better grounds for the conditions of correctness than the dispositions of an individual person?'

This is the reason why Kripke himself accepts the reference to the rules of a linguistic community only as a sceptical solution of the problem

[6]See also von Savigny 1983: 34.

of meaning. A substantial solution is, he thinks, impossible. Nothing in the world can account for the normative character, i.e. the conditions of correctness, of linguistic expressions. Therefore, in a strict sense, the conclusion is inescapable that no linguistic expression has the property of having a certain meaning. Hence it is nonsensical to ask what this property consists in. The only thing we can do is to *describe* under which conditions we ascribe which meanings to which words, and perhaps ask why we do it this way rather than another.

Following this line we then find, according to Kripke, that in the ascription of meaning we actually *do* refer to actions and dispositions of members of linguistic communities. And, what is more, Kripke also holds – in common with many Wittgenstinians – that it simply does not make any sense, i.e. does not serve any intelligible purpose, to ascribe meaning to the utterances of an isolated individual person and that therefore our reference to social practices is not accidental but in a certain way inevitable.

However, if one *were* to investigate the problem of meaning in a way which is concerned, not only with the *description* of a practice of ascription, but also with an *explanation* for this practice, then there might be more alternatives available.

In this spirit I will explore in the following, whether there are not some good reasons after all, for the practice of many cognitive scientists who regard certain physical (e.g. neuronal) structures as representations with a certain meaning. If it should turn out that this in fact is so, this would in my opinion also show that speaking of a language of thought (or the brain) – notwithstanding the arguments of Hacker and others – has a perfectly intelligible sense after all.

3. However, I would like to begin with a concession. Hacker has made it very clear that according to our normal use of the word 'language' a language can only exist if there are beings who speak this language and that it can only be said of a being that it employs a language if it masters a certain broad range of behavioral patterns.[7] One of his arguments against the idea of a language of the brain was precisely that neither the brain nor parts of it *can* master such a behavioral repertoire. And in this he is certainly right.

A language of thought, therefore, can only exist if it is – in a certain way – radically different from all normal languages, for a language of thought, if it exists, is a language which is not spoken by anyone, nor understood by anyone – it is not even heard by anyone. (If some people talk as if the brain would speak or understand this language, then this mode of speech can only be meant metaphorically.) A language of thought is, as it were, a language which simply happens. Sentence tokens of this language

[7]In my opinion it is a very interesting question whether the whole behavioral range is *really* a necessary condition for the possession of a language, or whether we would not be inclined (or even forced) to attribute a language to beings who only possess a part of the skills Hacker mentions. Unfortunately I cannot pursue this question here any further.

just arise in the brain under certain conditions, are altered in accordance with certain formal rules and – together with more sentence tokens – cause certain actions. The sentence tokens need not be uttered in order to exist and have (causal) effects. All this happens, one is almost tempted to say, as if by itself. Given these conditions however, the question suggests itself: To what extent can one speak here of a language at all? This question is certainly justified and I am not absolutely certain whether it can be answered convincingly. However, I would like to begin tentatively with the following consideration: A language can first of all be simply conceived of as a system of structured sentences with combinatorial semantics. The sentences have a meaning (truth conditions) and this meaning depends in a systematic way on the meaning of its constituents. One can distinguish between sentence types and sentence tokens. Sentence tokens are physical structures concerning which one can tell which sentence type they realize. If one accepts this, one can perhaps agree with the following as well: If a number of physical structures exist in a system which can – with good reasons – be conceived of as tokens of certain sentence types and insofar also as having certain truth conditions, then there exists an internal language in the system. Perhaps, someone might claim that the term 'language' would be inappropriate in such a case and would instead prefer to speak of a system of internal representations. To this I would have no objections since systems of internal representations are all the cognitive scientist needs. And I am quite sure that no cognitive scientist ever took a language of thought to be something more than this. On the other hand, Hacker's arguments, as he makes clear enough,[8] are meant to count against systems of internal representations in the same way as against the idea that there could be more fullblooded languages in the brain. To opt for the former alternative, therefore, does not change the overall dialectical situation.

In the remaining sections I am going to argue for the thesis, that there really are good reasons for conceiving of certain systems in the way explained in the last paragraph (or that it is at least possible that there are some) and that therefore the idea of a language of thought (i.e., of systems of internal representations) in the sense described above is not at all nonsensical. The introduction to this will be a very general remark from the field of the philosophy of science.

4. If we try to explain and understand the behavior of complex systems it is often not enough to take only the *physical stance*, as Dennett[9] calls it. Often an adequate understanding is reached only when we also understand the *functional organisation* of these systems. That this is so becomes especially clear in the field of biology: There explanations are frequently given on the

[8]See for example Hacker's claim: "Nothing in the cortex constitutes a 'symbolic representation' of the creature's environment." (Hacker 1984: 497)

[9]The distinction between physical, functional and intentional stances goes back to Dennett 1971.

functional level alone while anatomical and physiological details are hardly mentioned. Let us take an example – for instance temperature regulation in the human body, which is explained in the textbook *Biological Psychology* by Birbaumer and Schmidt as follows:[10]

> Thermoregulation can formally be viewed as a closed circuit regulatory system with a negative feedback loop. Body temperature is monitored by sensors, namely the thermoreceptors, which feed information into the central regulator. The latter checks whether the body temperature (the *actual value*) has deviated from its *desired value* and alters the control medium by sending *control signals* until feedback from the thermoreceptors signals that the mismatch has been compensated.
>
> The body's core temperature is registered at different sites through temperature sensitive cells or sensory neurons, the surface temperature through thermoreceptors within and beneath the skin. The hypothalamus, especially the posterior hypothalamic area, is likely to be the integration centre of temperature regulation. Central effector neurons control (probably via a chain of interneurons) the final control elements for the production and extraction of warmth (production of warmth, insulation of the body surface, sweat production and behaviour). They receive their afferent input from peripheral and central thermoreceptors. Cold receptors directly activate the effector neurons for the production of warmth and inhibit, via some interneurons, the final control elements for the extraction of warmth. Warmth receptors are wired in exactly the opposite way to the two types of effector neurons. (Birbaumer and Schmidt 1990:117–121)

The almost exclusive use of functional vocabulary ins obvious. Sensors, control media and feedback control systems are mentioned as often as integration centres, thermoreceptors and effector neurons. The only genuinely physiological concepts seem to be anatomical expressions like 'posterior hypothalmic area', and this is the case even though the story *could* be told completely also in purely physiological terms. But – apart from the fact that this story isn't known to us in all its detail – *this* story alone wouldn't satisfy us, because what we are really interested in is the question of how the body manages to maintain a relatively constant temperature under extremely differing conditions. And we only understand *this* if we realise that the physiological processes interact in the form of a feedback control system and therefore can be described with the help of the corresponding conceptual scheme. Functional concepts, therefore, are brought in especially if one isn't mainly interested in explaining individual physical states or activities, but in understanding how *successful* behaviour comes about, i.e. how a system manages to produce, under the most varied conditions, behaviour which meets certain standards. We can sum this up as follows:

[10] I'm abridging this description strongly.

Thesis 1 *Often we can only explain and understand the* successful *behaviour of systems adequately if we proceed from the physical stance to the functional stance with regard to those systems.*

I'd like to add here a short remark concerning functional systems: in this context, it is important to note that properties such as being a sensor or being a final control element are not natural properties in the usual sense of the term. This means that we cannot ascribe concepts such as 'sensor' and 'final control element' – as opposed to concepts such as 'pyramidal neurons' or 'neuromuscular synapses' – on the grounds of normal observable or measurable neurobiological characteristics. This is so because the applicability of these concepts to certain neurobiological phenomena depends on whether these phenomena interact in such a way that a circuit pattern results which can be *interpreted* as a closed circuit feedback control system. To put it in a rather simple – though somewhat misleading – way: functional properties do not exist in the world, we read them into the world.

5. The example of thermoregulation, however, is a little too unspecific to allow conclusions about the sense or nonsense of the idea of a language of thought. Another example might be a bit closer to the point – namely, the example of a chess computer which Dennett has often used for the purposes of illustration.[11] For such an electronic device, it is possible – in principle, at least – to explain every move in purely physical terms: one can ascertain how certain local states of silicon chips change through pressing certain letter or number keys; one can further deduce the sequence of the states these chips will go through after pressing 'Enter' from the circuit and the initial states of these chips. In the same way one can finally calculate which state will end this sequence and which of the diodes which make up the display will be lit. What can be achieved in this way, however, is only the explanation of certain concrete final states on the basis of of knowing the concrete initial conditions. What cannot be achieved is an understanding of the mechanisms that enable the device to produce outputs which correspond to moves which are plausible, or even successful, in the relevant situation of the game.

Such an understanding can again only be reached if we move on from the physical to the functional stance. In this particular case, this amounts to analysing the *program* which underlies the behaviour of the chess computer. Because only then is it possible to conceive of what happens between input and output not just as a sequence of states of silicon chips. Only in the functional stance can we interpret certain local states of these chips as representations of possible configurations of pieces on a chess board. Only if we presuppose the functional stance can we describe the occurrences between input and output in a way which is almost familiar by now: the computer first calculates the representations of all possible successive configurations of the actual situation whch would result from the moves possible for the

[11] First in Dennett 1971.

computer; then it repeats this calculation for all the moves open to the oppnonent to respond to these, and again for the computer's own moves to respond to the opponent, and so on until a certain number of moves and countermoves has been reached. The individual configurations are evaluated according to given criteria, before finally the computer produces, as an output, the move which leads to the configuration with the highest evaluation, taking into account moves of its opponent.

This way of telling the story makes it possible, for the first time, to understand that our computer normally makes plausible, or even successful, moves, because it can be shown that the evaluative function underlying the choice of moves indeed results in plausible, or even good, moves under the conditions in question. If the computer in the end makes the move which leads to the highest evaluation, its moves must, on average, be rather good ones. I say 'on average', because there are configurations which are objectively disadvantageous notwithstanding a high evaluation. If such a situation occurs, the computer often doesn't choose a particularly good move. This, however, need not surprise us, because we know, of course, that the computer sometimes makes mistakes. So the description of what takes place between input and output, with the help of the program outlined above, is doubly helpful in explaining the behaviour of the computer: the description explains why the computer normally chooses good moves, and it also explains why it sometimes makes grave mistakes.

The example of the chess computer, as well as many examples of biological systems, show that we can often understand the behaviour of complex systems only if we proceed from the physical to the functional stance, and that this is particularly so if the behaviour in question is of a kind which, measured against certain standards, can be classified as successful. More important than this general point, however, is a point which comes to our attention if we take seriously the functional stance towards certain systems, e.g. towards a chess computer.

I have already mentioned that assuming the functional stance with respect to a chess computer amounts to analysing the program implemented by this computer. And this in turn means two things; firstly, we conceive of certain steps which take place between input and output as the execution of certain instructions, and, secondly, we reconstruct how the system organizes the sequence of these steps. The execution of a certain instruction normally consists in the production or manipulation of a certain data structure. This means that we can only conceive of certain physical processes as the execution of an instruction if we also at the same time view certain physical structures as data structures. With respect to the functional analysis of the chess computer, this means concretely: We can reconstruct the program which underlies its normally successful behavior only if we conceive of certain physical structures within the system (the local states of certain silicon chips) as representations of possible configurations and of other physical structures of this kind as representations of evaluations. If we generalize

this result we arrive at

Thesis 2 *The functional analysis of a system is in some cases only possible if one conceives of certain physical structures within the system as representations.*

Following Hacker's line of argument one might be tempted to object that the argumentation up to now does not take into account the fact that chess computers are artifacts which have, indeed, been programmed by their manufacturers with a certain purpose. With respect to these artifacts, one can therefore say that they carry out programs and hence that within them there exists something like representations, because in this case there is someone – namely the programmer – who intends to represent certain configurations of chess pieces by means of certain physical structures. Representations without a person who uses them, however, would still be impossible.

This objection however, would miss the very point of my argumentation. Since this point is precisely that, with regard to some systems – independent of their origin – we *must* assume that there exist representations in them if we want to understand how the successful behavior of these systems comes about. We therefore would have to describe chess computers in exactly the same way as I have explained above, even if they were to grow on trees.

And it can be easily shown that this explanatory strategy is, indeed, pursued in neurobiology. I remember vividly a discussion in the course of which I once asked the Göttingen physicist Manfred Schroeder which neuronal mechanisms are responsible for the localization of sources of sound. His answer began with the sentence: "Firstly the crosscorrelation of the signals of the two auditory nerves is calculated in the brain". Another example of the same type can be found in J. Koenderink's article "The Brain a Geometry Engine"; it is Koenderink's central thesis that the best way to understand the mechanisms of the visual cortex is to take as a starting point the two dimensional intensity distribution of the light quanta which strike the retina and then to interpret the following neuronal processing as the calculation of the first, second and higher differentiation of this distribution.

> ...you may understand a large part of the structure of the front-end visual system as an embodiment of differential geometry of the visual field... Instead of the concrete 'edge detectors' and 'bar detectors', one speaks of the abstract first- and second-order directional derivatives. (Koenderink 1990: 125)

I cannot go into more detail here, but I hope it becomes clear even from these sketchy examples that in fact many neurobiologists take the functional stance in order to attempt to explain the amazing achievements of the brain, and that they go even further and try to reach explanations on the basis of the assumption that in the brain certain calculations really take place.

7. From this it is only a small step to my general conclusion: Just as it is necessary to assume that, within chess computers, there are representations and evaluations of configurations in order to understand how these devices succeed in producing successful moves, it may be necessary, with respect to other systems, to assume that there are sentence-like representations within them, if one wants to understand what enables these systems to behave successfully. (In this context no more is meant by the expression 'sentence-like representations' than 'structured representations with combinatorial semantics'.) This would – for example – apply to all AI systems, the problem solving behavior of which is based upon automatic theorem proving, because we cannot adequately understand the behavior of these systems without interpreting some of the processes taking place within them as *inference processes.* And inference processes are processes in which sentence-like representations are derived from sentence-like representations. That is to say, we cannot conceive of some processes within the system as inferential if we are not prepared to interpret some of the physical states within the system as sentence-like representations. Here – as in the example of the chess computer – it can be seen that the interpretation of processes has priority over the interpretation of states: Certain processes in a system cannot be adequately understood if we do not interpret certain states in a corresponding manner.

And now I think it is also clear under which conditions we are virtually forced to assume that there are sentence-like representations or symbols within certain systems, i.e. that these systems contain a language of thought. We are forced to assume this if we can only understand what underlies the successful behavior of those systems if we interpret some of the physical processes within them as processes of production and manipulation of sentence-like representations. To sum this up in a last thesis:

Thesis 3 *The assumption of sentence-like representations is not only plausible, but in a certain sense unavoidable if we can explain the successful behavior of a system only by means of the assumption that it is founded upon functional processes which can only be understood as processes of the production and manipulation of sentence-like representations.*

Speaking of sentence-like representations therefore is neither one of the little quirks of certain cognitive scientists, nor a habit which springs from a fundamental confusion. Rather it is a consequence which results from the attempt to understand the functional architecture of some systems underlying their successful behavior.

By way of conclusion I want to emphasize strongly that Thesis only formulates a *condition. If* this condition is satisfied, then we can say that within a system there exists a language of thought or a system of internal representations. This thesis however, does *not* imply that Fodor, or other cognitive scientists, are right in believing that intelligent behavior can only be explained within the symbol processing paradigm. However, my purpose

was not to defend this paradigm, but rather to save it from the charge of conceptual confusion.

References

Birbaumer, N. and Schmidt, R.F. 1990 *Biologische Psychologie*, Berlin: Springer.

Boghossian, P. 1989 "The Rule-Following Considerations", *Mind* 98, 507–549.

Dennett, D. 1971 "Intentional Systems", *Journal of Philosophy* 68, 87–106. Reprinted in D. Dennett, *Brainstorms* Montgomery, VT: Harvester 1978, 3–22.

Fodor, J.A. 1975 *The Language of Thought*, New York: Thomas Y. Crowell.

Fodor, J.A. 1978 "Propositional Attitudes", *The Monist* 64, 501–523. Reprinted in J.A. Fodor, *Representations*, Cambridge, MA: MIT Press 1981, 177–203.

Fodor, J.A. 1981 "Introduction - Something on the State of the Art", in J.A. Fodor, *Representations*, Cambridge, MA: MIT Press 1981, 1–31.

Fodor, J.A. 1987 *Psychosemantics*, Cambridge, MA: MIT Press.

Hacker, P. 1987 "Languages, Minds and Brains" in C. Blakemore and S. Greenfield (eds.), *Mindwaves: Thoughts on Intelligence, Identity and Consciousness*, Oxford: Blackwell 1987, pp. 485–505.

Harman, G. 1973 *Thought*, Princeton: Princeton University Press.

Hart, H.L.A. 1961 *The Concept of Law*, Oxford: Oxford University Press.

Hyman, J. (ed.) 1991 *Investigating Psychology*, London: Routledge.

Koenderink, J.J. 1990 "The Brain a Geometry Engine", *Psychological Research* 52: 122–127.

Kripke, S. 1982 *Wittgenstein on Rules and Private Language*, Oxford: Blackwell.

von Savigny, E. 1983 *Zum Begriff der Sprache*, Stuttgart: Reclam.

Stillings, N.A. *et al.* 1987 *Cognitive Science: An Introduction*, Cambridge, MA: MIT Press.ls

Distinguishing Perceptual from Conceptual Categories

Rita Nolan

I

The area between sensation and conceptualization is grey and confusing. Despite abundant philosophical and empirical research, results about how to understand this area which command widespread assent are very scarce. One contributory source to this impasse is the fact that, for mature and intact humans, the sensory, the perceptual, and the conceptual seem merged in consciousness. Perception is phenomenally so "cognitively penetrable" – so infused for humans by discursive understanding – that experimental and theoretical efforts to distinguish between it and conceptualization, and consequently between it and sensation, often seem constrained only by whatever favored theory drives the effort. In what follows, I consider reasons for distinguishing perceptual from conceptual categories and suggest a way of making the distinction. First, however, some preliminaries will help make clearer just what topic is under discussion.

II

Another approach to the problem of my concern can be made through the Wittgensteinian problematic: Is all seeing, seeing as; and, more generally, does all perceiving require interpretation? On the account suggested by the considerations I shall make, both questions are obscure; the notions of seeing as and interpretation that are engaged by them fail to distinguish between non-conceptual categorization and conceptualization. It may be that all visual perception requires categorization, even though not all categorization is conceptual. This failure is symptomatic of widespread unclarity about how to understand the differences among sensation, perception, and conceptualization.

Sometimes the Wittgensteinian problematic is taken as inviting an account of "seeing an aspect", where this is understood as equivalent to an account of seeing something "under an aspect". Recent discussions of "perceptual content", for example, of the perceptual content motivating frogs's leaps at flies, suggest this interpretation. Talk about perceptual content, as of mental content in general, invokes the metaphor of mind as container, standardly, as container of information, and directs one to provide a discursive account of the purported content. To give such an account, however, of

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 221-231

what is presumed to be phenomenal "content" – how it seems to the frog for example – is to engage in conceptualizing that content, and from the frogs's point of view as it were. But this is to suppose that it is the frog, and not we, who conceptualizes this content, a suggestion that can hardly be supposed transparently true, however compatible with some favored metatheory it may be.

Considerations like this one recommend that casting questions about differences between perception and conception in terms of mental content be avoided, whenever it is possible to do so. For it may be natural to suppose that the frog perceives, but not natural, absent more compelling evidence, to suppose that it conceptualizes that which it perceives. In general, whether or not it turns out that the ability to conceptualize is in fact dependent upon having a public language (I cannot imagine that it is not), one is on safer ground to limit clear examples of conceptualization to creatures who manifest their conceptualizations in explicit judgments.

III

Several sets of circumstances conspire to obscure the differences between perceptual and conceptual categories. Here is one such set.

Normally we distinguish between sensation and perception in complex and multifunctional organisms according as the response that is evidence for either is a local response by some part of the organism or a general response of the whole organism. Such reflex responses as withdrawing the hand from a hot object are clear examples of the former while fleeing from a predator is normally an example of the latter. A response that is a general response of a whole complex organism, and thus evidence of perception, requires for its explanation some reference to central processing by the organism because a complex organism is presumed to require some way of coordinating its various parts in order to make a whole-organism response. There is, thus, the following dilemma: In humans, the *best* evidence for categorization is linguistic evidence – what category the subject says an object belongs to. But in non-language-using species, the *only* evidence that categorization has occurred is the occurrence of a non-linguistic, whole-organism response. So, either one makes the *ad hoc* decision to count only linguistic evidence as acceptable evidence of categorization, treating all categorization as conceptual, or one treats perceptual and conceptual categories as substantially equivalent. To take the first course seems to prejudice the investigation in favor of humans, supposing perhaps gratuitously that only humans could enjoy the privilege of categorization. But to take the second course seems to prejudice the investigation against humans, supposing perhaps gratuitously that language is, after all, merely an efficient technology for communicating information that is intrinsically language-independent. In either case, perception and conception are conflated.

Perception and conception are conflated, in particular, whenever it is supposed that if there is a general, whole-organism response, then a generalization that "sends", as it were, appropriate information to each part of the organism involved in the response constitutes a judgment about the character of the stimulus. For, in this way, judgment – construed as generalization – is projected onto pattern recognition and perception. Categorization is thus construed as conceptualization and taken to be a single central function of an organism in which distinguishable stimuli are understood and responded to as if they were identical. In one example of this conflation, Miller and Johnson-Laird (1976), studying perception and language, discount any developmental distinction between a response and an assertion on grounds that there is no evidence that such a development occurs. Consequently, they build judgment into perception itself, dispensing with any need to account for assertion as a distinct cognitive practice. But what would count as evidence of such a developmental difference in the course of language acquisition? Certainly, anything that is a response, linguistic or non-linguistic, is categorial in as much as it is a whole-organism response to some generalized perception, one, that is, which involves pattern recognition by the organism as a whole, and, thus, also involves central processing. Nevertheless, there may be a difference between perceptual categorization and conceptual categorization that is discounted if all evidence of categorization is taken as evidence of conceptualization.

Another example of the tendency to conflate perceptual and conceptual categorization is suggested by Medin and Barsalou's comparative study, "Categorization Processes and Categorical Perception" (1987). They begin this review of experimental data on categorization presuming a distinction between "sensory perception" (SP) categories and "general knowledge" (GK) categories and identifying general knowledge categories with (linguistic) semantic categories; thus, they begin by assuming that perceptual and conceptual categorization are different phenomena. They propose to compare the two types because most empirical research has been on one or the other, but not both. The conclusion of their comparison, however, seems to bring them close to conflating perceptual and conceptual categories. For the conclusion of their comparison is that there are "deep similarities" between SP and GK categories, and they urge further study at this intersection. They thus end by suggesting that the distinction with which they had begun may not be a clear one after all.

A second set of circumstances encourages the conflation of sensation and perception. While most philosophers are careful not to conflate perception and conception, a favored route for marking this difference is to distinguish between sensory-perception and conceptualization as between analog and digital representations. Such an account, however, may conflate sensation and perception, and, indeed, seems to foreclose on acknowledging perceptual categories by assimilating perception to sensation.

Stephen Palmer's work on representation theory (1978) raises specific objections to carefree assignment of analog and digital properties to alleged representations. Palmer argues that the two types of representation, ana-

log and what he there calls "propositional", are informationally equivalent as representations and differ rather in their inherent structures as representations. Because of their informational equivalence, he claims that any controversy concerning whether a (type of) mental representation is analog or propositional cannot be resolved without physiological psychologists "looking inside the head" (p. 298). Palmer claims that manifested behavior can reveal only what information a subject has but does not reveal the form in which the information is stored. He maintains that cognitive psychology is concerned only with matters at "the level of abstraction defined by informationally equivalent systems" (p. 277).

Palmer distinguishes between analog and propositional representations according to whether a representation is intrinsically or extrinsically related to what it represents. In his example, to represent the distribution of ages in a population by rectangular columns of different heights would be to use an intrinsic representation, since the columns represent ages in virtue of the relations of column heights to one another, a relation that inheres in the representation. The inherent structural characteristics of the representation constrain what it can function as a representation for. A proposition, on the other hand, is said, on his scheme, to represent something only in virtue of a relation to something external to it – that for which it is a true description. A proposition is an "extrinsic" representation because it is constrained in what it can represent by its relation to something external to it. But, Palmer argues persuasively, whether a type of mental representation is digital (here, "propositional") or analog is not discernible solely on the basis of information content that behavior reveals the agent to have.

I will return to Palmer's distinctions a little later, after considering recent behavioral evidence that suggests that perceptual categorization in humans may become digitalized by, while remaining distinct from, acquired conceptual categories. For the moment, I conclude that widespread acceptance of the analog-digital account of the contrast between perception and conception is assisted by conflating sensation and perception in the notion of sensory perception; and that such accounts are by no means transparently appropriate. But what I suspect is stronger than this: it is that perception may acquire digital properties in the course of conceptual (semantic) development.

IV

Why should perceptual categories be distinguished from conceptual categories? Since sensation, perception and conceptualization are normally phenomenally "merged" for mature and intact humans, appeals to our normal phenomenal experiences provide no compelling evidence for this distinction. The case for making this distinction must rely, then, on its theoretical utility and explanatory power. There are four types of considerations that have bearing, three of which I describe only briefly here; the fourth will be discussed in greater detail in later sections.

The first consideration is that a notable increase in economy, simplicity, and intelligibility in the theory of language acquisition can be effected by acknowledging that there is a level of cognitive development in which new perceptual categories are acquired in the course of early language learning but prior to genuine linguistic semantic conceptualization. The argument for this claim is developed at length elsewhere (Nolan 1994), but the next considerations suggest the focus of this argument.

A second consideration is that a principled distinction here would provide an explanation of prototypicality effects, effects that prompted Rosch (*et al.*, 1976) to introduce the notion of "basic categories", in particular during cognitive development and the emergence of language, but also in general. The distinction invites the framing of the following account. A degree of mastery of a category such as *dog* is achieved largely on the basis of perceptual schemata which constitute a perceptual category. That such a perceptual category cannot be "manufactured" (as it were) by humans merely as the output of sensation, we have been amply assured by the history of failures of sense datum theories; it is the categoriality of the category *dog* that escapes such theories. A perceptual category (such as *animal*) may then be subject to transformation during the course of development to yield a conceptual category (such as *mammal*), perhaps together with a replacement prototype as its correlative perceptual category. Prototypes can thus be understood as perceptual categories, even in the absence of semantic, conceptual understanding of the related concept.

A third consideration in favor of the distinction is that it makes transparent the inadequacies of purely extensional accounts of natural linguistic meaning, even extensional accounts couched in the idiom of possible worlds semantics. Thus, a degree of mastery of a category such as *triangle* can be achieved largely on the basis of perceptual schemata. In all possible worlds, however, equilateral and equiangular triangles are coextensive. The conceptual category of *triangle*, thanks to Euclid, is what makes possible the intensional non-equivalence of equilateral and equiangular triangles; what Bealer (1982; see also Bealer and Mönnich 1989) has called "fine-grained intensionality".

One may add to the foregoing considerations that the distinction may provide a basis for understanding the differences between human cognitive aptitudes and the aptitudes of members of non-human species. For we may allow that some of the latter may be susceptible to developing new perceptual categories as a result of experience without acceding that their development is conceptual. Distinguishing perceptual from conceptual categories may suggest, as well, an account of perceptual phenomena to which Gesalt psychologists have called attention. But there are problems with each of these possibilities that require more detailed attention than can be attempted here. For example, it is terribly unclear whether or when distinct sensory and perceptual levels play roles in the lives of members of non-human species; and Gestalt psychologists have called attention to phenomena which

seem, oddly, cognitively *im*penetrable (cf. Kanizsa, 1979) in the sense that they are not susceptible to correction by true beliefs which contravene them. It would thus seem premature to pronounce on these issues in the present context.

A fourth set of considerations is provided by experimental results in neuro-behavioral psychology. These results seem to mandate the acknowledgment of perceptual categories as distinct from conceptual categories, and I will describe them shortly.

V

The idea that there are non-conceptual perceptual categories that have a central function in our cognitive lives is closely related to the notion of non-conceptual perceptual *content* introduced by Evans (1982: 151n). However, the example Evans gives to illustrate the occurrence of non-conceptual perceptual content is our use of indexicals like 'here', which he calls "egocentric spatial terms" (p. 154). So, while he is concerned to distinguish an element in perception that is not conceptual, he does not seem to be referring to non-conceptual, perceptual categories, as I have used that term so far. He is discussing, he says, "the spatial element in the non-conceptual content of perceptual information" (p. 154). What Evans seems to be describing in these passages is the character of our conscious awareness of the location in external space of some stimulus (auditory, in Evans's examples) when we attend only to that phenomenon of location, rather than to the character of the stimulus as conceptualized. He appears to be attempting to isolate in our conscious (phenomenal) experience a non-conceptual, perceptual element – what he calls "information content" – regardless of whether this element is veridical, since he means to include mis-information in the scope of his "information": "The spatial information embodied in auditory perception is specifiable only in a vocabulary whose terms derive their meaning partly from being linked with bodily actions" (p. 157). He is, he says, "talking about... information whose content is specifiable in an egocentric spatial vocabulary."

But these perceptual information states with non-conceptual content "are not", he says, "*ipso facto* perceptual *experiences*" – that is, conscious states of a conscious subject. "However addicted we may be to thinking of the links between auditory input and behavioural output in information-processing terms – in terms of computing the solutions to simultaneous equations [he takes Fodor's *Language of Thought* as an example of such addiction] it seems abundantly clear that evolution could throw up an organism in which such links were established, long before it had provided us with a conscious subject of experience" (pp. 157-8). To illustrate this possibility, he cites "the case of a brain-damaged patient studied by L. Weiskrantz, who was able to point to a source of light despite claiming that he could not see anything at all" (L. Weiskrantz *et al.*, 1974). "'But always he was at a loss for words to describe any conscious perception, and repeatedly stressed that he saw nothing at all in the sense of "seeing", and that he was merely guessing"' (Evans, p. 158). What this shows, Evans says, is that "a conscious adult may display fairly normal responses to stimuli...and yet have no associated conscious experience". Elsewhere, Evans describes the "informational states which a subject acquires through perception" as "non-conceptual, or non-conceptualized" (p. 227). In contrast, "Judgments *based* upon such states necessarily involve conceptualization: in moving from a perceptual experience to a judgement about the world (usually expressible in some verbal form), one will be exercising basic conceptual skills" (p. 227). In other words, Evans describes the usual relation between non-conceptual perceptual states and conceptual states as a relation between two types of informational states such that one ordinarily moves from the former to the latter.

I conclude this synopsis of Evans's account of non-conceptual perceptual content by noting that while his notion of non-conceptual perceptual content is not clearly a notion of non-conceptual perceptual *categories*, his related notion of a non-conceptual perceptual information state (understood as neutral with respect to veridicality of information) may be useful to describe a perceptual state in which a perceptual category is invoked by a stimulus-situation.

VI

For moral reasons and more, the case of Weiskrantz's patient is the closest we can come to satisfying Palmer's stipulation that we must "look into the brain" if we want to distinguish different structures of mental representation. More recently, another medical accident has revealed information about mental structures of a type that provides empirical support for the distinction proposed between perceptual and conceptual categories. I note at the outset that this data, while systematically collected, was also an accidental result of a hospitalization and that the elderly patient succumbed to illness three months after these results were collected.

Hart and Gordon (1992) report behavioral research results with a subject who sustained cerebral damage and whose resulting knowledge deficits are directly relevant to the hypothesis that there are non-conceptual, perceptual categories. The authors maintain that the results they report "mandate the existence of two distinct representations" of the physical attributes of animals "in normal individuals, one visually based and one language-based" (1992: 60). The neurologically-impaired patient (K.R.), a retired librarian, exhibited category-specific dysnomia for the category of animals "despite the input modality (visual or non-verbal sound) or response route (oral or written)" (p. 60). In addition, the patient was unable to describe verbally the physical attributes of animals. Nevertheless, she was able to distinguish between visual representations of animals that depicted their physical properties correctly and incorrectly, and "her knowledge of other animal properties was completely intact" (p. 60). The modality-independence of

the patient's deficits lead the researchers to conclude that the deficit is a higher-level, central processing impairment. These results are also said to "establish that knowledge of physical attributes is strictly segregated in the language system from knowledge of other properties" (p. 60).

Let me highlight somewhat less technically a few of the specific results upon which the authors base their conclusions. The patient, K.R., was unable to name animals nor to describe verbally either the correct colors of animals or other visible, physical (as opposed to functional) attributes of animals, such as number of legs or size. For example, when asked what the color of elephants was, she responded, "Orange". However, she could pick out pictures which depicted correctly-colored, -legged, and -sized animals from those which depicted these incorrectly; K.R. could also correctly match, visually, animal bodies to their respective heads. Her success on these last two types of task indicate that she had visually-based knowledge of these features of animal categories, and that her knowledge-deficit was language-based. Her language-based impairment was limited to the visual attributes of animals (as well as to their names); she had no knowledge-deficit about other kinds of things nor about non-visual attributes of animals. For example, K.R. was able to answer correctly questions about functional properties of animals such as "Is an elephant edible?" and "Is it a pet?".

Hart and Gordon present other data from K.R. which they take as showing, on standard neuro-behavioral criteria, that her deficit was not merely a language-access deficit but was a language-representation deficit. They point out that any other hypotheses consistent with the data must nevertheless acknowledge that the deficit was category-specific to animals and that it mandates the existence of two distinct knowledge-representation systems for the visible physical properties of animals, one language-based and one visually-based (p. 63). The case of K.R. presents an elaboration over the minimalist "look into the brain" (to discern different representational structures) that is afforded by Weiskrantz's patient. Although philosophers will have further questions about the case that Hart and Gordon describe, K.R., unlike Weiskrantz's patient, provides direct evidence of the categoriality of non-conceptual perceptual "information". While Hart and Gordon focus their report on the character of K.R.'s *linguistic* deficit, the fact that her accurate, non-linguistic responses were occasioned by visual representations–line drawings in fact–of the properties of animals for which she retained no evident linguistic representations makes apparently unavoidable the conclusion that her visual perceptual information was categorial–about animals, about specific animals (animal species), and about their visible physical properties, but not about any particular individual animal. Indeed, their categoriality lies in this fact about them, that they are generic. In this respect, they differ markedly from the indexical, egocentric spatial non-conceptual perceptual content to which Evans calls attention.

Hart and Gordon end their report with a brief summary (p. 64) of what is known of the anatomical bases of category-specific deficits like K.R.'s, noting that relatively little pathological evidence is available. "Most patients have shown temporal lobe pathologies... Pathologies limited to the fronto-parietal lobes have been associated with impairments of the nonliving things categories". Of K.R.'s case, they mention "diffuse, mild inflammation attributable to a paraneoplastic syndrome that involved the cerebral cortex, including both temporal lobes" and "Incomplete or patchy disruption, rather than a complete and sweeping disruption". Their conclusion is that "A number of neural processing architectures (hierarchical or distributed networks) could produce the processing distinctions and anatomical assignments K.R.'s case requires."

VI

How might a distinction between perceptual and conceptual categories be drawn, supposing that the above considerations suggest that such a distinction should be made? How, that is, might we conceive of the difference between such types of categories so as to capture the chain of intuitions linking the considerations raised above? And, further, can we conceive of such a difference in a way that might provide some conceptual advance about the several issues mentioned here?

If we have, in the responses of K.R., a case in which intact non-conceptual perceptual categories have been manifested in a person's behavior in the absence of, and hence in isolation from, any correlative conceptual categories, then we should be able to describe those respects in which the one type of behavior differs from the other type. Here, I do not mean that the relevant differences are merely behavioral; obviously they are not. But it is unlikely, at least, and perhaps inconceivable, that any "looking into brains" alone that we might now do could reveal relevant differences without some further macro-description of the differences.

I propose that the focus for such a distinction may be found in adopting a principle introduced by Evans, what he called the "Generality Constraint", as a characterization of conceptual categorization and as distinguishing it from perceptual categorization. Although Evans introduced his principle in a context different from this one and with quite different theoretical goals from those that I am concerned with here, Evans was also keenly aware of the importance, and the difficulty, of distinguishing between perception and conception. Indeed, Gillett has argued (1987) for the stronger conclusion that there is a conceptual relation between Evans's constraint and the idea of a conscious thinking subject. The constraint is introduced in Evans's work during in the course of discussing Russell's theory of singular terms and was intended to contribute to an account of what is required for a person to be able to make a predicative judgment about a particular individual. Hence, the sentence schemata used in its statement are schemata for particular statements, and, so, use individual constants (i.e., a). The present proposal is that the same constraint, with some minor changes which include changing its individual constants to variables (e.g., a to x) in the schemata, can be

used also to characterize the ability conceptualizers have to make predicative judgments using general terms, concepts, or conceptual categories. Here is Evans's constraint:

> ... if a subject can be credited with the thought that a is F, then he must have the conceptual resources for entertaining the thought that a is G, for every property of being G of which he has a conception. This is the condition that I call 'The Generality Constraint' (Evans 1982: 104).

(To adapt the constraint to the use for which it is proposed here, suitable limitations to its generality would also need to be made. For example, it is not necessary that one be able to entertain the thought that x is G for *every* property of being G of which one has a conception, but only for some selected subclass of such properties of which one has a conception.) According to this constraint, what distinguishes true conceptual thinkers from "mere responders or information processors" (Gillett, 1987, p. 20) is that predicates available to the thinker must stand in contrastive relations *for the thinker* to other predicates available to the thinker:

> Even readers not persuaded that any system of thought must conform to the Generality Constraint may be prepared to admit that the system of thought we possess – the system that underlies the use of language – does conform to it. (It is one of the fundamental differences between human thought and the information- processing that takes place in our brains that the Generality Constraint applies to the former but not to the latter...) (Evans, 1982, p. 104, n. 22).

The suggestion that Evans's constraint provides is that, while both perceptual and conceptual categories are abstractions, only conceptual categories must satisfy the Generality Constraint. For a subject to have a concept, hence to be able to use a general term in a predicative judgment, the subject must satisfy the Generality Constraint (suitably modified) with respect to that concept or term.

Here, then, are some tentative consequences of the general proposal I have here considered. One may satisfy the Generality Constraint (suitably modified) with respect to some conceptual categories but not others; such, one can suppose, is the plight of each of us. And one might fail to satisfy the constraint with respect to certain categories with which one might nevertheless be able to perform sorting tasks; such was K.R.'s plight in particular. When this is the case, one has a perceptual category but not a conceptual category. Members of other species may not satisfy the constraint for any categories, lacking conceptual categories entirely, but be able (as surely most must) to perform sorting tasks on the basis of non-conceptual perceptual categories. Some conceptual categories have noteworthy relations to perceptual categories: the conceptual category of animal to the perceptual category of dog, furniture to chair, vegetable to carrot; the conceptual category of

equiangular triangle to the perceptual category of triangle. One consequence can be stated less tentatively: neither conceptual not perceptual categories can be generated mechanically from sensory processes alone.

References

Bealer, G. 1982 *Quality and Concept*, Oxford: Clarendon Press.

Bealer, G., Mönnich, U. 1989 "Property Theories", in D. Gabbay and F. Guenther, eds., *Handbook of Philosophical Logic*, vol. 4, Dordrecht; Boston, MA: D. Reidel, pp. 133–251.

Evans, G. 1982 *The Varieties of Reference*, John McDowell (ed.), Oxford: Clarendon Press.

Gillett, G. R. 1987 "The Generality Constraint and Conscious Thought", *Analysis* 47, 20–25.

Hart, J. Jr. and Gordon, B. 1992 "Neural Subsystems for Object Knowledge", *Nature* 359, 60–64.

Kanizsa, G. 1979 *Organization in Vision*, New York: Praeger.

Medin, D. L., Barsalou, L. W. 1987 "Categorization Processes and Categorical Perception", in Stevan Harnad (ed.), *Categorical Perception*, Cambridge: Cambridge University Press, pp. 455–490.

Miller, G. A., Johnson-Laird, P. N. 1976 *Language and Perception*, Cambridge, MA: Harvard University Press.

Nolan, R. *Cognitive Practices*, Oxford: Blackwell 1994.

Rosch, E., *et al* 1976 "Basic Objects in Natural Categories", *Cognitive Psychology* 8, 382–439.

Weiskrantz, L. *et al* 1974 "Visual Capacity in the Hemianopic Field Following a Restricted Occipital Ablation", *Brain* 97, 709–728.

Naturalizing Intentionality through Learning Theory

J. Proust

Dretske's endeavour – to give a naturalistic account of intentionality, i.e. of the representational capacity of an entitity – is currently considered as being among the most interesting and innovative proposals, and rightly so. What is particularly interesting is that Dretske wants his theory of intentionality to bridge the "explanatory gap" characteristic of theories of intentionality which do not derive the causal powers of a system from its internal states intentionally described.

I will concentrate here on the latest version of Dretske's account of intentional states in terms of learning (Dretske 1988). Insofar as it relies heavily on a specific approach to learning theory, i.e. Skinner's so-called "operant conditioning", Dretske's account may miss essential features while giving undue importance to characteristics now – in contemporary cognitive theories of learning – considered inessential; in particular, too an important a role seems to be devoted to the concept of immediate reinforcement of behavior. The very idea of recruiting an internal state as a cause of a bodily movement seems to have to be revised in substantial ways.

One of Dretske's aims in his theory of intentionality is to explain why "meaning itself (and not just the structures that have meaning) is supposed to play an important role in the explanation of an *individual*'s behavior (as beliefs and desires do)" (Dretske 1990a: 14). To explain this, Dretske suggests that one has to look to a restricted set of intentionally characterizable behaviors, i.e. to those in which "meaning is instrumental in shaping the behavior that is being explained". Dretske's proposal consists in articulating two levels of causal relations. At the first level, there is

(i) a nomic correlation between an external condition F and an internal state C in virtue of which C indicates F; this can be stated more exactly, following Kim's corrections (Kim 1991), by saying that a token-state C with the neurobiological property N is caused by F in S's vicinity at the time;

(ii) a causal connection between the internal token-state C with property N on the one hand, and a (token-)motor output of kind M. This causal connection ensures that C is not simply an indicator of F (which only carries information about F), but is also a representation of F. For as

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 233-245

Dretske (1988) shows, a representation is an indicator whose function it is to indicate.

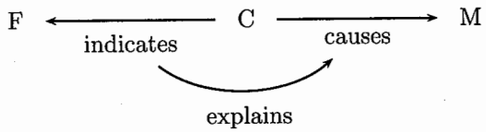
The work to be done at the second level consists in connecting the two preceding first-level causal connections together through a second-order causal link. At that level, it must be shown that it is in virtue of having informational content F that the token-internal state C triggers the bodily movement M. The notion of learning here provides the required causal link: learning is supposed to demonstrate that it is in virtue of the nomic correlation (i) that C represents F, and that because of that representational content C was recruited to cause M. C caused M because M was advantageous to the organism when F occurred. Therefore C's indicative content was given the function to cause M, and both:

1) explained misrepresentation through the distinction between actual indication (natural meaning) and representation (functional meaning);

2) explained the recruitment of C as a causal link in motor output.

In other words, the fact that a C-token represents F through its property N is a structuring cause of C's (by having N) causing M. It became a structuring cause because this recurrent association between an event in perception leading to detect an F and a bodily movement appropriately related to there being an F was *reinforced* by instrumental conditioning.

These considerations are summarized in the following figure (Dretske 1988: 84).



I will only mention here a difficulty of this account when one tries to make completely explicit the type/token distinction in this model. As we saw, it is in virtue of their sharing neurobiological property N that the various C-tokens are able to cause M. But in fact what do they cause? They each cause a token of motor output, i.e. a particular bodily movement. But notoriously, it is hopeless to try and give an exhaustive list of tokens of bodily movements which could qualify as belonging to the type of the anticipated motor output M. On the other hand, one cannot describe the *type* of bodily movements involved in the motor end of a behaviour without using some general characterization of the object which the action considered aims at grasping or fleeing. But the solution seems to be to discard any verificationist stance to the effect that what there is depends circularly on what we can know about what there is. There may be a motor output type with a property P *that we don't characterize*, such that each bodily

movement instantiating it reinforces the causal link beween the internal C-token and a token of bodily movement with property P and thereby ensures the recruitment of C as causing M.

Now according to Dretske, the explanatory fact through which *genetically determined* behaviors happen to relate to external conditions is not located at the appropriate second level indicated in this figure. In a genetically determined behavior, the internal state responsible for the motor output was recruited not because of the presence in the environment of S of events with properties that the individual learnt to use, making thus the internal indicator into a representation; but because the past continued link between events of type F and advantageous reactions M fixed in the species a mechanical, hard-wired, F-C-M sequence in which the representational contents of C-(present) tokens play no role.

For example, noctuid moths have receptors designed to respond to high-frequency sounds, the latter being normally emitted by a bat. The response is either turning away from the source, in the case of low frequency sounds, or diving and spiralling in the case of high-frequency sounds. Evolution theory explains why internal states of the frequency detectors cause the animal to turn away or to dive: this mechanism confers on the moth a competitive advantage. But it is not able to explain why this particular moth does whatever it does; in other words, the selectional explanation here says whatever there is to say. No particular fact about this moth can help explain why it does what it does. The reason is that the relation of indication holds essentially between a type of internal state and a type of condition which has long enough been part of the moth's environment to be depended on as what the genetically induced response responds to. If the external environment suddenly changed, the moth would still respond to high frequencies in the same way, and the relationship between this internal state C token *and the external world* would cease to explain why the moth does what it does. In other words, C's informational content is causally idle. This kind of behavior is a tropism, which fails to give *reasons* for the token behavior, because "what the indicators indicate is irrelevant to what movements they produce" (Dretske 1988: 94). Again, the system does not move as it does now because of what C means about external condition F, but because of some mechanical or feedback process which cannot be modified by learning. One cannot then attribute the behavior of the organism to any reason *it* could entertain.

In order to find a "genuine case" of intrinsic intentionality, we have to look for a case in which the behavior is actually shaped (or structured) by the relation of indication holding between an internal state and an external condition. My strategy in the present paper will be to take this opposition between "genuine intrinsic" and extrinsic intentionality at face value and show what the notion crucial for getting intrinsic intentionality – learning – involves in turn, which kind of competences and functional equipment it minimally presupposes. I will nevertheless reexamine in a later work the

plausibility of matching the strict opposition between intrinsic and extrinsic intentionality, with the opposition between developmental and selectional arguments.[1]

One can find cases of intrinsic intentionality, Dretske says, in places "where individual learning is occurring, places where internal states *acquire* control duties or *change* their effect on motor output as a result of their relation to the circumstances on which the success of this output depends" (Dretske 1988: 95). Then how should we spell out the conditions for such an individual learning? Dretske dismisses the cases of elementary conditioning, i.e. habituation and sensitization, for they are mediated by peripheral mechanisms, and do not seem to rely on any particularly stable internal condition.[2]The first condition he proposes is "having the appropriate sensory capacities":

> The rat must be able to hear, able to distinguish one tone from another, if it is to learn to respond in some distinctive way to a particular tone. The pigeon must be able to *see*, to distinguish visually, one color from another if it is to learn to peck when the light is red. (Dretske 1988: 97)

This first condition bears explicitly on sensory receptors – such as sensory capacities: being able to be affected by visual or auditory information, and being able to discriminate between various classes of stimuli. Does such a condition amount to "being equipped with an F-indicator"? So says Dretske (1988: 98). Sensory receptors allow a system to form indicators in a highly modular form of information processing. But so far these indicators do not yet have a representational *function*. They are necessary, and not sufficient, conditions for contentful inner states.

In (Dretske 1990b), Dretske distinguishes between sense and cognitive perception. Our sense perception of objects is direct and unmediated; not in the sense that no information processing is necessary to build a representation of these objects, but in the sense that the objective, world-to-mind, causal link between the object represented and the internal state is all that there needs for it to be the case that the subject sees, e.g. a cat. Mediation comes in only at the level of "cognitive" perception, in which one comes to see an object as what it is (or looks like). It is clear for the present that what is needed for a detector to acquire motor control is the cognitive variety of perception, which allows the system to detect an F *as* an F. For simply seing an F, without having the ability to categorize it as such, would not allow the system to form a stable internal indicator *with its functional role* in controlling motor output.

Thus Dretske seems to be committed to the view that detection belongs to the sensory part of perception, whereas categorization belongs to the

[1]See on this question the stimulating objections in Garcia-Carpintero "Dretske on the causal efficacy of meaning" (in press), pp. 18ff.

[2]For a different view on this question, see Hawkins & Kandel (1984).

cognitive part. While a rat may well see a condition F around him, he will only see it *as* an F if he has learned how to use F in some way (Dretske 1993: 196–7). Dretske's first condition only states that learning to see it *as* an F presupposes that an F can be perceived in the sensory mode.

Let us see whether the second condition may bring us all the functional structure which is needed for an internal representation to be built. The second requirement is "harnessing this indicator to effector mechanisms in such a way that appropriate movements are produced when and only when the indicator positively registers the presence of condition F". (Dretske 1988: 98) This harnessing, however it is achieved, appears to be what is essential for learning to occur: "Learning cannot take place *unless* internal indicators of F are harnessed to effector mechanisms in some appropriate way. Since this learning *does* occur, the recruitment *must* take place". Learning of this kind, known as "instrumental" or "operant conditioning" is thus taken to represent the essential core of learning: it consists in recruiting certain information-carrying internal states for control duties in virtue of what they mean, i.e. of what they indicate about environmental conditions.

There are several questions one is tempted here to ask Dretske. First, is operant conditioning a good representative of learning? Why choose this one and not customary Pavlovian conditioning? Second, is it not relevant to look at "how the nervous system manages to accomplish this trick"? Maybe after all only a nervous system endowed with certain other intentional properties is able to display this capacity. In such a case, the learning condition brings in much more than the two causal steps indicated above. I will leave these questions aside for now, and concentrate on two "relatively unproblematic facts", which, according to Dretske, are needed to substantiate his claim about the central role of learning. Indeed those two facts are simply a restatement of the two hypothetical steps just given.

One is Thorndike's Law of Effect, according to which successful behavior tends to be repeated. One may feel reluctant to accept Dretske's observation that it is "not important that we get clear about the exact status of this law" – a law which has been considered tautological and thus void of any empirical significance insofar as it only states that what increases the probability of a behavior (a reward) tends to increase it. A reward being defined by its operational role, there is no more to the Law of Effect than what we put in this definition and later use to describe behavior. Again Dretske rejects any epistemological discussion concerning the status of Thorndike's Law only because he is only interested in there being some kind of contingency of one event – a response of pecking – on another – a stimulus or reward. But surely there must be some interpretations of Thorndike's Law of Effect which are incompatible with the attempt at reducing intentionality to an explanatory (causal) link between two causal chains. For example, if it could be proved that the very description of a behavior – as contrasted to a bodily movement – already involves an intentional content, or if what counts as "the same circumstances" can be stated only given a certain interpreta-

tion of what it is, for this organism, to use a particular piece of information, Thorndike's Law of Effect would do no more than beg the question of intentionality. By no means would it provide us with a means of accounting for it.

The other fact is very briefly recalled : "Such learning requires, on the part of the learner, a sensitivity to specific conditions F.... There must be something *in* the animal to 'tell' it when conditions F exist". (1988: 101). It is clear that the mystery of Thorndike's Law extends to this requirement: what is it in the animal which "tells" him whether condition F obtains? We saw earlier that to be affected by some feature of an object (to detect a change in the environment) does not amount to being able to use that feature in categorization; if "telling" implies explicit identification, then by Drestke's own standards (in Dretske 1990b), no such output can be expected at the informational level of sensory perception. The philosopher's (and psychologist's) job is to determine what is required to get feature identification, over and above pure detection. How does the task modify the weight given to such and such a feature dimension? Is categorization sensitive to decision? How is memory involved? In what kind of representational format is the information stored? How and in which respects is motivation causally involved?[3] Undoubtedly, it takes a lot more for the internal state to "tell" something about an external condition F than simply having been recruited to produce the bodily movement contingent upon F's being the case. We begin to see that we have to endow our intentional system with a capacity to "cognitively" perceive, i.e. categorize, memorize, plan and be motivated.

The "two relatively unproblematic facts" therefore seem to raise a substantial number of philosophical, epistemological and psychological questions. It is obviously crucial for Dretske's purpose of naturalizing intentionality to give a detailed account of what happens in learning, when the main challenge is to exclude from the causal account of representational states any process already involving contentful states. Speaking of "contingency" for example would not be radical enough, if the very causal efficacy of what counts as a stimulus was shown to presuppose the intentional core to be explained. It is generally admitted that a realistic approach to properties allows us to account for what a bodily movement is contingent upon, properties being out there in the outer world independently of what the organism – or, for that matter, of what anyone – may know about them.[4] Another potential difficulty for the "contingency" approach lies in identifying the dynamics of the coupling between a given stimulus and a response. Would a behavior be adequately explained if nothing else was said besides the fact that the animal

[3]By the end of the chapter, Dretske seems to wave in this direction, by explaining that the function of the internal indicator is not exactly to produce M, but to produce M "on the right motivational and conative conditions" (Dretske 1988, 105).

[4]This realistic move was used above as a rejoinder to the verificationist point that the role of the bodily movement type which counts as reinforcer for the C-M connection needs circularly the understanding of the meaning of F to be expressed.

learnt to associate a response with a given condition? Various facts about conditioning show that such an atomistic approach to learning is untenable.

But let us concentrate now on the use which Dretske makes of *instrumental learning.* As we saw earlier, Dretske's concept of having a mind involves much more than carrying out adequately various informational tasks. No mind, he claims, without internal states able to acquire control duties or modify the existing control system. Only in that case can the internal indicators be given a function – which transforms them into representations; only in that case can we have an explanatory/causal relation between what there is out there, the internal state being activated and what the animal does.

Let us observe here that the role of the bodily movement does not consist in providing us with an *overt* proof of the existence of the internal indicator: Dretske's claim does not belong to the verificationist family. It does not assert simply that only behaviorally relevant internal states can be recognized as having a meaning. His claim is rather that the bodily movement is intrinsically intertwined with the internal state's acquiring a function, i.e. a representational content; no function of an internal contentful state, in other words, without some kind of control of the behavior of the organism: this is here a conceptual, and not an empirical truth. My strategy in what follows will be to question the conceptual, constitutive character which is given to motor activity in defining the function of an internal indicator. I will do so, mainly, by using empirical data showing that motor output has no essential role to play in learning. Learning in Dretske's theory "shapes the causal role" of the indicators and thereby endows them with a specific representational content. It is crucial again that this Skinnerian shaping results from the reinforcement of some bodily movement, which is supposed to be initially produced as part of a random general activity. Recruitment in this view may be taken as an independent, atomistic process, whose actual implementation may be left to neurophysiological discussion. I will suggest that there are strong empirical reasons to reject this atomistic view of learning. But I will first question the contingency of learning on "timely reinforcement of certain [motor] output".

It is an essential feature of Skinner's so-called operant conditioning that only the *rate of responding* and the *probability of response reinforcement* are causally involved in learning. But recent work in animal learning (Gallistel 1990: ch. 11) shows that the response as well as the probability of its reinforcement are "largely irrelevant" in learning. In other words, it is not what the animal does which affects its future behavior; moreover, whatever reinforcement it gets (with which probability) is not the kind of factor which *causes* learning. What causes learning is what the animal in the right motivational state is allowed to *observe*, regardless of whether he indeed benefits directly from his observation, and his behavior is only relevant insofar as it may influence what he is able to observe.

A number of experiments suggest that the correlation between the rel-

ative amounts of time spent foraging in particular locations and the relative rates of reward occurrence at those locations is only a special case of a more general law in which the variable is *net food availability*; the latter is defined as the product of the average magnitude of the observed morsels and the average rate at which they are delivered. For example, Gallistel reports Neuringer's experiment (Neuringer 1967) in which it is shown that pigeons in an operant conditioning situation chose one among two illuminated keys with a percentage of choice corresponding to the relative total access to reinforcement *for a key*, the latter variable being defined as the number of obtained reinforcements on a key multiplied by the duration of each reinforcement (Gallistel 1990: 363). Animals seem to "compute the product of morsel magnitude (or access duration) and the number of morsels (accesses) per unit time". What is thus being learnt by animals in operant conditioning is not a simple relationship between a response and a state of the world, "condition F", but a certain dynamical fact involving the rates of returns of different foraging patches, whether actually used by the animal or not. It thus seems that there is no direct, atomistically established relationship between, on the one hand, the fact that a certain internal fact indicates a condition F, and, on the other hand, the fact that it is being recruited to trigger some bodily movement.

Before examining whether some extended version of learning accommodating these new findings would still do the job which Dretske's theory requires, I now turn to the very notion of "recruiting" which holds between the internal state C carrying the information that F obtains, and some bodily movement it has been recruited to cause. In the classical conditioning studies, it was assumed that there was a simple mapping between the strengths of association and observed behavior. Typically, the bodily movement which the internal state C is recruited to cause is just some bodily movement allowing the organism in question to perform some unconditioned response (like attacking, salivating, pecking, fleeing etc.). In Dretske's most recent statements of his own theory, it appears clearly that recruiting a movement is definitional or constitutive for an internal state being representational, but also that it is causally operative in the acquisition of the indicator function: the model has to provide the causal mechanism which explains why the function was established in the first place. Saying that there is some way of establishing the function would not be enough if step two of the model is to be adequately carried out. It could well be that although motor control is definitional for function, no such function can be causally established. In other words, one has to show the very possibility for a motor output to help recruit an internal indicator in its control function, in a way which is purely causal.

> Beliefs have the function of indicating, a function which they acquire during learning by causing M – but it is not their job to cause M. (Dretske 1991: 115)

Now contemporary theorists of learning believe that no single theory

may account both for acquiring internal indicators and for exploiting these indicators in action. To develop a theory of behavior, it is assumed, one first has to have a theory of learning. The question to know "how what is learnt gets translated into what is done" is highly complex and presupposes that a theory of learning is already at hand (Gallistel 1990: 388). To this one might object that the philosopher may idealize and give an abstract scheme for possible empirical theories. But the point of giving a naturalistic theory of intentionality seems to require that the elements included in the scheme are not only plausible, but also necessary and sufficient conditions for what they are supposed to achieve. In other words, it would not be enough to state that acquiring a function in controlling motor output transforms the indicator into a contentful structuring cause; it must also be shown that there is no causal gap between the two segments of first order causality. But the difficulty of the link suggested by Dretske is that acquiring a function of indication and using the indicator in a behavior are two independent processes as far as causality is concerned.[5]

I will only give here one example showing that the crude, atomistic, "on-line" notion of operant conditioning cannot be the proper idealization for "an internal state acquiring the function to indicate". A much weaker notion of "recruitment" could be built in dispositional terms – but from such a notion, as we will see, the concept of a representational function ceases to be causally explained in the required, backward sense of the term.

The notion of a "cognitive map" occurs often in Dretske's work. I will examine one concrete case of such a map being built and put to use. In the current use of the term in cognitive science, a cognitive map is some record in the brain of geometric relations between surfaces in the environment used to plan movements. Gallistel (1990) reports studies on birds migratory navigation showing that certain species, like indigo buntings, have to have learnt a stellar map as unfledged nestlings (i.e. twhile they are still in the nest, in their first spring) to be able to take off in the appropriate direction. It was shown by Emlen (1967) that the birds only learn the configuration of certain circumpolar stars, which are seen the year round in temperate latitudes. Now in this case the relationship between the indicator acquired and the bodily movement is particularly intricate: hormone treatment allows one to manipulate the direction to which the bird will take off. But one has also here a case in which "there is no necessary connection between the behavioral context in which information is acquired (the context in which learning occurs) and that in which the animal makes behavioral use of the information (the context in which what was learned becomes manifest in what is

[5]Dennett also has a theory of intentionality which involves an external event and a behavior: "What an event or a state 'means to' an organism also depends on what it does with the event or state". (Dennett 1969: 76) Reference is given on the stimulus side, while sense is given on the efferent side: behavior guarantees that the ascribed intentional content is content for the organism. But Dennett's account does not require an extra-causal link between the two.

done)" (Gallistel 1990: 86). It is difficult not to count stellar orientation as learning that a certain condition holds. Some experiments show that indeed a bird raised in a planetarium is able to respond to other, counterfactual star orientations. And he does use this information when in the appropriate physical and motivational states.

It should be clear that these kinds of examples – which abound in animal psychology – undermine the early behavioristic assumption "that reinforcement is central to learning, that the *immediate* beneficial consequences of particular behaviors stamp into the nervous system the circuit changes that increase the likelihood of future performance of the same behavior". At the time when learning occurs, the bird has no feathers, no competence to fly. At the time when he can fly, he can't learn any more about star configurations. Clearly, this example is not "Dretskian" for a second reason: the relationship between the indicator and what it indicates has to be mediated by still other external referents and other internal states. Gallistel observes that "The center of rotation of the sky cannot be derived from a single look": no single pairing between an internal state and a condition F, on the one hand, and that state and some bodily movement, on the other hand, can be sufficient in explaining why the animal learns what it does. One could argue that my example does not show that other, more elementary kinds of learning, do not occur. My answer will be to show that even the elementary, so-called associative learning, in which one external condition is learned to predict a second one, requires a network of other capacities, and cannot be modelled as a pure pairwise association.

Again the contemporary models of learning shatter the atomistic conception of pairing inherent in the traditional model. Let us consider an unconditioned stimulus US, and let us pair it with a conditioned stimulus SC1, say, a tone. Now let us try and pair SC1 with another conditioned stimulus, say, a light. For a long time, theorists believed (1) that animals adapted their behaviors to temporal contingencies in their environments; (2) the associative strength was supposed to represent all there was for an animal to learn, in endowing it with the capacity to predict an unconditioned stimulus from a conditioned one: for example to predict a prey from its natural indicators. In that view, learning is triggered by individual events in the world. (3) It was further supposed that it is the probability of US's occurrence which determines causally the strength of the associative bond.

Unfortunately, these three assumptions are disconfirmed. First the various contemporary models of associative learning posit the *multidimensionality of the memory state space* in which learning occurs. In other words, associative strength depends on the various associative links *already established* with the unconditioned stimulus in the system. For example, the Rescorla-Wagner model (Rescorla & Wagner 1972) states that any change in the associative strength of a given CS depends on the amount of associative bonds already established with the stimulus *and* on the specific saliency of the given stimulus dimension. The more numerous associations there are,

the less associative force is left for new associations to be established. Therefore, assumption (2) is at fault as well: secondly, learning involves more than pairing events, and forming associated representations. It is shown that animals which learn regularities in their environments are able to record the rates of occurrence of the events. Thus organisms represent the external world not initially by discrete states corresponding to discrete events, but by rates of occurrence, which themselves are found to be canvassed as "the consequences of Poisson processes" (Gallistel 1990: 422).[6] Finally, it is false that the probability of occurrence of the US is what determines associative learning. The key notion in classical conditioning is probability of response. We will see that one does not get fine-grained correlations between associative strengths and responses using this notion, whereas the idea of response *rates* – which involves memorization of time lapsed between stimuli (or ISI) *and* between trials (or ITI) – does a better job. The rate of conditioning depends thus on the ITI/ISI ratio rather than on the ISI alone.

Dretske's model fails to take into account these findings in animal and human learning. They are nevertheless relevant to the extent that they suggest that no simple causal connection can allow motor output to fix the representational function of an internal indicator. Two points seem of particular importance.

First, it is not plausible to maintain that the relation of indication is *atomistic*, at least in the sense in which a large functional substructure has indeed to be involved in order that a certain internal computational state relates adequately to variations in the world. What then is relevant for an indicative function to be established is not a one-one, event-internal state pairing, but a mapping between *the variations on external contingencies* and *their internal computational counterpart*, expressible by changes on preformed variation patterns. Philosophically speaking, the idea is to make intentionality a property of a dynamic, ever-adjusting, system – an idea which is explored, although only superficially, in Dretske's own work. Indicating has to occur in space and time, and requires as well capacities for representing and storing facts about space and time.[7] In other words, it seems that learning is only possible on the assumption that the information stored by the system is multi-layered. In Gallistel's terms, "learning is inherently hierarchical" (Gallistel 1990: 87). The indigo buntings studied by Emlen, for example, had to store information about circumpolar stards in their various successive positions to be able to locate the center of rotation of

[6] This assumption is the default; it corresponds to the case in which there are many uncorrelated causes, giving rise to a random rate process. See Gallistel 1990: 422ff.

[7] The pressure for introducing history in intentionality surfaces at several points in Dretske's account, in particular in the contrast he makes between selectional and developmental explanations (Dretske 1988: 92), and in his appeal to associative learning (Dretske 1986) But he fails to see the tension between his concept of indication, which applies to whole types of events, and the need for some time-dependent account of associative linkage, which should make prediction possible in cases when tokens of events are correlated in a less than lawful way.

the planets. The same holds for conditioning: at a low level, an animal has first to record the time of occurrence of events, to be able later to compute a prediction function of the occurrence of the unconditioned stimulus.

It may then be that individual learning always involves, at least at the lower levels, some innate equipment which, according to Dretske's criterion, does not qualify as being endowed with intentionality. Exploring the relationship between genetically acquired response schemas and individually acquired intentional contents seems one of the main problems which naturalistic philosophers have to put on their agendas. I will try myself to address it somewhere else.

Second, the *explanatory link* between indication and behavior is not grounded in learning (which insures the internal state being *recruited* for performing a particular bodily movement), but in the informational content *per se* plus functional architecture (including motivational states).

Contrary to what Dretske's model suggests, bodily response does not have a causal role in shaping the indicator function of internal states. Then what does? One could perhaps say that causality flows in the other direction: internal states contribute to explaining behavior in virtue of the informational content which they carry. The previous discussion seems to point to a dispositional conception of behavioral responses: bodily movements do not bring about by themselves any belief fixation; they are parts of an action to which the animal may be disposed by way of a set of antecedent beliefs and desires.

What informational content internal states have is not independent of the capacities of the system *as a whole* to extract the relevant information. And these capacities seem to be present in the organism without any previous learning. These innate perceptive and motor skills may well put top-down constraints on the causal linkage between, on the one hand, the indicator formation, and, on the other hand, the "motor connection" (from meaning to bodily movement). If this is true, then the philosopher would have to reconsider his theory of what a "reason for behaving" is.

If what insures the stability of an internal state as an indicator for some external event (or relation between events, or a dynamic function on either of these) is not an advantageous motor output, but some systematic integration of that state in a set of other functional states, then the final philosophical puzzle will be to elude the threat of circularity. To show, that is, that none of the postulated subserving skills involve themselves any hidden intentionality of a derivative variety. This may prove an impossible task.

References

Baker R.L. 1991 "Dretske on the Explanatory Role of Belief", *Philosophical Studies* 63, 99–111.

Dennett D. "Ways of Establishing Harmony", in E. Villanueva (ed.). *Information, Semantics and Epistemology*, Oxford: Blackwell, pp.18–27.

Dretske F. 1986 "Misrepresentation", in *Belief*, ed. R.J. Bogdan, Cambridge MA: MIT Press.

Dretske F. 1988 *Explaining Behavior: Reasons in a World of Causes*, Cambridge MA: MIT Press.

Dretske F. 1990a "Does Meaning Matter?" in E. Villanueva (ed.), *Information, Semantics and Epistemology*, Oxford: Blackwell, pp. 7–17.

Dretske F. 1990b "Seeing, Believing and Knowing", in D. Osherson (ed.), *Visual Cognition and Action* vol. 2, Cambridge MA: MIT Press.

Dretske F. 1991 "How Beliefs Explain: Reply to Baker", *Philosophical Studies* 63, 113–117.

Dretske F. 1993 "The Nature of Thought", *Philosophical Studies* 70, 185–199.

Emlen, S.T. 1967 "Migratory Orientation in the Indigo Bunting, *Passerina cyanea*. Part I. Evidence for Use of Celestial Cues", *Auk* 84, 309-342.

Gallistel R.C. 1990 *The Organization of Learning*, Cambrige MA: MIT Press.

Garcia-Carpintero S.-M. 1993a "Dretske on the Causal Efficacy of Meaning", draft.

Garcia-Carpintero S.-M. 1993b "The Teleological Account of Content", draft.

Hawkins R.D. & Kandel E.R. 1984 "Is There a Cell Biological Alphabet for Simple Forms of Learning?", *Psychological Review* 91, 375-391.

Kim J. 1991 "Dretske on how Reasons Explain Behavior", in B. McLaughlin (ed.), *Dretske and his Critics*, Oxford: Blackwell.

Rescorla R.A., Wagner, A.R. 1972 "A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement", in A.H. Black & W.F. Prokasy (eds.), *Classical Conditioning* II, New York: Appleton-Century-Crofts.

Troubles with Heterophenomenology

Eduard Marbach

The problem of consciousness is back on stage in present-day cognitive science. As someone with a strong background in Husserlian phenomenology, which in a first approximation can be seen as a descriptive science of the phenomena of consciousness based on reflection, I am very happy with this resurgence of interest in consciousness. Here and there I already detected more or less explicit and elaborate references to and discussions of Husserl's work and its significance in the context of this new development - with mixed feelings, I might say.[1]

In this paper I wish to engage in a discussion of Daniel C. Dennett's proposal of what he calls a *heterophenomenology* devoted to the empirical, scientific study of the phenomena of consciousness.[2] To get right away a feeling for what hetero-phenomenology as "a method of phenomenological description" (Dennett 1991: 72) aims at, the following comment from Dennett (1993c: 211) is helpful:

> The original use of the term "phenomenology" was to mark the cataloguing of everything that happened regarding some phenomenon, such as a disease, or a type of weather, or some other salient source of puzzlement in nature, as a useful preamble to attempting to explain the catalogued phenomena. First you accurately describe the phenomena, as they appear under all conditions of observation, and then, phenomenology finished, you – or someone else – can try to explain it all.
>
> So my heterophenomenology is nothing more nor less than old-fashioned phenomenology applied to people (primarily) instead of tuberculosis or hurricanes: it provides a theory-neutral, objective catalogue of what happens – the phenomena to be explained. It does assume that all these phenomena can be observed, directly or indirectly, by anyone who wants to observe them and has the right equipment. It does not restrict itself to casual, external observation; brain scans and more invasive techniques are within its purview, since *everything that happens in the brain* is included in its catalogue of what happens.

[1] See, e.g., H. Dreyfus 1979: 65; 1982: 9ff.; 1992: xvii; R. McIntyre 1986 who, while rightly criticizing some of the more implausible claims in Dreyfus (1982), unfortunately still appears to attribute a so-called "methodological solipsism" to Husserl's enterprise (101f.); D. Münch 1990; Chokr 1992, with whom I largely sympathize.

[2] In a number of publications over the last 15 years, Dennett has forcefully argued for heterophenomenology; see, e.g., Dennett 1978, 1982, 1987, 1991, 1993a, 1993b, 1993c.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 247-264

A discussion of Dennett's method is of quite some importance to me and seems philosophically instructive, because, on the one hand, I am very much in favor of Dennett's attempt to devise ways of making the scientific study of the phenomena of consciousness possible and even attractive for scientists (see Marbach 1988, 1993). Moreover, interestingly, for my purpose, Dennett himself likes to introduce his carefully elaborated heterophenomenology by pointing out that "the method has close kin in the history of philosophy and psychology" (see Dennett 1982: 159), especially in the traditional form of phenomenology as represented in the Brentano-Husserl line.[3] On the other hand, however, it does seem correct to say that, on Dennett's view, his own method of heterophenomenology deserves to *displace* the traditional form of phenomenology.[4]

And this, of course, is not exactly what some of us would like to happen, for it could readily amount to falling back behind a wealth of results of seriously conducted traditional phenomenology. Dennett's proposal presents thus a healthy challenge to Husserlian phenomenology from the standpoint of cognitive science. To be sure, my present attempt at meeting Dennett's challenge cannot be much more than seeking to pave the way for *using*, rather than letting behind, Husserl's work in contemporary investigations into the phenomena of the conscious mind.

I see shortcomings of basically two sorts in Dennett's proposal which lead me to speak of troubles with heterophenomenology. First, Dennett seems to me to underrate the potential of traditional phenomenology, and this leads him – too hastily, in my view – to the proposal of displacing the traditional way of doing phenomenology. Second, as I perceive Dennett's proposal of heterophenomenology, there is the following philosophically more intriguing shortcoming, which, given the difficulty it turns upon, I would like to put in the form of a question: Is the very complexity of any method of *hetero*-phenomenology, taken as a method for studying the phenomena of consciousness, not underestimated, *if*, as Dennett appears to be doing, the method is presented on a par with "phenomenology" as applied to phenomena in the public domain of the natural world (see, e.g., Dennett 1993b: 153; 1993c: 211; 1993a: 50; 1991: 96)? For does it then not look as if heterophenomenology were to be brought to bear from the third-person perspective independently of taking into account the interpreter's own original perspective on the type of phenomena under study? This question could of course be examined quite apart from the shortcoming concerning Dennett's understanding of the Brentano-Husserl tradition alluded to first. For example, in his first *Critique*, Kant (1781: A 353) already has observed this: "It is obvious that, if one wants to have an idea of a thinking being, one must put oneself in its place and thus substitute (*unterschieben*) one's own subjectivity for the object which one wanted to consider (which is not the case in any other kind of investigation)". Similarly, I take it that contemporary philosophers such as Thomas Nagel, John Searle and others, each in his own way, have addressed basically the same philosophical difficulty quite independently of Husserlian considerations which I try to bring to bear in what follows.

[3] See Dennett 1991: 44; compare already Dennett 1978: 184. Regarding the link to the Brentano-Husserl tradition, see in particular also 1982: 159; 1987: 153ff. – Dennett likes to spell the word 'phenomenology' with an upper-case 'P' when he wishes to refer to the discipline of traditional phenomenology (see, e.g., Dennett 1991: 44); by contrast, following "recent practice", he adopts the term 'phenomenology' "(with a lower-case p) as the general term for the various items in conscious experience that have to be explained" (1991: 45, 65, et passim). In the present paper (except where I literally quote Dennett), I will refer to items in conscious experience by using the term 'phenomena' and I will reserve the term 'phenomenology' (with a lower-case p) to refer to the theoretical enterprise of studying such items (in analogy to 'zoology' as the discipline of zoologists studying animals; cf. Dennett 1991: 43ff.), and of course use Dennett's term 'heterophenomenology' where his version of such an enterprise is meant.

[4] See, for example, Dennett's paper "A Method for Phenomenology" at the *Colloque international, L'intentionalité en question. Entre les sciences cognitives et le renouveau phénoménologique* (Nice, June 1992) , where Dennett presented his "brand of phenomenology" by contrast to traditional phenomenology as being "in any case a close and worthy cousin – it deserves to displace the traditional form" (personal notes).

First I will briefly present Dennett's reasons for seeking to displace traditional phenomenology. By singling out what I take to be the most relevant criticism addressed to phenomenology in our context, I then try to meet the challenge by sketching how I think Husserl conceived of the study of the phenomena of consciousness.[5] Finally, with reference to examples taken from Dennett (1991), I will try to delineate how a philosophically more satisfying heterophenomenology would have to be worked out, illustrating thereby the troubles with heterophenomenology as it now stands.

1 Dennett's Main Criticism of Traditional Phenomenology and a Sketch of an Answer to the Challenge

No doubt the most important shortcoming of phenomenology in Dennett's view (and he is not alone) can be succinctly put as follows: "The tradition of Brentano and Husserl is *auto-phenomenology*" (Dennett 1987: 153). In other words, Dennett is attributing to phenomenologists "a special technique of *introspection*" (Dennett 1991: 44, my emphasis), or "some special somewhat introspectionist bit of mental gymnastics - called, by some, the phenomenological reduction" (Dennett 1987: 153). As he also puts it: "The standard perspective adopted by phenomenologists is Descartes's *first-person perspective*, in which *I* describe in a monologue (which I let *you* overhear) what *I* find in *my* conscious experience, counting on *us* to agree" (Dennett 1991: 70).[6]

Now, crucially, it is just in opposition to, as Dennett also has it, "the dubious introspectionist (*genuinely* solipsistic) method of the Phenomen-

[5] For some of the basic methodological tenets of Husserlian phenomenology, see, e.g., Bernet, Kern, Marbach 1993.

[6] For further passages presenting basically the same criticism of traditional phenomenology, see Dennett 1987: 154, 1991: 44.

ologists" that Dennett proposes a "third-person alternative": his *heterophenomenology*, where "we are concerned with determining the notional world of *another*, from the outside" (Dennett 1987: 157f., 153).[7]

For the moment, let us just retain the idea that Dennett's heterophenomenology is conceived as a method which, epistemologically speaking, is operating from the third-person point of view (see, e.g., Dennett 1991: 66ff.), by means of which Dennett aims at displacing traditional "introspectionist" phenomenology while still wanting "to save the rich phenomenology of consciousness for scientific study" (Dennett 1993a: 50).[8]

I am convinced with Dennett that an important modification and extension of the methodology of *pure* phenomenology is required for the envisaged purposes (Dennett 1991: 65), indeed, that something like a 'heterophenomenology' is needed.[9] I want therefore to say something regarding the charge of introspectionism which, especially if the charge of being a genuinely solipsistic method had to be accepted, would of course disqualify traditional Husserlian phenomenology forever from becoming part of a scientific enterprise.

To begin with it is worth recalling that Husserl himself repeatedly complained about the assimilation of phenomenology to a variety of psychological introspection. For example, in a text written in 1912, he speaks of "the basically perverted view that with phenomenology it is a matter of a restitution of the method of inner observation or of direct inner experience in general" (Husserl 1980a: 33). What does Husserl have in mind when he so adamantly rejects being associated with practicing a method of introspection or inner observation? As I understand the controversy, the main point to recall is that Husserl conceives of phenomenological analysis of conscious experiences in a mathematical spirit as reflection-based elaboration of the structures or forms of the experiences in accordance with their very possibilities, i.e. unconcerned with empirical matters of fact regarding the very phenomena under study (see, e.g., Husserl 1980a: § 8; 1987: 79f.). This emphasis on the conditions of the possibility of conscious experiences, as the matter may also be put, may well prove to be decisive in the present context; for it has consequences for one's assessment of such moot points as

[7] As Dennett explains, the term "notional" in expressions like "notional object", "notional world" etc. can be considered to correspond to "intentional", in Brentano's sense; see, e.g., 1987: 153.

[8] See also, Siewert 1993: 106ff. who has an interesting discussion of heterophenomenology, attributing an epistemological position to Dennett which he calls "third-person absolutism" and which Dennett (1993b: 153) accepts: "... I think he's got my epistemological position clear. Whatever the position is called, it is not a rare one. It is, in fact, the more or less standard or default epistemology assumed by scientists and other 'naturalists' when dealing with other phenomena, and for good reason".

[9] Husserl himself already conceived in outline how 'pure' phenomenology should, and could, be applied to empirical, scientific studies of consciousness, and he labelled this project 'empirical' or 'applied phenomenology', in analogy to the application of pure mathematics to physics (see Marbach 1988); Kern 1975: §15, §52 had already clarified the relationship between pure and empirical phenomenology.

privileged access, infallibility, incorrigibility, and even solipsism, which are commonly brought into play in connection with criticisms of the method of introspection (Dennett 1991: 67f.), criticisms which I do not think apply in the case of Husserlian phenomenology.

In order to appreciate the emphasis on possibilities in Husserl's method, one must clearly understand how much against entrenched habits and thus – unlike our practice of introspection – how counterintuitive phenomenologically reflective analysis of "consciousness of something" really is, as Husserl untiringly pointed out.[10]

In this respect, consider for example the following remarks from Husserl's *Ideas* I (1913) about how to find the right beginning in phenomenology; but do not let yourself be put off by the use of "the metaphor of vision" (see Dennett, 1991: 56), to which Husserl too, alas, frequently succumbs:

> As a matter of fact, the beginning is what is most difficult here, and the situation is unusual... It is not only that, *prior* to any method for determining matters within its field, a method is needed in order to bring, without exception (überhaupt), the field of affairs pertaining to transcendentally pure consciousness within the regard which seizes upon it; it is not only that this requires a difficult turning of the regard from the natural data...; but it is also that everything helpful to us in the case of the natural sphere of objects is lacking: familiarity by virtue of long-practiced intuition, the benefits of inherited theorizations and methods adapted to the subject-matter. (Husserl 1982: §63).

In addition consider the following passage, again from *Ideas* I, where, in the face of skepticism, Husserl addresses the issues of communication and error regarding phenomenological data:

> "Consciousness of something" is... something obviously understandable of itself and, at the same time, highly enigmatic. The labyrinthically false paths into which the first reflections lead, easily generate a skepticism which negates the whole troublesome sphere of problems... If the right attitude has been won, and made secure by practice, above all, however, if one has acquired the courage to obey the clear eidetic data with a radical lack of prejudice so as to be unencumbered by all current and learned theories, then firm results are directly produced, and the same thing occurs for everyone having the same attitude; there accrue firm possibilities of communicating to others what one has himself seen... of making known and weeding out errors by measuring them again against intuition – errors which are also possible here just as in any sphere of validity (Husserl 1982: §87).

[10] See already the Introduction to the *Logical Investigations*, §3 (originally published in 1901). For a theory to be counterintuitive is, by the way, nothing ominous but rather a virtue in Dennett's eyes; see 1991: 37f.; "radical challenges to everyday thinking" (see 45), then, are common to both Husserl and Dennett, although, of course, in different ways.

Keeping in mind such cautionary remarks concerning the phenomenological method, let me now delineate how Husserl envisaged the possibility of making consciousness of something the subject-matter of a science of consciousness. I wish to indicate why the Husserlian enterprise is by no means limited to private introspective validity, nor intent on claiming "papal infallibility" (see Dennett, 1991: 96; also 1993c: 211).

Consider first that when we are doing phenomenology we have to do "not with objects *simpliciter* in an unmodified sense, but with noemas as correlates of noeses" (see, e.g., Husserl 1982: § 133). Husserl's "turning of the regard from the natural data" in order to win "the right attitude" involves just this: to turn away from dealing straightforwardly with the intersubjectively available objects out there in the natural and cultural world, or in some abstract domain (e.g., mathematical entities), in order to study these objects in the modified sense of "intentional objects" ("noemas"), i.e. "as correlates of noeses". The point of method to be highlighted in this context is Husserl's idea of the ontological or noematic guide provided by the various kinds of objects – most of the time intersubjectively available – for uncovering the structures or forms of the noetic intentionality which, according to him, is "intentionality properly so called" (e.g., Husserl 1982: §104).

Now, the key to understanding the analysis of conscious experiences of something in terms of the "noetico-noematic correlations" (see Husserl 1982: Part Three, chapters 3 and 4) is just to see that Husserl's first concern was with the "eidetic data". And this just means that at first Husserl's concern was with analyzing philosophically the possibilities of conscious experiences of something as such and with the system of possible modifications of such experiences, rather than with these experiences and their intentional objects as actual matters of fact. By contrast, for someone interested in conscious experiences by making observations and experiments as, for example, a scientific psychologist or a neuroscientist, to be able to establish the factually given existence of the experiences would be important for grounding the empirical knowledge claims regarding one or another group of subjects; this step could not be substituted by merely imagining examples (see Husserl 1982: §7).

To be sure, in phenomenology too, a given conscious experience of something provides the experiential basis for the elaboration of its form in accordance with its very possibility. Thus, for example, a conscious experience of "imagining a purple cow" (Dennett, 1991: 27f.), or a case of "recollecting an episode" from one's own conscious experience (Dennett 1991: 406ff., 26f.), will be submitted to the analysis. However, the factually chosen case will be taken as an arbitrary example, a mere starting point for the analysis. It does not bind the phenomenologist *qua* this or that particular, factually existing subjective experience, which is such and so determined, showing for example this or that degree of vivacity and distinctness of content etc. The irrelevance of the matter of fact as such for the purpose at hand can also

be seen when we realize that we must engage in a process of varying the conditions in order to define which ones are invariably required for making the experience possible. Phenomenological analysis, then, is only interested in constituent parts or properties capable of being distinguished in reflection as belonging to the conscious experience under study in its own essence or nature, i.e. in accordance with the conditions of its possibility, and not of its factually variable actuality. In a text originally written in 1912, Husserl put the point regarding the "eidetic data" of conscious experiences as follows:

> We concern ourselves then with the eidos, the essence "perception", and with what belongs to a "perception as such", as it were to the sense, ever the same, of possible perception in general. ... We therefore differentiate the "possible" perceptions in general according to basic types; for each one we ask what belongs to it essentially and what it requires according to its essence as necessarily belonging to it, what changes, transformations, connections it makes possible purely through its essence ... Precisely the same problems result for recollections, phantasies, expectations, obscure ideas, processes of thinking of every sort, processes of feeling, of willing. (Husserl 1980a: 35)

This view has important consequences with respect to the status of the question of errors in our context. As we all know and as psychological research has confirmed, all of us all too often commit errors by misreporting what was experienced (Dennett 1991: 94). However, in trying to determine what in its very possibility makes up a conscious experience as such of a certain kind, an important liberation from the actual, but perhaps mistaken, fact of the experience's occurrence takes place. Indeed, the factual experience (my now imagining a purple cow; Dennett's then recollecting the episode of sitting in his rocker and looking out the window...) is not anything that I, when doing phenomenology, must posit and accept as the empirically existing fact as such, which I, perhaps erroneously, take it to be (or to have been). Whether the experience which I use as a basis for my analysis has actually occurred or not is a matter of fact which as such does not affect the determination of that which belongs to the possible experience in general.[11]

In the light of the preceding remarks the following important point can by now, I hope, be appreciated without unduly taking it for a claim in favor of a "constitutive" first-person authority in the sense of 'what I say goes' (Dennett 1991: 96). I have in mind a certain asymmetry of access that obtains with regard to the experiences which serve the person doing phenomenology as the basis for the reflective analysis; namely the asymmetry between *original* and *non-original* access (the latter being also called 'indirect access' or 'access by analogy'). As I see it, the asymmetry is linked to the

[11] To be sure, recall that as a working phenomenologist I may of course commit errors of analysis with respect to that which I believe to belong to an experience in its very possibility. But such errors are more like miscalculations in mathematics; they occur on another level than, e.g., to be erroneously believing in the actual existence of a certain experience which, in retrospect, will turn out to have been another one.

special way phenomena of consciousness occur. Phenomena of consciousness are, for example, to be perceiving a tree, or to be recollecting an episode from the past, or to be imagining a purple cow, or to be giving a verbal report on the basis of any of these experiences, etc., and they are first of all experientially (*erlebnismaessig*) given, that is prior to any reflection. Now, instead of actually, experientially being engaged in one or another conscious experience, thereby being interested in the object of the experience (in the thing, event, state of affairs, other person, number etc.), we human beings are able to reflect on what it is like to be consciously aware of something. In order to do so we must stop being engaged in the experience and must merely represent the experience by, for example, remembering or imagining it, or by attributing one or another of the experiences to someone else out there. Here I want to say that, among such representations of experiences, I myself have original access to those experiences of which it is possible to say that they can or could be given to me experientially, i.e. prior to the reflective work, and only I have such access to them. By contrast, among my representations of experiences, I have only non-original (indirect) access to those experiences which I attribute to others as being experientially (and thus for them, and only for them, originally) given conscious experiences.

Now when I want to determine what it is possibly like to be consciously experiencing something in one way or another, I will first of all focus on experiences which are possibly mine. For experientially given instances will implicitly contain just those distinctions of consciousness of which I am pre-reflectively aware and which phenomenology aims at making explicit as belonging to the possible experiences as such.

It is only in the cases of interpreting the behavior of others or one's own from the point of view of consciousness that factually occurring experiences matter as such; for then the phenomenologist tries to explain the empirically observable behavior by saying how conscious experiences are actually involved in this behavior. In the last section, to which I now turn, I will bring this procedure of applied phenomenology to bear in discussing the main trouble I have with Dennett's conception of heterophenomenology.

2 Towards a Phenomenological Modification of Dennett's Heterophenomenology

As we have seen, the method of Husserlian phenomenology should not be viewed as a private, even solipsistic, introspection of one's own conscious experiences, so that it would be of no avail for scientific investigations into the phenomena of consciousness. Rather, the fact that phenomenological reflective analysis of the conditions of the possibility of conscious experiences is in principle open to communication and intersubjective control leads me to think that Dennett should not so much try to *displace* traditional Husserlian phenomenology, but instead strive to *integrate* it into the project of explaining consciousness with the help of heterophenomenology. Why do I think so?

Well, the most serious trouble that I see arising from Dennett's method of heterophenomenology as it now stands (unenlightened, I dare say, by good phenomenology!) is related to this: So long as heterophenomenology is conceived as access to the phenomena of consciousness "from the outside", just as it is the case with the phenomenological level "when dealing with other phenomena" (Dennett 1993b: 153), it really (perhaps I should say paradoxically) makes a mystery of how conscious experiences, i.e. the *explananda* to be explained scientifically, could ever be detected as such. *This* mystery can be overcome, I would argue, if the objective, third-person perspective which Dennett proposes to adopt is itself submitted to methodical clarification in relation to reflective phenomenology of the Husserlian kind. For, the "heterophenomenological world... populated with all the images, events, sounds, smells, hunches, presentiments, and feelings that the subject (apparently) sincerely believes to exist in his or her (or its) stream of consciousness", even if it were presented "in the subject's own terms" and "given the best interpretation we can muster" (Dennett 1991: 98), could never be detected or identified as a world that presents itself to conscious experience in the subject's stream of consciousness, unless the interpreter, working from the outside, were actually in possession of conscious experiences that provide the experiential basis for the required knowledge. From where, indeed, if not from the interpreter's own conscious experiences, could the very idea of wanting to attribute a heterophenomenological world to a subject ever come? As I see it, we must therefore address, rather than avoid, the difficulty of how to combine the outside perspective and the reflective perspective on the phenomena of consciousness.

As Dennett (1991: 96) points out, the human subjects submitted to the heterophenomenological investigation get "the last word" when they report their beliefs about items of their heterophenomenological world. In what follows I want to draw on this privilege. Thus without much ado let us suppose that a Husserlian phenomenologist, say, I myself, be the subject submitted to heterophenomenological interpretation so that I get "the last word", rather than either Otto, – the "imaginary critic" and "spokesperson for folk psychology" whom Dennett introduces in *Consciousness Explained* (Dennett 1991: 230, 316)[12]– or even Dennett himself as he appears in the section "Getting back on my rocker" (Dennett 1991: 406ff.), where he says: "I apply my own theory to myself" (407). With this tactic I aim at bringing to the fore just those items of the heterophenomenological world which are overlooked from the outside perspective. In a certain analogy to Dennett's use of heterophenomenology, I might add that "this way of treating people", namely as being, potentially – if not actually – Husserlian phenomenologists, "is not our normal way of treating people" (Dennett 1991: 82). My purpose,

[12]Dennett may intentionally have called his critic 'Otto', so to speak *en hommage* to *auto*-phenomenology. At any rate Otto's interventions show affinities with Dennett's view of traditional phenomenology against which he presents his heterophenomenology.

however, is just to articulate the heterophenomenological world according to a Husserlian subject, and to ask how or to what extent the theorist is able to cope with it from the outside perspective.

I will in turn discuss two types of conscious experiences, namely imagining something and recollecting something, which both play a rather prominent role in contexts where Dennett applies his method of heterophenomenology.

First, let us turn to imagining something, which seems to involve "a particularly puzzling phenomenological item, the *mental image*" (Dennett 1991: 85). Suppose, then, for the sake of the argument, that I were to utter the following sentences taken from a passage in Dennett (1991: 97) which is likely to be originally attributable to someone like Otto:

> My *auto*phenomenological objects aren't fictional objects – they are perfectly *real*, though I haven't a clue what to say they are made of. When I tell you, sincerely, that I am imagining a purple cow... I am consciously and deliberately reporting the existence of something that is *really there*! It is no mere theorist's fiction to me!

When I, as a Husserlian phenomenologist, say this, what do I mean, which referents do I have in mind, and what is a Dennettian heterophenomenologist able to understand from the outside, when he tries to interpret my report, taking it, as he should, "as a record of speech acts"? (Dennett 1991: 76)[13]

As we have seen, phenomenological analysis requires that we reflect on the various ways of possibly experiencing an imagination of something like a purple cow by giving attention to the possible modes of givenness of the object (the purple cow) of the experience under study. And this cannot be done while I am actually undergoing the experience of imagining the cow, nor even when, in retrospect, I report what I have just seen and how it looked. So, when I, the phenomenologist, say that something is "really there" when I am imagining a purple cow, I at first simply refer to the pre-reflectively, experientially given act of imagining a purple cow. My "autophenomenological object", then, is a certain way of experiencing something which, if I am in my right mind, I would not confound with other possible ways of experiencing the thing in question. For this reason, I tell you that *I am imagining* a purple cow, rather than *I am remembering* a purple cow or *I am looking at a picture* of a purple cow, etc. However, these are just words, good enough in everyday life to let you know what I am doing and what you might expect from me in the given situation.

For the Husserlian phenomenologist, reflective analytical work then only begins. This work makes clear that the conscious experience of imagining something does *not* consist of looking at mental images, even though

[13] As might be expected on the basis of what I developed in the previous section, it is not relevant for the phenomenologist's purposes that I stress my sincerity in reporting what I am actually doing, namely, e.g., to be imagining a purple cow.

"people undoubtedly do believe they have mental images..." and may think of them as being "something that is really there" (Dennett 1991: 97f.; 1978: 178). If I understand Dennett correctly, I would therefore not hesitate to agree with him that "items thus portrayed" from the outside as "'intentional objects' in heterophenomenological worlds" – i.e. mental images – may "exist as real objects, events, and states in the brain", as empirical investigations would have to show, and that, "if suitable real candidates are uncovered, we can identify them as the long-sought referents of the subject's terms" (Dennett 1991: 98; see also 1991: 459).

However, I would not yet be content with such an identification. To indicate at least what is missing in a heterophenomenological world merely described in terms of "mental images" and of what may correspond to them in the brain, we must look more closely at the details of the properly phenomenological work, although I cannot expand on them within the scope of this paper.[14] To this effect it is instructive to take up Dennett's example of imagining a purple cow as it is introduced already early on in the book (Dennett 1991: 27f.). There Dennett invites the reader in the manner of cognitive psychologists like Kosslyn and others "to perform a simple experiment", following "this instruction: when you close your eyes, imagine, in as much detail as possible, a purple cow". As he previously announced, he then presents a few questions relative to how the imagined cow appeared and what the cow did. Dennett then proposes a second exercise: "close your eyes and imagine, in as much detail as possible, a *yellow* cow". This time, he asks "a different question", namely "What is the difference between imagining a purple cow and imagining a yellow cow?" (Dennett 1991:27), and he continues:

> The answer is obvious. The first imagined cow is purple and the second is yellow. There might be other differences, but that is the essential one. The trouble is that since these cows are just imagined cows, rather than real cows, or painted pictures of cows on canvas, or cow shapes on a color television screen, it is hard to see what could be purple in the first instance and yellow in the second. Nothing roughly cow-shaped in your brain (or in your eyeball) turns purple in one case and yellow in the other...
>
> There are events in your brain that are tightly associated with your particular imaginings, so it is not out of the question that in the near future a neuroscientist, examining the processes that occurred in your brain in response to my instructions, would be able to decipher them to the extent of being able to confirm or disconfirm your answers ...
>
> Suppose all this were true; suppose scientific mind-reading had come of age. Still, it seems, the mystery would remain: what is brown when you imagine a brown cow? Not the event in the brain that the sci-

[14] For a detailed account see Husserl 1980b, Kern 1975, Marbach 1993.

> entists have calibrated with your experiencing-of-brown. The type and location of the neurons involved, their connections with other parts of the brain, the frequency or amplitude of activity, the neurotransmitter chemicals released – none of those properties is the very property of the cow "in your imagination". And since you did imagine a cow..., an imagined cow came into existence at that time; something, somewhere must have had those properties at that time. (Dennett 1991:28; my emphasis)

I have quoted this passage at length, because it nicely permits me to show where Dennett's in many ways circumspect account just misses what I consider to be the very phenomena of consciousness involved in his example. Indeed, a number of remarks obtrude themselves for the phenomenologist. Most remarkable in Dennett's description of the experimental situation, I think, is the fixation on the object and its properties, in the following sense. Dennett focuses too one-sidedly on the object 'cow' and on some of the cow's properties (being purple, being yellow, being imagined, being real, being painted on canvas, etc.). Note that such an object-oriented ontological outlook is quite natural, and there is nothing wrong with it – except if the account is presented as addressing the issues which really matter in the context. And that seems just to be Dennett's view; as he puts it, "There might be other differences, but that is the essential one": that the first imagined cow is purple and the second is yellow.

Now if I, defending a Husserlian approach, were to report on the two instances of imagination, I would claim that this is obviously not the essential difference between the two situations. But note that to make the example a phenomenologically relevant one I assume that there are no purple cows to be found in the real world, whereas yellow cows could possibly be seen out there. What then really matters, I would argue, can be found out when we examine the question of how a cow, first as being purple, then as being yellow, is given to me when I imagine the cow. Similarly, what is it that I do when I imagine a cow as against when I see the cow out there in the real world, as against when I see the cow painted on canvas etc.? If we proceed in this way we hit upon what from the point of view of our conscious experiences matters decisively. In so describing the ways of givenness of the cow I must aim at consistently correlating the intentional object "imagined purple cow" to the ways of intentionally referring to the cow that are proper to the experience of imagining something. My reflection-based verbal report will thereby just focus on what, objectively and subjectively speaking, it is like to be experiencing an act of imagining a purple cow.

More specifically, I will point out to the Dennettian interpreter of my heterophenomenological world things like these: When I imagine (in as much detail as possible) a purple cow, I simultaneously actually perceive my present surroundings (acoustically, tactually, etc.) from my actually bodily occupied point of view, if only dimly and vaguely in the background of my

attention, and I represent (more or less vividly) a perceptual situation, in which the cow as it were appears in such and such a position at a certain distance and in a certain orientation to a represented, not actually occupied visual point of view of mine, whereby I am aware of the represented cow's merely imagined existence in the sense that I neither believe nor deny that I ever saw or will see the cow as really being purple, but that I rather remain neutral in this respect.

I might then continue as follows: Despite a number of similarities which I will not report again, I consciously experience a sharp contrast to the situation just described when I imagine (in as much detail as possible) a yellow cow; for then the intentional object 'imagined yellow cow' according to its very possibility is given to me in my imagination much more like an intentional object 'remembered yellow cow' would be given, but of course, the situation is consciously yet again radically unlike the one which would be given with a remembered cow. Let me just pick out the most relevant feature of the new situation in our context: When I imagine a yellow cow, I do it with an awareness of intentionally referring to an animal of which I know that it could be found in the real world, although the represented perceptual situation in which the imagined yellow cow appears as it were in relation to a represented visual point of view of mine is not taken by me to be one that actually obtains (or that did or will obtain), etc.

Let me just add that I would produce distinctly different reports with regard to the purple or the yellow cow's being given as "painted pictures of cows on canvas" etc. Again I would want to point out that the decisive differences are to be detected on the side of one's conscious experiences of seeing a cow as painted on a canvas instead of seeing a cow as appearing in a perception or in an imagination etc.[15]

Now similar remarks could also be made with respect to Dennett's (1991: 406ff.) auto-application of a heterophenomenological account to his own literary description of recollecting an episode where "the mystery of consciousness" had struck him anew (Dennett 1991: 26f.). At one point in the description Dennett says that his "musings were interrupted... by an abrupt realization. *What I was doing* – the interplay of *experiencing and thinking* I have just *described from my privileged, first-person point of view* – was much harder to 'make a model of' than the unconscious, backstage processes that were no doubt *going on in me* and somehow the causal conditions for what I was doing" (26, in part Dennett's own emphasis).

Here I can only briefly comment on this important observation which I find most pertinent. The description Dennett gives from his "privileged, first-person point of view" is again, and quite naturally so, predominantly directed to the objects and their properties, including his own bodily behavior and happenings in his brain. He only mentions several activities which

[15] In Marbach (forthcoming) I have tried to describe in detail what it is like for an object to be seen as painted in a picture.

he was engaged in during "the interplay of experiencing and thinking" at the time (e.g., reading a book, looking up from the book, being lost in thought, noticing something, listening to background music, skipping fleetingly over some dimly imagined brain processes, etc.). All this is by no means reflectively and analytically oriented in the phenomenological sense outlined above. Instead, the text is a relatively detailed description of how things appeared to him, how he behaved, and of what was going on in his mind and in his surroundings. I would say that the text is pretty much like an introspective or rather retrospective report of a past episode given by a sophisticated and sensitive person from his/her point of view – it could at least be taken to be just that.

Now, when Dennett (1991: 406ff.) applies his own theory to himself as practitioner of heterophenomenology, he says that his task is to take the text previously presented (Dennett 1991:26f.), to "interpret it, and then relate the objects of the resulting heterophenomenological world of Dennett to the events going on in Dennett's brain at the time" (Dennett 1991: 407). So far, so good, but, significantly for my purpose, Dennett seems to me to fail to attribute to himself just those sorts of conscious experiences which I, again speaking as a Husserlian phenomenologist, would detect as underlying the production of the very text (Dennett 1991:26f.) which he interprets as a heterophenomenologist, producing thereby still another text (Dennett 1991:407ff.) which, to me, is indicative of still other conscious experiences. Space here permits no more than to pick out a few examples for a very brief consideration.

Quite a lot comes to mind already regarding the fact that when the episode from the past is first introduced (Dennett 1991:26f.), it is presented as a report which appears as if composed while the author actually remembered the episode. One indication to this effect is the frequent use of the "I"-phrase together with past tense verbal forms. The description on the whole nicely evokes a past episode unfolding as if seen from somewhere there and then again by the narrator. We can assume then, that at the time of producing that text, the author was engaging in an activity of remembering while at the same time being consciously aware of his actually present surroundings (the word processor etc.).[16] Later, when Dennett interprets the text as heterophenomenologist from the outside, he rightly speaks about it in the third-person perspective, commenting on "the text", "the events about which *it* speaks", "the author" or "he", "the contents of the author's consciousness" as presented "according to the text", and the like (Dennett 1991:407ff.). Here the interpreter himself remembers how "the text was produced some weeks or months after the events... occurred", yielding a text that "portrays a mere portion (and no doubt an idealized portion) of the contents of the author's consciousness" at some time in the past. At this

[16]To be sure, the very experience of verbally reporting about the past is not just the same as the actually experienced remembering; for, as can readily be verified, remembering an episode from the past can occur with and without verbalization, whether inner or outer.

juncture, then, the interpreter's text is indicative of the fact that the objects of the heterophenomenological world of the subject, namely of the author of the text which is now plausibly recalled by the interpreter to have been an idealizing portrayal of a past episode from the conscious experiences of the author, could only in part still be considered to be the product of someone engaging in actually remembering the episode, and that it had in part to be taken as the result of "deliberate", no doubt more or less imaginative, "editorial compressions", abridgments due to memory gaps, etc. (Dennett 1991:407). Again, the interpreter, partly rememberingly, partly imaginarily, entertains also thoughts about how the text might have turned out had it been "produced... there and then", yielding thus a portrayal of something *present* at the time of the text production. To take just one more passage from "the text we have from Dennett", consider the following comment by the interpreter regarding the attention profile of the author at the time in the rocking chair:

> The attending he engaged in while rocking, and the concomitant rehearsal of those particulars that drew his attention, had the effect of fixing the contents of those particulars relatively securely "in memory" but this effect should not be viewed as storing a picture (or a sentence) or any other such salient representation. Rather, it should be thought of as just making a partially similar recurrence of the activity more likely, and that likely event is what happened, we may presume, on the occasion of the typing, driving the word-demons in his brain into the coalitions that yielded, for the first time, a string of sentences. (Dennett 1991:409)

This much must suffice here. The main point I wish to make can briefly be put like this. If I, the Husserlian phenomenologist, had "the last word", I would insist that the interpreter be very careful, when he relates the objects of the heterophenomenological world to the events going on in his subject's brain at the time of the reported episode, also to include and by the same token to differentiate among the various intentional activities corresponding to the intentional objects attributed to the subject – for there are, phenomenologically speaking, no intentional (notional) objects *without* specifically corresponding intentional experiences. As much as I agree with Dennett that having recourse to pictures, sentences and other salient representations is not required, as little am I satisfied by simply being told that "a partially similar recurrence of the activity" may explain what I would consider to be distinctly differently experienced activities of, first, attentively perceiving something out there in the world and, later on, to be remembering or, perhaps, in part only imagining some of the things and events once attentively perceived. Admittedly, all of this would require much more elaboration than I could here provide.

To conclude, what I have tried to argue could be put like this: Unless the interpreting heterophenomenologist is prepared to make use of reflect-

ively clarified knowledge about originally accessed conscious experiences, he or she will fall short of interpreting in verifiable ways the phenomenologically trained subject's verbal reports about phenomena of consciousness such as attentively perceiving one or another detail when looking out the window, imagining a purple cow, recollecting an episode from one's own conscious experiences, remembering to have recollected such an episode, imagining how a recollected episode might have been reported differently, etc. For a true verification of the referents of the phenomenologist's statements about such conscious experiences – namely a verification in the spirit of Dennett (1993a: 57), when he himself emphasizes: "If you want to avoid being taken, you will have to think it through for yourselves" - can only take place with the help of reflective clarifications of what a conscious experience possibly is, and by verbal communication concerning an occurrent instance of a conscious experience on the basis of reflection. A Husserlian phenomenologist can tell you in detail what these and other conscious experiences are like in accordance with their very possibilities. The text thereby produced deliberately refers to differences which can only reflectively be made explicit. And that means, as we have seen, that a represented instance of the conscious experience in question must in principle be available for the interpreter of the text as well. For only by reflecting on what the interpreter implicitly knows from experiencing what it is like to be perceiving, to be remembering, to be imagining, or to be remembering that one imagined, etc. can he/she verify what the Husserlian subject reports.

The moral of these considerations is: Heterophenomenology yes, but if with your method for the study of the phenomena of consciousness you want to avoid the sort of troubles discussed above, then it will not do to limit yourself to gather and assess phenomena from the outside in the style of other sciences. The simple but baffling reason is that phenomena of consciousness as such are not available out there to be observed, directly and indirectly; nor are they in here to be privately introspected in some sort of Cartesian theatre, as Dennett (1991) rightly criticizes. Rather, phenomena of consciousness are first of all experienced or lived through in the manifold ways of being active, doing something and reacting to something; and for creatures who are able to reflect phenomena of consciousness are also available for reflective articulation and systematization according to their possible forms or structures of being conscious of something.

I agree with Dennett (1991: 22) that "the phenomena of consciousness ... do not need to be protected from science"; but in order "to *save* the rich phenomenology of consciousness for scientific study" (Dennett 1993a: 50), the indirect third-person access of heterophenomenology must be combined with the original, pre-reflective and reflective perspective of (auto-)phenomenology. In just so combining the original perspective regarding conscious experiences and the perspective from the outside relative to the observable behavior, including everything that happens in the brain, the Husserlian phenomenologist in cooperation with science is able to apply what is known

about possibly being engaged in this or that conscious experience by way of attributing an instance of the kind of experience as being actually involved in the behavior given from the outside – whether rightly or wrongly in a given case is an empirical matter of fact, on which phenomenology alone cannot decide.[17]

References

Bernet, R., Kern, I., Marbach, E. 1993 *An Introduction to Husserlian Phenomenology*, Evanston, Illinois: Northwestern University Press.

Chokr, N.N. 199) "Mind, Consciousness, and Cognition: Phenomenology vs. Cognitive Science", *Husserl Studies* 9, 179–197.

Dahlbom, Bo, ed., 1993 *Dennett and his Critics. Demystifying Mind*, Oxford: Blackwell.

Dennett, D.C. 1978 *Brainstorms. Philosophical Essays on Mind and Psychology*, Montgomery, VT: Bradford Books.

Dennett, D.C. 1982 "How to Study Human Consciousness Empirically or Nothing Comes to Mind", *Synthese* 53, 159–180.

Dennett, D.C. 1987 *The Intentional Stance*, Cambridge, MA: MIT Press.

Dennett, D.C. 1991 *Consciousness Explained*, Boston: Little, Brown.

Dennett, D.C. 1993a "Caveat Emptor", *Consciousness and Cognition* 2, 48–57.

Dennett, D.C. 1993b "Living on the Edge", *Inquiry* 36, 135–59.

Dennett, D.C. 1993c "Back from the Drawing Board" in Dahlbom (ed.), 203–235.

Dreyfus, H.L. 1979 *What Computers Can't Do*, New York: Harper and Row.

Dreyfus, H.L. 1982 "Introduction", in H.L. Dreyfus, ed., *Husserl, Intentionality, and Cognitive Science*, Cambridge MA: MIT Press 1982, 1-27.

Dreyfus, H.L. 1992 *What Computers Still Can't Do*, Cambridge, MA: MIT Press.

Husserl, E. 1970 *The Crisis of European Sciences and Transcendental Phenomenology*, Evanston: Northwestern University Press.

Husserl, E. 1980a *Ideas Pertaining to a Pure Phenomenology and To a Phenomenological Philosophy*, Third Book, The Hague: Nijhoff.

Husserl, E. 1980b *Phantasie, Bildbewusstsein, Erinnerung. Zur Phänomenologie der anschaulichen Vergegenwärtigungen*, The Hague: Nijhoff.

Husserl, E. 1982 *Ideas Pertaining to a Pure Phenomenology and To a Phenomenological Philosophy*, First Book, The Hague: Nijhoff.

Husserl, E. 1987 *Aufsätze und Vorträge (1911-1921)*, Dordrecht: Nijhoff.

Kern, I. 1975 *Idee und Methode der Philosophie. Leitgedanken für eine Theorie der Vernunft*, Berlin: Walter de Gruyter.

[17] It is a pleasure to thank Daniel C. Dennett, not only for his stimulating and perceptive remarks on this paper, but also for having had the patience to engage, over a period of several years, in a dialogue across philosophical traditions. He has raised some difficult questions, some of which I cannot deal with now, concerning the relationship between "reflectively clarified knowledge" – as claimed by phenomenologists – and the naive introspective protocols of non-phenomenologists. In particular, Dennett asks whether or not these reflective phenomenological findings really are material of a *different* kind from what the Dennettian heterophenomenologist is dealing with. If so, then – as my text suggests – an enlargement or revision of the heterophenomenological method would have to be envisaged.

I also want to thank the Swiss National Science Foundation for support of this work (grant Nr. 11-30049.90).

Marbach, E. 1988 "How to Study Consciousness Phenomenologically, or Quite a Lot Comes to Mind", *Journal of the British Society for Phenomenology* 19, 252–268.

Marbach, E. 1993 *Mental Representation and Consciousness. Towards a Phenomenological Theory of Representation and Reference*, Dordrecht: Kluwer.

Marbach, E. (forthcoming) "On Depicting", in G.Haefliger/P.Simons (eds.), *Analytic Phenomenology: Essays in Honour of Guido Küng*, Dordrecht: Kluwer.

McIntyre, R. 1986 "Husserl and the Representational Theory of Mind", *Topoi* 5, 101–113.

Münch, D. 1990 "The Early Work of Edmund Husserl and Artificial Intelligence", *The Journal of the British Society for Phenomenology* 21, 107–120.

Why Parallel Processing?

Jaakko Hintikka

Strong claims have recently been made for parallel distributed processing (PDP) as offering a better model for human information handling than sequential processing.[1] PDP is claimed to be "what makes people smarter than machines". In order for such claims to be correct, there must be something that parallel distributed processing does better than sequential processing. It is far from clear, however, what this alleged superiority of PDP is based on. Sometimes it is apparently thought that the advantages of parallel processing are due to an enhancement of the speed of the requisite computations. Sometimes it is thought that parallel processing systems with a large number of units are the right models of actual neutral activities. The superiority of such large systems would then presumably be based on their statistical properties.[2]

These claims are hard to evaluate. For instance, I am not aware of any explicit definition of computability on the basis of which one could show that parallel processing systems can do more than linear systems. Hence it might seem that the advantages of parallel processing as a model of human information handling over old-fashioned Turing machine representation would be manifested only in the statistical behavior of large-scale PDP systems.

I shall not discuss these problems in their whole generality. Instead, I shall offer a simple but subtle logico-combinatorial reason why distributed parallel processing is in a certain sense more powerful modelling tool than sequential processing. This reason is independent of the statistical properties of PDP systems, and is in fact manifested already in the smallest parallel processing systems.

My line of thought utilizes a partial parallelism between logical formulas and distributed processing systems, no matter whether sequential or parallel. Such parallelisms can undoubtedly be set up in many different ways. I shall rely on only one such parallelism.

Consider, for the purpose, a distributed processing network. It consists of a number of processing elements called units. These units can interact with each other.

Consider now one such unit. What does it do, and how can we describe its activity? Here we are not interested in its internal operations. For us, it is

[1] Rumelhart *et al.* 1986

[2] In Rumelhart *et al.* 1986: ix, the idea of parallel distributed processing is characterised as "the notion that intelligence emerges from the interaction of large numbers of simple processing units".

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 265-272

the proverbial black box. All that it does is to take an input x and transform it into an output y. What x and y are does not matter. The values of these variables are simply all the relevant kinds of inputs and outputs. We can think of them as numerical variables, but my line of thought is not predicated on that assumption.

I shall assume, however, that processing units are all homogenous, i.e., that the output of a unit can always be the input of another unit.

I shall assume that certain conditions can be imposed on inputs and outputs, i.e., certain things can be said of them. If the inputs and outputs are numerical, then the conditions on them will be suitable numerical relations.

I shall assume that these relationships can be expressed by means of certain basic relations and their Boolean combinations. In other words, I shall assume that we are given the nonquantified part of a first-order language to describe the inputs and outputs.

What a unit U does can now be described as follows:

When given an input x (or several inputs, $x_1, x_2, \ldots, x_k$) and produces from it (or them) an output y (or several outputs $y_1, y_2, \ldots y_l$), related to the inputs in the way expressed by the condition

$$C[x, y] \tag{1}$$

or, more generally,

$$C[x_1, x_2, \ldots x_k, y_1, y_2, \ldots y_l]. \tag{2}$$

Then what the operator U does can be "expressed" by the first-order statement

$$(\forall x)(\exists y)C[x, y] \tag{3}$$

or, more generally

$$(\forall x_1)(\forall x_2)\ldots(\forall x_k)(\exists y_1)(\exists y_2)\ldots(\exists y_l)C[x_1, x_2, \ldots x_k, y_1, y_2, \ldots y_l]. \tag{4}$$

For instance, an addition device would correspond to the formula

$$(\forall x)(\forall y)(\exists z)(x + y = z). \tag{5}$$

This correlation between computational systems and first-order statements is very much in the spirit of the explanation by J. L. McClelland *et al.*, that "in some cases, the units [of a PDP model] can stand for possible hypotheses...".[3]

I am not raising the question here as to which of the prima facie descriptions of processing units of the form (5) correspond to actually realizable computational devices or even abstract devices like Turing machines.

It is well known that there are first-order statements which cannot be satisfied by any computable Skolem functions. In other words, these are descriptions of the form (5) of processing units that cannot be realized by

[3]Rumelhart *et al.* 1986:10.

any digital computing device (Turing machine). This makes the correlation between certain first-order formulas and actual computing devices only a partial one. In other words, only some of the first-order statements we have considered are true. Such descriptions nevertheless make sense even when there is, say, no Turing machine realizing the requirement which the formula expresses. We need such descriptions among other things for the very purpose of discussing which of them are realizable by different kinds of actual computing devices.

Furthermore, we might even want to keep open the possibility that a processing unit is an analogical device which is able to realize a condition not realizable by any Turing machine.[4]

In any case, the logical relations between first-order sentences will be mirrored by certain relations of processing power between the corresponding units. For instance, if

$$(\forall x)(\exists y)C_1[x, y] \tag{6}$$

logically implies

$$(\forall x)(\exists y)C_2[x, y] \tag{7}$$

(with no quantifiers in either $C_1[x, y]$ or $C_2[x, y]$), then the processing power of the unit correlated with (6) is in an obvious sense greater than that correlated with (7). Thus the processing power is in a crude but obvious sense measured by the logical strength of the correlated first-order statement.

Instead of first-order sentences like (3)-(7), we could also use the corresponding Skolem function representations, such as

$$(\exists f)(\forall x)C[x, f(x)] \tag{8}$$

$$\begin{gathered}(\exists f_1)(\exists f_2)\ldots(\exists f_l)(\forall x_1)(\forall x_2)\ldots(\forall x_k) \\ C[x_1, x_2, \ldots x_k, f_1(x_1, x_2, \ldots x_k), f_2(x_1, x_2, \ldots x_k]\end{gathered} \tag{9}$$

$$\begin{gathered}(\exists f_{11})(\exists f_{12})\ldots(\exists f_{21})(\exists f_{22})\ldots(\forall x_{11})(\forall x_{12})\ldots(\forall x_{21})(\forall x_{22})\ldots \\ C[x_{11}, x_{12}, \ldots x_{21}, x_{22}, \ldots f_{11}(x_{11}, x_{12}, \ldots), f_{12}(x_{11}, x_{12}, \ldots), \ldots \\ f_{21}(x_{11}, x_{12}, \ldots x_{21}, x_{22}, \ldots), f_{22}(x_{11}, x12, \ldots x_{21}, x_{22}, \ldots), \ldots].\end{gathered} \tag{10}$$

These are equivalent to (3)–(5), respectively.[5]

That we are dealing with abstract specifications of what a unit might be able to do and not with actual recipes for computation is reflected by the fact that in (6)–(8) we have (existentially quantified) function variables rather than specific functions.

The operation of a complex processing system which involves several units can be likewise correlated with a logical formula. However, when such

[4]There ar deterministic systems of classical mechanics whose mode of operation can produce a nonrecursive sequence of numbers. If such a system is thought of as an analog computing device, it can do what no Turing machine can do. See, for example, Ekeland 1988: 59–64 for references to the original papers.

[5]This equivalence depends on the axiom of choice. However, for a variety of reasons, I see no reason to worry about this dependence.

a system is analyzed as being sequential or parallel (or otherwise involving some internal structure), we have to see what this entails concerning the form of the corresponding logical representation of the system.

For instance, consider the system which consists of two sequentially connected units, as follows:

$$x \overset{U_1}{\longmapsto} \begin{array}{c} y = f_1(x) \\ (\forall x)(\exists y)C_1[x,y] \\ (\exists f_1)(\forall x)C_1[x,f_1(x)] \end{array} \overset{U_2}{\longmapsto} \begin{array}{c} z = f_2(y) \\ (\forall y)(\exists z)C_2[y,z] \\ (\exists f_2)(\forall y)C_2[y,f_2(y)] \end{array}$$

The sequential combination corresponds to the statement

$$(\exists f_1)(\exists f_2)(\forall x)(C_1[x,f_1(x)]\&C_2[f_1(x),f_2(f_1(x))]) \tag{11}$$

This second-order statement has a first-order equivalent, viz.

$$(\forall x)(\exists y)(\exists z)(C_1(x,y)\&C_2[y,z]) \tag{12}$$

But consider now a system consisting of two parallel units U_1, U_2. Its output will be an ordered pair of numbers (or other objects of computation) y,u obtained from two separate inputs x,z. The four objects x,y,z,u will satisfy a certain condition $C_o[x,y,z,u]$.

The situation can be represented as follows:

$$\begin{array}{ccc} (\forall x)(\exists y)C_1[x,y] & = & (\forall x)C_1[x,f_1(x)] \\ & xU_1y & \\ & C_o[x,y,z,u] & \\ & zU_2u & \\ (\forall z)(\exists u)C_2[z,u] & = & (\forall z)C_2[z,f_2(z)] \end{array}$$

Then the *modus operandi* of the combined system U_o can obviously be described by the following second order formula:

$$(\exists f_1)(\exists f_2)(\forall x)(\forall z)C_o[x,f_1(x),z,f_2(z)] \tag{13}$$

But it is known that statements of the form (13) do not in general have a first-order equivalents, not even if we restrict $C_o[x,y,z,u]$ to arithmetical relations. What (13) says can be expressed by the following branching-quantifier expression:[6]

$$\begin{array}{cc} (\forall x) & (\exists y) \\ (\forall z) & (\exists u) \end{array} C_o[x,y,z,u] \tag{14}$$

It can also be expressed in a notation where, game-theoretically speaking, informational independence of a quantifier (Q_1) from another one, say (Q_2),

[6]For partially ordered quantifier structures, including branching quantifiers, see, for example, Henkin 1961 or Barwise 1979 (which has further references). See also notes 7, 9, 11 and 12 below.

can be explicitly expressed. I have proposed using a slash notation (Q_1/Q_2) for the purpose.[7] In this notation, (13) is equivalent to

$$(\forall x)(\forall z)(\exists y/\forall z)(\exists u/\forall x)C_o[x,y,z,u] \tag{15}$$

The connection between either of these notations and the idea of parallel processing is obvious.

Such examples can easily be generalized. What they show are the advantages of parallel processing for the purposes of human reasoning, including logical reasoning. By spelling out the parallel processing character of a system we can express stronger requirements that can be imposed on such systems. This means being able to represent through actual or possible processing systems logical statements which go beyond first-order logic. In other words, we can thus reach means of representing by means of possible parallel processing systems modes of thinking and reasoning which are beyond the purview of first-order logic. This is the reason for preferring parallel processing as a modelling device to sequential processing which I promised in the beginning of this paper.

For instance, in any computational level theory in the sense of David Marr we have to discuss what the computational tasks are that a machine – or an organ – has to perform.[8] In Marr's own words, we have to study "the logic of the strategy by which it [the task of computation] can be carried out". The correlation of certain statements in a formal language and certain computational tasks that I have expounded can serve as a tool for studying in general terms the logic of such strategies. By means of this correlation, we can express the kinds of requirements or computational tasks which a computational level theory deals with. The main thesis propounded here concerns the kind of logic that is adequate for this task. What I am proposing is that as soon as parallel processing is involved, any adequate logic must allow for quantifier independence.

It is important to be clear about what is involved here. The advantages which the theory of parallel processing offers as a modelling device over, say, the theory of Turing machines do not lie in the fact that parallel processing devices offer possibilities of computation which Turning machines do. No one has, to the best of my knowledge, given a clear reason to think that they offer such possibilities. They don't. What is involved is that the analysis of computational information processing through Turing machines does not enable a theorist to analyze the computational processes in the right way, i.e., in a way which would facilitate the modelling of non-elementary logical relations by means of computational processing systems.

The combinatorial character of the reasons for preferring parallel processing as a modelling device reason can be seen especially clearly from the

[7]See Hintikka and Sandu 1989.

[8]See Marr 1982, especially p. 25

general conditions on which a second-order statement of the form:

$$(\exists f_1)(\exists f_2)\dots(\forall x_1)(\forall x_2)\dots C_o[x_1, x_2, \dots f_1(x_{11}, x_{12}, \dots), \dots] \qquad (16)$$

where each set of arguments $\{x_{i1}, x_{i2}, \dots\}$ is a subset of the set $\{x_1, x_2, \dots\}$, can be reduced to a usual first-order form.[9]

It can in fact be shown that (16) reduces to a linear first-order form if and only if the functions $f_1, f2, \dots$ can be ordered in a suitable way. Assuming that $f_1, f_2, \dots$ itself is the order, this critical condition is simply

$$\{x_{i1}, x_{i2}, \dots\} \subseteq \{x_{i+1,1}, x_{i+1,2}\} \qquad (17)$$

In other words, the arguments of earlier functions are always included among the arguments of later functions.

This way of spelling out the combinatorial situation is especially suggestive in that a violation of the ordering requirement in effect means that the computations codified in the functions fi are done partly in parallel. Hence the step from ordinary first-order logic to a logic which allows for quantifier independence is connected very closely with the very idea of parallel processing.

These observations prompt a few interesting further conclusions. First, first-order logic is not adequate as a tool in the theory of parallel processing. We need a logic that is at least as strong as the logic of partially ordered quantifier structures. And, since parallel processing is highly important in AI, the logical tools needed in the study of AI must be much stronger than first-order logic.

This observation is interesting in view of the widespread use of first-order logic in AI, database theory, etc.[10]

A second conclusion is prompted by the question: What is the natural logic that corresponds to parallel processing if first-order logic fails, as we saw? We could resort to a fragment of second-order logic or to the logic of partially ordered quantifiers. Neither is very intuitive nor easily delineated. There is a more systematic way of building the requisite logic starting from ordinary first-order logic. This way consists simply in a systematic use of the independence (slash) notation explained above, while sticking to first-order logic in all other respects. The result might be called the independence-friendly (IF) first-order logic.[11]

It is to be noted that in an IF logic the slash notation can be applied to logical constants other than quantifiers, especially to propositional connectives. In order to keep the back references to different occurrences of the same connective separated from each other we may also have to use also some simple coindexing method.

[9]This follows from the results of Walker 1970.

[10]The contrast I have in mind here is that between ordinary first-order logic and higher-order logic (or such extensions in the direction of higher-roder logic as IF logics). The usual modal logics, epistemic logic, etc., are in the sense intended here in the same boat as ordinary first-order logic.

With this proviso in mind, one can see that the majority of papers in such AI oriented volumes as Halpern 1986, or its successor Vardi 1988, use an essentially first-order logic.

[11]See Hintikka forthcoming.

Thus the natural logic of information processing is the IF first-order logic, not the fragment of this logic which is the usual (independence-free) first-order logic. This naturalness is illustrated by the case at which the IF first-order logic captures the relevant relationships. It is for instance easy to see how (16) can be expressed in the IF notation. The representation can have the form

$$(\forall x_1)(\forall x_2)\dots(\exists y_1/\forall x'_{11}, \forall x'_{12}, \dots)(\exists y_2/\forall x'_{21}, \forall x'_{22}, \dots)\dots \qquad (18)$$
$$C_o[x_1, x_2, \dots y_1, y_2, \dots]$$

where$\{x'_{i1}, x'_{i2}, \dots\}$ is the complement of $\{x_{i1}, x_{i2}, \dots\}$ with respect to $\{x_1, x_2, \dots\}$.

Even though notationally an IF first-order logic is but a mild extension of the ordinary first-order logic, it adds a great deal to the logical power of our logic. Indeed, it has been shown that the decision problems of the theory of partially ordered quantifiers, which is a part of the IF first-order logic, is as difficult as the decision problems for the entire second-order logic (with standard interpretation).[12] Hence the increase in logical power in the transition to an IF logic is truly formidable. As a consequence, the IF first-order logic cannot admit of a complete axiomatization.

In view of the naturalness of an IF first-order logic as the logic of computational information processing, how come its role has not been recognized earlier? A partial explanation lies in the fact that in natural language independence is not marked at all syntactically. (The reasonsy for this fact are likely to be quite interesting from a linguistic viewpoint.)[13] Hence, when information processing, including parallel processing, is discussed in normal English prose, the role of informational independence easily escapes one's notice.

Some of the proponents of PDP as a tool of modelling human thought or as tool in AI have belittled the applications (and applicability) of logic for these purposes. If by logic they simply mean customary first-order logic, I can see the point of their criticisms. However, it seems that the underlying reason for the superiority of parallel processing can only be spelled out clearly by reference to logic.

References

Jon Barwise 1979 "On Branching Quantifiers in English", *Journal of Philosophical Logic* 8, 47–80.

Ivar Ekeland 1988 *Mathematics and the Unexpected*, Chicago: University of Chicago Press.

Joseph Y. Halpern (ed.) 1986 *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1986 Conference*, Los Altos, CA: Morgan Kaufmann.

[12]A proof is sketched in Hintikka 1974.

[13]See Hintikka 1990.

Leon Henkin 1961 "Some Remarks on Infinitely Long Formulas", in *Infinitistic Methods*, Oxford: Pergammon, pp. 167–83.

Jaakko Hintikka 1974 "Quantifiers vs. Quantification Theory", *Linguistic Inquiry* 5, 153–77.

Jaakko Hintikka 1990 "Paradigms for Language Theory", *Acta Philosophica Fennica*.

Jaakko Hintikka (forthcoming) "Independence-friendly Logic as Elementary Logic".

Jaakko Hintikka and Gabriel Sandu 1989 "Informational Independence as a Semantical Phenomenon", in Jens Erik Fenstad *et al.* (eds.), Logic Methodology and Philosophy of Science VIII, Elsevier Amsterdam: Science Publishers, pp. 571–89.

David Marr 1882 *Vision*, San Francisco: W. H. Freeman.

Moshe Y. Vardi (ed.) 1988 *Theoretical Aspects of Reasoning About Knowledge: Proceedings of the 1988 Conference*, Los Altos, CA: Morgan Kaufmann.

W. Walker, Jr. 1970 "Finite Partially Ordered Quantification", *Journal of Symbolic Logic* 35 (1970), 535–55.

Automated Deduction and Artificial Intelligence

Neil Tennant

As a central normative component of cognitive science, computational logic should direct its interests first and foremost to the design of ideal inference engines for artificial intelligence. The computational logician wants the computer to solve deductive problems. A deductive problem is of the form X ?– A, where X is a finite set of sentences (the premisses) and A is a sentence (the conclusion). If the computational logician succeeds, the resulting programs can be used as components in the design of an artificial intelligence. For in designing an artificial intelligence we want a *cybernetic theorizer*: we want cybernetic development of the logical consequences of sets of mathematical axioms; cybernetic unfolding of the predictive content of our scientific hypotheses; and cybernetic revision of those theories in the light of contradictions discovered, whether internally or with recalcitrant observation. In short, wherever logical reasoning features in our own framing, testing and revision of theories, we should aim to have computational surrogates for that exercise of our own logical intelligence. I say this in full awareness of the undecidability of many logics, and of theories based on them. What we need is *eine Einstellung zur Seele als Maschine*, an attitude that concedes the possibility of cybernetic emulation of our logical abilities.

Computational logic can take two forms. First, one could devise logic programs that would prompt the human user interactively when the human user is seeking a proof of a given argument. In this rôle the computer with its program is no more than a desk-mate, relieving the human user of some of the formal drudgery involved in precise rule applications, and reminding the human user of the full gamut of permissible moves that might be made at any stage with the deductive materials at hand. It would still be up to the human user to make the strategic decisions as to where to look next for an unbroken path through the logical foothills. This is called *interactive* theorem-proving, and I am not interested in it.

What interests me rather is *automated* theorem proving. On this approach, the computer is programmed to carry out the whole proof search entirely on its own, following the search algorithm already embodied in the program. Such programs can be logically complete only if the logical system (or the theory within which one is working) is *decidable*, and not merely itself complete. With an undecidable system, every such (terminating) program for problem-focused proof search will be incomplete. But so too would be

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 273-286

that of any human mathematician for the system concerned; for the mathematician is just another machine. So even with respect to an undecidable system one could emulate the abilities of the human logician or mathematician. Even if we cannot match each and every human reasoner within one all-encompassing program, then at least we can furnish for each and every human reasoner a matching program. The aim, then, is to make an automatic theorem-proving program that could take its place within the human mathematical community, and earn its colours by passing the sort of Turing Test that mathematicians constantly apply to each other in sustaining their sense of community.

As we know from applications of computers in other areas, their great advantage is their astonishing speed and their formidable accuracy. They can take the drudge out of the logician's task of formal pattern-matching when invoking axiom schemata and applying rules of inference. They can also make sure that no logical slips are made – that everything done so far has been done correctly, according to the formal rules of the system concerned. Computers keep our logical toils honest. They force us to abide by the rules we have laid down. These rules specify what inferential moves we may make; and what moves the doubting audience must accept as licit. There is astonishing hardness in the silicon must.

Computational logic, then, even conceived as a part of cognitive science, is a normative enterprise. One is seeking a model of competence, not of performance. The methodology is *a priori*, and involves expert introspection. The challenge is to identify, formulate and implement constraints governing search for proofs within given theories.

In a project like this, one has to get three things right:

- a theoretically well-informed choice of logic (where the theoretical considerations can concern such matters as computational complexity, philosophical grounding, adequacy for science and mathematics, etc.)
- a decision algorithm that is no more complex than it has to be for the logic in question (where the algorithm can be one that searches for proofs and/or counterexamples, and where the proofs can be in various formats)
- an implementation of that algorithm that is as efficient as possible (where the implementation should be suitably general and uniform across all possible input problems)

If you want a novice to become a good architect, let her start with Lego. If you want to build a super-sophisticated deductive system, start with simple p's and q's. I shall now describe some theory developed to devise automated proof search for certain decidable propositional logics: logics deserving to be called systems of constructive and relevant reasoning. (A *propositional* logic deals only with logical operators that form sentences from one or more sentences. The most familiar of these are $\sim$ (negation), & (conjunction),

$\vee$ (disjunction) and $\supset$ (implication). A logic is *decidable* when there is an effective method for deciding, of any finite set of premisses and a conclusion, whether the latter can be deduced from the former within that logic.)

Deep and interesting problems emerge for the automation of proof search even with these simple materials: problems the shape of whose general theoretical solution one must discover and appreciate if one is to hope for any progress later with more complicated systems. The computational explosion involved in proof search, even in decidable propositional logics, is awesome. One has to learn how to tame it before tackling the quantifiers, which take one into the realm of the undecidable.

The computational logician is interested not just in the question of whether a logic is decidable, but also in the question whether its decision problem is *tractable*. Every tractable problem is decidable, but not conversely. Tractable problems are conventionally taken to be those that can be solved in polynomial time. Thus, the decision problem for a logic if there is a polynomial function $f(\cdot)$ such that for any problem $X \vdash A$ of length n, it could be decided within $f(n)$ steps (or units of computing time) whether there was a proof of the conclusion A from the set X of premisses within that logic. Alas, none of the usual propositional logics turns out to be tractable in this sense.

Classical logic C properly contains *intuitionistic* logic I, which properly contains *minimal* logic M. M consists of just the introduction and elimination rules for the logical operators. I has in addition the absurdity rule, allowing us to infer anything from a contradiction. C goes even further by adding one of the classical rules of negation: the law of double negation, or classical reductio, or the rule of dilemma, or the law of excluded middle. Within I, M has an overlapping neighbour: the system IR of intuitionistic relevant logic. From M there is an easy theoretical transition to IR. This system, I have argued elsewhere, has strong meaning-theoretic, epistemological and methodological claims to adequacy and correctness. I maintain it is the right logic. Like M, it lacks the absurdity rule; but unlike M, it contains disjunctive syllogism and *lacks* the inference $A, \sim A \vdash \sim B$.

Because of its extra negation rules, C has the tamest decision problem of all: although it is not tractable in the sense just explained, it is nevertheless *NP-complete*. What does it mean when one says that a computational problem class (such as deciding whether a given deductive problem admits of proof) is NP-complete? It means that it can be solved in non-deterministic polynomial time (that is, it is NP-easy); and that any other problem class that can be so solved can be transformed into the one at hand in (deterministic) polynomial time (that is, it is NP-hard). This in turn raises definitional demands regarding non-deterministic polynomial time and (deterministic) polynomial time. A problem class can be solved in *non-deterministic polynomial time* just in case there is an algorithm α for solving it, and a polynomial function $f(\cdot)$, such that for any input problem π of length n, the search tree generated by the branching (choice) points of

the algorithm α applied to the problem π contains a solution at the end of a branch of length less than $f(n)$. An intuitive way of thinking of this is that if, in the execution of the algorithm α (which requires us at branching points to make single choices from a range of possible alternatives) we were so it lucky as always to choose "correctly" – and thereby solve the problem as quickly as possible – then indeed we would do so in no more than $f(n)$ units of time.

P is the class of problems (more exactly: problem classes) that can be computed in polynomial time. NP is the class of problem classes that can be computed in non-deterministic polynomial time. It is an open problem whether P = NP. The orthodox conjecture is that P is properly contained in NP; for no-one has ever provided, for any one of the hundreds of problem classes now known to be NP-complete, an algorithm that runs in polynomial time.

C is NP-complete because it provides a very replete set of proof methods. Both I and M, by contrast, are PSPACE-complete. This means that the amount of memory (rather than the number of units of time) needed for the relevant computations is bounded above by a polynomial function. Computations in general need more time than memory, for it takes time to exploit memory storage. So PSPACE-completeness is likely to be worse than P(time)-completeness; indeed, likely to be worse even than NP-completeness.

Let us look more closely at the problem of proof search for systems of constructive and relevant reasoning. For systems such as M, I and IR, there are no such results as conjunctive normal form or disjunctive normal form theorems. The de Morgan laws and dualities break down. Double negations cannot in general be eliminated. One cannot simplify a problem by normal-forming of formulae, or pre-processing its premisses and conclusion before embarking in earnest on proof-search with the simplified forms. Thus the so-called resolution method in automated theorem proving, insofar as it relies on such normal-forming of formulae, is confined to classical logic.

The computational credentials of IR are as compelling as its philosophical and metamathematical ones. By this I mean that its exploitation of the notion of relevance will not make it strictly more difficult to find proofs than it already is in intuitionistic logic. My method of relevantising a system of logic will produce a relevant system whose decision problem is no more complex than the decision problem of the parent system. The method of relevantising the system I of intuitionistic logic to get the system IR also produces, in a exactly similar fashion, the system CR of classical relevant logic from the system C of classical logic. Both IR and CR are decidable, and have decision problems no more complex than those of I and of C respectively.

By contrast, the propositional relevance logic R of Anderson and Belnap is undecidable. Its well-known decidable fragment LR, obtained by dropping the distributivity axiom, has an awesomely complex decision problem: at best ESPACE hard, at worst space-hard in a function that is primitive recursive in the generalised Ackerman exponential. From NP to ESPACE or worse, courtesy of "relevance"! – so much the worse, then, for this brand of relevance.

These complexity results for LR (and the undecidability of R itself) bring out an important tension between one motivation for studying relevance, and the Anderson-Belnap paradigm for treating it. The *motivation* is that a proof system obeying some constraint of relevance of assumptions to conclusion would admit of faster proof search. (The untutored idea is that a suitable relevance constraint will somehow narrow the search space.) The Anderson-Belnap *paradigm* for treating relevance is their family of systems, clustering around R, that enjoy unrestricted transitivity of proof, eschew disjunctive syllogism, and involve so-called "intensional" connectives. If the motivation and the untutored idea are sound, we must conclude, in the light of the complexity results, that the Anderson-Belnap approach to relevance is not. The possibility remains that some other characterization of relevance will yield the sought reduction in complexity of proof search (or, at least, no increase). Here is an opportunity for computational concerns to inform philosophical preferences for one system of relevance logic over another.

No-one interested in the logical reconstruction of mathematics should undertake it in the relevance logic of Anderson and Belnap. For it forces one to give up disjunctive syllogism – because it holds on to unrestricted transitivity of deduction. Those who would "relevantise" mathematics in R have to re-derive every well-known mathematical theorem from the accepted axioms. They have no general metatheorem to the effect that if a result holds classically, then there is a relevant proof of it.

By contrast with R, we have metatheorems, for the systems IR and CR of relevance logic, guaranteeing the relevantisability of consistent mathematical theories:

Theorem 1 *Any proof in I [C] of a conclusion A from a set X of premisses can be transformed into a proof in IR [CR] of either A or* Λ *(absurdity) from (some subset of) X.*

This guarantees epistemic gain. On relevantising any classical or intuitionistic proof, one obtains either a proof of the sought conclusion from the set X of original premisses, or a proof of that conclusion from some proper subset of X, or a proof that (some subset of) X is inconsistent. We may sum up by saying:

- we may relevantise without loss on consistent sets of premisses
- we may relevantise without loss on logical truths (i.e., on the empty set of premisses)
- we may relevantise without loss on inconsistent sets of premisses

So on the assumption that mathematics is consistent, we are assured that every mathematical theorem (classical or intuitionistic) admits of the corresponding kind of relevant proof. And if mathematics is not consistent, we are assured further that we shall be able to prove this relevantly. We are assured, moreover, that every logical truth can be proved in relevant logic. Finally, we are assured that the hypothetico-deductive method of science, which involves the logical pursuit of inconsistencies between hypotheses and observational evidence, can be carried out using relevant logic.

The systems IR and CR can do everything that any intuitionist or classicist, respectively, could wish their logic to do. Furthermore, we have now the prospect of exploiting the relevance relation (as these systems explicate it) so as to speed up proof-search – or, at the very least, not slow it down. Propositional CR should be no harder than NP; propositional IR no harder than PSPACE. And this is indeed the case.

To summarise, then, we have the following contrasts between the Anderson-Belnap approach to relevance and the approach that I favour:

- They can relevantise mathematics only piecemeal. By contrast, there are metatheorems guaranteeing the relevantisability of mathematics in IR and in CR.
- They keep unrestricted transitivity of deduction, and abandon disjunctive syllogism. By contrast, in IR and CR the transitivity of deduction is controlled in an epistemically gainful way, and disjunctive syllogism is retained.
- Their propositional logics are either undecidable or have decision problems of awesome complexity, compared to those of their parent systems. By contrast, my propositional logics lead to no increase in the complexity of the decision problem.

A central methodological question in connection with any logic we may try to treat computationally is: Do we search for proofs by brute methods or refined ones? By machine-driven merry dance, as with resolution or model elimination methods – or by somehow emulating or simulating competent human interests and methods?

Considered as part of cognitive science, computational logic faces a double challenge. First, one still has the old challenge of using computers as prosthetic devices. That is, one programs computers to find solutions to difficult problems faster and much more accurately than the unaided human mind. One can meet this challenge without any concern for naturalness, simplicity, elegance or "human-like" features of one's search algorithms, except insofar as these features might conduce anyway to greater speed and efficiency in the execution of the programs on the available hardware.

As part of cognitive science, computational logic faces a new challenge: that of programming the machines to *emulate* human reasoning by *simulating* it. That is, one tries to design algorithms for proof search, to be executed by the machines, that are as isomorphic as possible to whatever collection of methods is employed by competent human reasoners in search of suasive arguments. To face this new challenge, as a cognitive scientist, is to give hostage to fortune in the competition to design proof-finders that do their work simply as fast as possible. For the available hardware, because of its radically different physical construction from the human brain, might be much better at some tasks than the human being; and, correlatively, might be much worse at others.

The human brain, as a product of natural selection, has highly evolved abilities to recall and match visual patterns. This ability no doubt features crucially in higher order human competence in schematic reasoning – at least on reasonably short problems. Likewise, we are able to remember past attempted moves and their outcomes when solving a problem. (And the word 'remember', remember, is ambiguous between record and recall.) Our relatively effortless memory of what we have recently done enables us to learn rapidly from past mistakes.

In both these respects we are, arguably, somewhat different from today's programmable hardware. Depending on one's programming language, there may be a disproportionately higher cost, in the case of a machine, associated with recording and consulting results of recently past computations. And pattern-matching and other forms of associative learning may be severely hampered by the physical design of the hardware. We may have to wait for an engineering revolution in the design of neural networks before our prosthetic devices' profile of relative competences begins to match our own. Just having the theoretical assurance that any Turing machine can be modelled by one of today's digital computers offers no comfort to the cognitive scientist endeavouring to use those computers to simulate our own range of competences in a way that would be real-time faithful.

With the human profile of competence possibly drastically skewed with respect to that of the digital computer, it may turn out that unnatural algorithms can be executed faster than the natural ones that reflect specifically human techniques, abilities, interests and methods. This must constantly be borne in mind when assessing the models of reasoning, or of proof search, offered by computational logic as a branch of cognitive science. Only then can we properly compare the achievements of researchers who employ any brute-force or machine-friendly method, with those of researchers who want to "make the machine think the way we do". If the latter can come close to matching the achievements of the former as far as execution times are concerned, that would already be cause for considerable satisfaction. (Especially when one considers how late has been the entry of "natural deduction"-minded logicians into the field of computational logic.) But there is the exciting prospect also of achieving the added benefit of, say, finding the very proofs that human beings would find, by following methods that human beings themselves deploy. So I would venture to suggest that, in addition to execution times, one consider the nature of the process and features of the

output – in particular, the length and structure of a natural deduction – before entering a decision as to which computational logic program is optimal as a model of human deductive reasoning.

Deductive logic is the centrepiece of any model of (ideal) cognition and reasoning. One's metaphysical stance can influence choice of methodology: e.g. choice of syntactic, proof-theoretic methods over semantic, model-theoretic methods which in their full extent can deal with infinitary objects. A cognitive scientist whose metaphysical position is basically materialist, and who is impressed by the finitude of the neural network, will incline towards models of cognition and reasoning that involve effective transformation of finitary representations.

Now from the standpoint of one interested in human cognitive competence – and in particular the ability to reason logically – "natural proof search" methods seem strikingly underdeveloped. Computational logic needs to explore the algorithmic gains in efficiency on offer from over five decades of proof theory. The kind of proof theory I have in mind here is what might be called intra-systematic proof theory. Its main concern is to achieve a thorough understanding of what a given proof system is like "from the inside". It studies the structure, in the system, of proofs in normal form. The system is characterized by its rules of inference and by the way steps according to them can be patterned so as to form proofs.

It has been the central concern of the work reported on here[1] to explore what proof theory can offer computational logic. Successful proof-search can be very fast, when it is guided by constraints deriving from a deeper understanding of the structure of proofs in normal form. There are also some unexpected benefits for proof theory in confronting the exigencies of computation. The main one is a hybrid system of proof that can be characterised as midway between a Gentzen sequent system and a Prawitz-style natural deduction system.

Another closely connected concern is to develop methods to deal with propositional logic that will generalise smoothly to first order logic. We wish emphatically to avoid any hacker's devices that will not survive the lift to first order. We want, as far as possible, to keep our algorithmic principles uniform across the whole class of input problems. It is this concern that gives us further reason to explore systems of natural deduction, and attack the problem of how to find or generate proofs as suitably structured patterns of sentences.

When searching for proofs we are seeking to construct tree-like arrays of sentences satisfying local or global constraints on their syntactic patterning. Intelligent – that is, highly constrained – search would be best secured by applying the knowledge we have from proof theory about the shape of proofs in normal form, and the transformations that convert proofs into normal form.

[1] N. Tennant 1992.

Systems of natural deduction offer a rich variety of what might be called *completeness-conserving constraints*. The main theoretical investigations to be presented below concern the existence of various kinds of normal forms for proofs of given problems. Suppose P is an effectively decidable property of (X, A) and F is an effectively decidable and non-trivial – that is, constraining – property of proofs. Then what I shall call a *PF-normal form theorem* is a result of the following form:

> for all X and for all A, if $P(X, A)$ then for all proofs Π of A from any subset Y of X, there is a proof Σ of A from some subset Z of Y such that $F(Z, A, \Sigma)$.

Such a theorem gives constraining heuristic guidance in the search for proofs for the problem X ?– A. One checks whether $P(X, A)$. If so, then one confines one's search to proofs with property F. Ordinary *normalization* theorems are special cases of results of the above form:

> for all X, for all A, and for all proofs Π of A from any subset Y of X, there is a proof Σ of A from some subset Z of Y such that Σ is in normal form.

Note that the precondition P in their statement is trivial, and F is the property of normality as usually understood (that is, "not containing any maximal formula occurrence – an occurrence standing as the conclusion of an introduction rule and as the major premiss of the corresponding elimination rule"). Naturally we exploit this conventional normal form theorem, in that we seek only proofs in such conventional normal form. But we supplement this obvious focusing of our search with further PF-normal form theorems, for non-trivial preconditions P, tailored for service in computational logic.

Another special case of PF-normal form theorems is where the relational property F is restricted to the arguments Z and A and is *persistent*, in the sense that if $F(Z, A)$ and Z is a subset of Y, then $F(Y, A)$:

> for all X and for all A, if $P(X, A)$ then, for all proofs Π of A from any subset Y of X, there is a proof Σ of A from some subset Z of Y such that $F(Z, A)$.

Results like this are called *filters*. They say, essentially, that if $P(X, A)$ then the problem X ?– A has a proof only if $F(X, A)$. So if we are given X ?– A, and can determine that it has property P but lacks property F, then we know that there is no proof to be had.

One has to be careful to prosecute the enquiry into constraints with great care, so as to avoid producing an incomplete proof-finder for one's chosen logical system. One has to pay attention, that is, to the *compossibility* of constraints. For suppose one has a series of P_iF_i-normal form theorems $(i = 1, \ldots, n)$. One may have a provable problem that satisfies all the P_i. If one has constrained one's search by using all of the corresponding F_i, then

one must be assured that the F_i are compossible – that is, that there will indeed be a proof satisfying all the properties F_i.

Think of a completeness-conserving contraint F as a spotlight on a surface whose points are proofs. Compossibility then amounts to this: with several spotlights in play, one wants to be sure that there is a region that they all illuminate. When one's constraints are not thus compossible, then one is forced to choose different combinations from among them that are. And here lies the prospect (for computational logic as a branch of cognitive science) of being more or less faithful to the repertoire of human logical competence. Some completeness-conserving constraints may force proofs into a form that is highly unlikely to be happened upon by human reasoners. Others may turn up proofs on which all human avenues of logical enquiry converge. The aim is to produce highly readable proofs that are rigorous and detailed formalizations of intuitive lines of human reasoning. The more succinct arguments that human reasoners would produce in actual logical or mathematical discourse will be homomorphs of these formal proofs under a very natural projection.

Another advantage for a computational logician in working with systems of natural deduction rather than, say, with Beth tableaux or the resolution method, is that debugging one's proof-finding programs is much easier. Suppose, for example, that one discovers a provable problem that one's program fails to prove. When one examines the trace of the computation one is much better able, in a natural deduction system, to locate and isolate the characteristic errors through failure to construct various subproofs of the would-be proof. Clauses of the program correspond to rules of inference; calls of clauses correspond to attempted applications of these rules of inference. The subproblems generated correspond to the subproofs required for successful application of those rules of inference. Diagnosis and debugging are very easy in such a nested environment.

A decidable logic is the simplest case of a decidable theory based on it. One can implement a decidable theory in a variety of ways, ranging from the evidentially miserly to the evidentially generous. Take the *decision problem* for a theory T:

> Find correct Yes/No answers to problems of the form X ?– A, where X is a finite set of premisses and A is a conclusion, and the question mark concerns the relation of deducibility within T.
>
> There is a minimal response to this problem:

Bare oracles One could give a bare oracle for the decision problem: a program that computed Yes/No answers and gave nothing else as output. This would be a case of extreme evidential miserliness. (It goes without saying that the answers would have to be correct. This is true also of the programs involved in the remaining responses.)

Then there are two intermediate responses:

Proof-finders: One can give full reasons for Yes answers, in the form of proofs. Proofs are finite objects that we can check for correctness. It matters not whether we check the correctness of proofs "by hand" or by means of yet another program – a proof-checker. Proof-checkers are not to be confused with proof-finders. The former verify proofhood. The latter find proofs. If we know that the proof-finding program is correct, however, we will not have to check them. Instead, we can use them to convince others who may be sceptical about positive answers. I call such a program a proof-finder.

A proof-finder for a given theory may be complete or incomplete. A complete proof-finder is one that gives at least one proof for each true statement of deducibility in the theory. An example, by contrast, of an incomplete proof-finder for a theory would be one which implemented an axiomatic system of arithmetic (such as Peano-Dedekind arithmetic), for the theory consisting of all true sentences of arithmetic.

A proof-finder could be *complete* – insofar as it would eventually, for any given provable problem, find a proof of it – and yet fail, on some unprovable problems, to yield even bare negative decisions. On these problems it would not terminate. A complete and *bounded* proof-finder is one which will eventually terminate with a negative verdict on any unprovable problem. Even with a complete and bounded proof-finder one cannot tell, from its failure to respond by any given time, whether it had not yet had time to find a proof, or whether indeed there was no proof to be had. One simply has to wait. All one knows is that one will not have to wait for ever.

A *hypercomplete* proof-finder would do even better than a merely complete one: it would provide a distinct formal but faithful representation, in the form of a proof, for each of the possibly many different informal arguments that might serve to establish the validity of the transition, in the theory, from premisses X to conclusion A. Hypercompleteness is of necessity an informal notion, like that of computability; but it is important to bear in mind as an ultimate desideratum. It will turn out, however, that if complete proof-finders written in any version of Prolog based on a depth-first strategy are to have remotely tractable tasks, they must abjure hypercompleteness. Otherwise the proliferation of alternative proofs on backtracking will greatly delay the making of correct negative (and positive) decisions.

Best of all proof-finders would be a hypercomplete and bounded one, which was able to arrange all alternative proofs in some order of ascending complexity. The notion of such order, however, has yet to receive a satisfactory theoretical analysis.

Counterexample-finders: One can give full reasons for No answers, in the form of counterexamples. Finite counterexamples, like proofs, may be checked for correctness. Counterexamples may be used to convince sceptics about negative answers (but see below).

In the case of first-order mathematical reasoning, a counterexample will be a structure – finite or infinite – that forces all the premisses (makes

them true), but does not force the conclusion. There are interesting philosophical problems, however, concerning the status of counterexamples to X ?– A as supposedly semantic objects distinct from proofs. One can have philosophical reservations about the epistemic force or persuasive power of an infinite counterexample *qua* semantic object. In the case of an infinite counterexample, it could be maintained that what we really have in mind is a well-known theory, given by a set of axioms widely assumed to be consistent, and enjoying the counterexample as a model, and that the counterexemplification of X ?– A consists in the proof-theoretic facts that there are proofs of each of the premisses in X from those axioms, and a proof that the conclusion A is inconsistent with those axioms. And one can easily contend that a finite counterexample is, once again, simply another way of coding proof-theoretic facts: facts concerning the existence of proofs of each of the premisses in X from (an obviously consistent set of) axioms categorically describing the finite counterexample, and the existence of a proof that the conclusion A is inconsistent with those axioms.[2]

In the propositional case, another kind of counterexample may be *logical matrices*. These provide functional interpretations of the connectives over a set of designated and undesignated values. They have to be sound for the theory in the sense that every true statement of deducibility in that theory preserves designated value from premisses to conclusion under every assignment of values to the propositional variables involved. A matrix counterexample to X ?– A is then an assignment of values to the propositional variables involved in X and in A on which each premiss in X takes a designated value and A takes an undesignated value. This is a generalization of the familiar matrix $\{T, F\}$ for classical propositional logic, with T designated and F undesignated, and the usual truth-functional interpretations for the connectives. A class of sound matrices is complete for the theory if every false statement of deducibility in that theory has a counterexample using a matrix in that class. A sound and complete class of matrices is said to be characteristic for the theory. Another example of a characteristic class of matrices for a theory is the class of Jaskowski matrices for intuitionistic propositional logic. One well-known way of showing that an axiomatisable propositional theory (including a logic) is decidable is to show that it has a characteristic class of finite matrices. For then the decision procedure can exploit two enumerations: one of proofs, by virtue of axiomatisability, and one of the finite sound matrices, testing the latter effectively to see whether they counterexemplify the problem in hand. Eventually either the first enumeration hits on a proof of the problem, or the second enumeration hits on a counterexemplifying finite sound matrix.

Like a proof-finder, a counterexample-finder for a given theory may be complete or incomplete. A complete counterexample-finder is one that gives at least one counterexample for each false statement of deducibility in the

[2]For a fuller develoment of this view, see Tennant 1986.

theory. Even a sound but incomplete class of matrices can be useful for an algorithm of this third kind. But such a class is usually exploited for the fortuitous benefits it might bestow in attempts by a proof-finder to prune the search tree in the space of possible proofs. Its members could be used to counterexemplify sequents generated by inductive breakdown of deductive problems in those logics. This can yield speed-up in proof search in the context of the proof-finder of which the counterexample-finder is a module even though the matrices be drawn from a class that is incomplete for the logic concerned.

It is possible for a complete and bounded proof-finder also to be a counterexample-finder (and of finite ones at that) without much further ado. The trace of a terminated and unsuccessful search by such a proof-finder for a proof in response to X ?– A is, after all, a finitary object that bears effective scrutiny. It can play the epistemic rôle of a finite counterexample to X ?– A to one who knows how to read it.

Finally, we have the full-blown response to the problem, which is to provide programs that I call:

Judicial reasoners: These are both proof-finders and counterexample-finders. That is, they give full *rationes decidendi* for their positive or negative verdicts on deductive problems. The best among these would be the *rationauts*. A rationaut would be able to give the shortest or simplest proof possible as its answer to any provable problem, and be hypercompletely ready to give, on request, all alternative proofs, in normal form, in ascending order of length, complexity, roundaboutness or what one will; and would be able to give the shortest or simplest counterexample possible as its answer to any unprovable problem.

The way I set about the task at hand is to aim, modestly, for a complete and bounded proof-finder for a well-chosen decidable propositional logic. It is not hypercomplete. But its terminated unsuccessful searches can deliver traces playing the role of counterexamples to the unprovable problems concerned. The algorithm for finding proofs is written in such a way, however, as to allow one to insert various filters on the sub-problems generated during the search. We have such filters anyway in connection with past recorded successes and or failures. In just the same way we could incorporate filters exploiting matrices, say, if we so wished. But we do not actually do so. The only filters we employ, apart from those recalling past successes and failures, concern the various sorts of syntactic relationships that subformulae can bear to containing formulae. That is, we try to exploit only the sort of syntactic evidence that it is reasonable to suppose the human reasoner can easily (and perhaps often subconsciously) detect.

One scientific theory can possess the pragmatic virtues of elegance and simplicity to a greater or lesser degree than another. It is a matter of trained taste on the part of practising scientists to decide between them on the basis of such considerations. So too in cognitive science: one can

prefer computational models of human logical competence that exploit only what meets the eye syntactically, so to speak, to models that employ, say, nine-element Heyting algebras in their filtrations during search.

Our search heuristics will encode the effect of generally available transformations on proofs. It is desirable to get as many of these as possible to be invariant across choice of logical system, so that the proof-finder is more easily adaptable to whatever system one chooses to work in.

A first major challenge will be to combine aspects of the *bottom-up* and *top-down* kinds of search that human logicians undertake when trying to construct suasive arguments.

A second major challenge is to program a proof-finder that is short enough to admit of an *informal proof of correctness*. On our approach we justify the inclusion of every clause in the Prolog program that embodies the proof-finder, by appeal to proof-theoretic considerations concerning the logic in question.

A third major challenge is to exploit the notion of the relevance of premisses to a conclusion in a manner that I would call it *endogeneous* to the proof-finder. One does not want the proof-finder to find proofs indiscriminately, and only thereafter produce one that exhibits genuine relevance in all its inference steps. One wants rather to use the requirements of relevance to avoid the irrelevant inferential directions and focus on the relevant.

References

N. Tennant, *Autologic*, Edinbugh: Edinburgh University Press 1992.

N. Tennant, "The Withering Away of Formal Semantics?", *Mind and Language* 1 (1986), 302–318.

Towards Psychoontology

Jerzy Perzanowski

Introduction

1 Psychoontology is the ontology of the psyche and of related matters. Hence it is a case of particular and applied ontology.

2 Here, following Leibniz's idea, ontology is defined[1] by its characteristic question: How is something possible? More exactly: How is x possible? Now the level of generality of a given ontology depends on the generality of its characteristic question, i.e. on the scope of the variable x. If this is the most general of all, we obtain general ontology, which is the study of the following, most general, version of the ontological question: How is what is possible, possible? To answer it we must provide a reason for being possible, as well as a framework for the study of the ontological space of all possibilities. [2]

3 Particularizations follow by specification of the range of x. For example, by asking How are facts possible? or: How is the world possible?, that is, by searching for reasons for the existence of the world, its mechanism and basic principles, for sources of its regularity, we define *metaphysics* which, by definition, is the ontology of reality. To be precise: description and investigation of the world is a business of science, whereas investigation of its basic and most general principles is a subject of metaphysics.

4 Psychoontology is even more specific. It concerns the specifically human part of the world, the realm of human beings, understood as wholes composed of, *inter alia*, their psyche and their body. The following questions are therefore characteristic for psychoontology: How is a psyche possible? How is cognition possible? How are soul-body or mind-brain connections possible? How is consciousness possible?, and so on. What, however, do these questions mean? Leibnizian questions, to be sure, sound strange to laymen.

[1] For a discussion of general ontology in comparison with particular ontologies see Perzanowski [8].

[2] This is what Wittgenstein in the *Tractatus* named *logical* space.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 287-296

5 Take, for example, the second question on our list: How is cognition possible? To understood it we need, first of all, *categorization.* Cognition is

- a relation (of which arity?, which objects are related?);
- a process (of what type?);
- transfer, or processing, of information.

What more? Now it is clear that a proper framework for the investigation of cognition must include at least all the above items: relations, their arguments, i.e. subjects and co-subjects (things? situations? facts? persons? institutions?), processes, transfer itself, and information. By such a descriptive and conceptual analysis we obtain a quite complex domain, organized in some way. Its investigation is a business of the cognitive sciences, including a suitable applied ontology. To illustrate: In science we are interested in laws governing information transfer, in technology – in rules enabling us to transfer information economically, whereas in ontology we search for those components and features of the world which makes such transfer possible. The psychoontology of cognition deals therefore with the most basic part of this investigation, i.e., with the mechanism of the field under investigation, with that which makes cognition possible, with the most primitive components and the most general principles of the cognitive universe.

6 Psychoontology is not a new subject. It has an at least three-hundred year old tradition starting with the Cartesian problem of psycho-physical connection and growing through the great contributions of Descartes himself, of Spinoza, Leibniz and Kant, and in the last two centuries through the works of many philosophers and scientists. The paper's title seems thereby to be misleading, producing the false impression. Why "Towards"?, if psychoontology is, in a sense, already flourishing? The reason is simple: For further progress we need an exact, formal, complex and sophisticated psychoontology. But, unfortunately, contemporary psychoontology is not sufficiently developed for this to be achieved.

The Task

7 We still need a general ontological framework in which the basic notions of psychology and of the other cognitive sciences can be defined clearly and rigorously, thus enabling a formal machinery in which psychic phenomena emerge in a natural and clear way. To express our task by means of an example: we would be prepared for a formal discussion and development of the ontological content of standard books in the field, like Popper and Eccles' *The Self and Its Brain.*

8 Usually, such an apparatus is borrowed either from physics and from the other natural sciences, or from mathematics, logic and computer science, or from the humanities, and then applied to questions of psychoontology. These

starting points, however, were produced in other domains, with different reasons in mind. In most cases they are too narrow, and hence inappropriate. For example, use of the ontological apparatus prepared for the ontology of physics, which is a part of metaphysics, usually ends with the notorious difficulties of physicalism.

9 To overcome these difficulties and similar shortcomings we must rethink the ontological questions of psychology, find a more general and hence more natural and proper formal apparatus, and next develop, step by step, a logical psychoontology. This should not be done in opposition to tradition. On the contrary, we should borrow as much as is reasonable from the ideas of the old masters, reshaping them to meet the present standards.

10 The above task can be realized in many different ways, which, taken to an extreme, might even seem to be incoherent. The reason is simple: There is a large number of types of ontology, which are so different that they must produce differences in description and explanation. In what follows, I will try to outline several clues leading to a combination psychoontology, i.e. a psychoontology based on a combination ontology.

Two Examples

11 Let us start with a remark concerning what is probably the most popular post-Aristotelian ontology, that of things and properties. Clearly, it is a descriptive ontology. In its way of looking at the world, everything is classified with respect to the basic relation connecting things with their properties. Therefore, everything is either such-and-such a thing or such-and-such a property or, in some cases, both. Therefore, for post-Aristotelian (or, more specifically, post-Brentanian) psychoontology the basic question is to find an adequate definition of psychological properties (and also of things). Basic psychological and cognitive notions have also be defined in this framework. Take, for example, a mind.[3] Is it a thing? Which one? Or is it a property? Again, which one? In general: which properties (things) are psychological?[4] This is quite a lot of obscure questions; they are difficult to answer even for the most serious Aristotelians, such as Chisholm.

12 Consider now an example taken from Wittgenstein's *Tractatus.* Thoughts are defined there twice: in thesis

> 2. A logical picture of facts is a thought;

and in thesis

> 3. A thought is a proposition with a sense.

[3] Mine, or yours, if you like.

[4] See the lecture "The Marks of the Purely Psychological" given by R. M. Chisholm at the 9th International Wittgenstein Symposium, 1984; Chisholm [2]-[5].

One conclusion is immediate:

(1) A logical picture of facts is (or equals) a proposition with a sense.

What more? If we really want to capture Wittgenstein's idea we must answer several fundamental questions concerning the *Tractatus*: What is a fact? What is a picture? What is a proposition? What is a sense? What is a configuration? What is its structure? What is its form? What is form in general? What is logical form?, and so on. In the *Tractatus*, these matters are explained *informally* in terms of the Tractarian version of combination ontology, which is rather obscure. Its clarification and full understanding needs therefore a formal development of the general combination ontology with a discussion of its Tractarian peculiarities.

Combination Ontology

13 Assume that, with respect to the primitive relation *simpler than*, all objects are divided into two families: simple objects, called *elements*, and non-simple objects, called *complexes*. The chief idea of combination ontology[5] is that complexes are combinations built up from simpler objects according to their internal traits, or determiners. The traits of an object constitute its form. By the fundamental idea of combination ontology, everything happens because of form. In particular, form determines all the possible combinations in which an item can be involved, and – in the case of a complex – form determines the net of bonds fusing its components into one in such a way as to form a whole.

14 Combination ontology is a rather complex enterprise. It is, in fact, an advanced and complex theory of analysis and synthesis. Its starting part is therefore a general theory of analysis and synthesis, in which the ontological universe of all objects **OB** is treated as ordered by two conjugate relations: an analytical one, *simpler than*, $<$, and a synthetic one, *to be a component of*, $\subset$. Two associate operators: the *analyser* α and the *synthesiser* σ are also considered. The ontological universe of analysis and synthesis can thereby be understood as the quintuple $\langle \mathbf{OB}, <, \alpha, \subset, \sigma \rangle$. The general theory of analysis and synthesis considers interconnections between the two relations introduced above and suitable operators. Its basic observation is that analysis and synthesis are dual, but not invertible in a simple way.

15 They also differ from a methodological point of view. The theory of analysis is rather straightforward and immediate, whereas the theory of synthesis is more complex and subtle. Difficulties concerning synthesis are connected with the characterization of wholes, their unification and with different types of synthesis. To meet these difficulties two approaches are introduced: the first, external and purely relational, and the second, internal and deeply

[5]For more details see Perzanowski [7], [9] and [12].

modal. The relational approach has itself two editions: one, which is simply the general theory of analysis and synthesis outlined above, and another, locative and connectional, which starts with a very basic reflection on the nature of combination.

16 What is a combination? Clearly, it is a complex of a special type. It is any natural configuration of objects, i.e., each collection of correlated and connected objects. Each combination has thereby five primitive correlates: its *stuff* (i.e. the class of all its parts), its *substance* (i.e. the class of all its simples), its *structure* (i.e. the way in which its components are related), its *form* (anything which makes its structure possible), and its *network* (i.e. the net of bonds fusing its components together). Several secondary correlates are also present.

17 The idea of combination can therefore be decomposed into three more primitive ideas: location, correlation and connection:

$$\text{combination} = \text{location} + \text{correlation} + \text{connection}$$

To be in a combination means to be located in it, and to be correlated and connected with other components. Two ideas – that of location and that of correlation – are purely relational, whereas the idea of connection is more dynamic and modal. Now, if we like to have a really general combination ontology, sufficient to cover both the realm of psyche and the realm of matter, we must have suitably general ontologies of location, correlation and connection. And this is our task here.

Correlation

18 The ontology of correlation is immediate. It is the general ontology (or calculus) of relations,[6] for correlation simply means relation. Its algebraic version, which depends on certain assumptions, is the lattice theory of G. Birkhoff,[7] which is currently the best description of the structures of combinations which we have.

Location

19 The ontology of location is also available.[8] It formalizes location in *any* given relational frame using the thesis: an item x is located in y iff each part of x is related to y.

20 This idea can be expressed formally in the following way: Fix a non-empty set U and a binary relation E on U. Next consider all derivative binary relations defined in this framework by means of the universal quantifier and implication.

[6]Schröder [14] and Tarski [15].
[7]Birkhoff [1] and Grätzer [6].
[8]Perzanowski [10] and [12].

20.1 In this way we obtain at least two parthood relations:

$$xPy \quad := \quad \forall z(zEx \to zEy),$$

Leśniewski's parthood relation: x is a part of y iff each item which is E-related to x is also E-related to y, or, in a more familiar way, x is a part of y iff everything in x is also in y.

$$xCy \quad := \quad \forall z(yEz \to xEz)$$

the covering relation dual to the Leśniewskian one: x is covered by y iff each container of y also contains x.

In fact, two further relations of this sort can also be introduced.

20.2 Extending the above procedure further we reach the following two formulas expressing the idea of location:

$$xLy \quad := \quad \forall z(zPx \to zEy);$$

x is located in y iff any part of x is related to y, and

$$xAy \quad := \quad \forall z(yCz \to xEz)$$

x is allocated in y iff it is related to any cover of y.

Again two further locative relations can be considered.

21 We can check that the above definitions formalize basic intuitions concerning location. Among other things we can prove that

(2) Both location and allocation are logically stronger than the starting relation: $L \leq E, A \leq E$.

Hence the proper locative structures are those in which the starting relation E is equal to its locative counterparts L and A:

IL $E = L$: To be is to be located in, or

EL $E = A$: To be is to be allocated in.

Now it is only a question of routine calculation to see that locative spaces satisfying at least one of the above axioms are quite similar to the usual preorders, hence they are regular and rich structures. They indeed offer a very natural framework to study location, which is defined for *any* starting relational frame. Therefore, the relational location is not limited to the usual cases of space-time, physical location and it is general enough to be used as a basis for studying cases of extra-physical location. [9]

[9] For example the location of my thoughts when writing this essay.

Connection

23 The characterization of connection is more difficult, for it is not a purely relational matter. Now we are interested in questions concerning the internal structure of a given combination, which is treated here as a whole. We are looking for a mechanism for its unification and therefore we search after determiners forming the combination from its components, i.e., for its form.

24 There are several approaches to the investigation of a given object's form. A standard mathematical technique is to choose a group of transformations and characterize its invariants, whereas a logical method[10] proceeds by treating suitable determiners or traits as ontological modalities which are subjects of logical treatment.

25 To this end, let us use the most primitive pair of ontological modalities: making possible $\mathrm{MP}(\cdot,\cdot)$, and making impossible $\mathrm{MI}(\cdot,\cdot)$. Take a given combination x and two arbitrary components of it, y and z. We can think of the latter as connected if each of them can be combined with the other: i.e. y and z are both connected in a combination x iff there is a smallest part of x containing both y and z, say $y \cup z$, such that each of them make $y \cup z$ possible:

$$\begin{aligned} C(x;y,z) \quad := \quad & yPx \wedge zPx \wedge \exists(y \cup z)(yP(y \cup z) \wedge zP(y \cup z) \wedge \\ & \forall u(yPu \wedge zPu \to (y \cup z)Pu) \wedge \mathrm{MP}(y, y \cup z) \wedge \mathrm{MP}(z, y \cup z) \end{aligned}$$

Clearly, properties of the connection operator C depend on properties of the ontological modality *making possible*, the theory of which is known to be complex and rich.[11]

Production

26 Any production is by analysis and synthesis. The basic product of a given synthesis is, of course, combination itself. There are also secondary products, including properties of the combination and some phenomena caused by the synthesis and connected with the combination's occurrence. For our purposes some secondary products are even more important than the basic one. They include, *inter alia*, the usual properties of things, their determiners, traits and other qualities, several types of fields including physical ones, and several dynamical characters and states, like propensities, homeostasis, equilibrium and stabilization.

27 Restrictive syntheses, in which only combinations are produced, should be distinguished from nonrestrictive ones.

[10] They are used in my [7], [9] and [12] .

[11] Perzanowski [7] and [12].

Emergence

28 Emergent syntheses are special cases of nonrestrictive ones, in the course of which the rules for the process of synthesis itself are changed. They are like games during which we not only produce game-situations, but sometimes also change rules of the game itself. The most important case of an emergent synthesis we know is emergent evolution.[12]

29 Now the basic question concerning emergence is the ontological question: How is it possible? In the framework of combination ontology the answer is rather straightforward. Emergence occurs by means of a nonrestrictive synthesis in which *new* qualities are produced. Qualities form the combination's form which, we remember, fully determines its synthesis. When new qualities are in play then the process of synthesis can, and usually is, changed by something which sometimes seems to be a case of downward causation.[13]

30 There is nothing mysterious in either the idea of emergent evolution (synthesis) or in that of downward causation, if a suitable theory of qualities is provided. Such a theory is a chapter of the logic of qualities, which, according to the fundamental insight of Leibniz, is the crux of combination ontology, i.e., the general theory of analysis and synthesis.

Existence

31 Existence is certainly among the most important products of a special kind of synthesis. The following conditions are necessary for the existence of an object x:

i) x must be a combination, and so *a fortiori* it must be a complex;

ii) x must be coherent, i.e. possible;

iii) x must be condensed and stable; and, according to Leibniz:

iv) x must be compatible with a maximal number of other possibilities.

32 The idea of existence is therefore quite complex and can be decomposed[14] into at least eight more primitive ideas:

$$\text{existence} = \text{combination} + \text{coherence} + \text{condensation} + \text{stabilisation} + \text{maximalisation} + \dots,$$

or

$$\text{existence} = \text{location} + \text{correlation} + \text{connection} + \text{coherence} + \text{condensation} + \text{stabilization} + \text{maximalization} + \dots$$

[12] Popper & Eccles [13].
[13] Popper & Eccles [13].
[14] Like the idea of combination; see §17.

Thoughts Revisited

33 The justification of all research, like any enterprise, is by its fruits. Our ontological machinery is so developed that, if reasonable, it should offer us an adequate and natural way to deal with basic cognitive notions. Is our ontological framework correct and useful? To see the point, let us return to an example of *thoughts*, discussed previously in §12 in the ontological terms of Wittgenstein's *Tractatus*.

34 First of all, thoughts are (logical) pictures of facts, hence facts, hence *existing combinations*. Notice that by §§17, 31 and 32 we now have a much more developed and sophisticated apparatus to study combinations. On the other hand, the theory of synthesis and its subtheory of qualities (see §§16, 23 25, 28 29) are prepared especially to deal with the notion of form which, as any true Wittgensteinian scholar knows, is the crucial notion of the *Tractatus*.

35 Observe also that by §§19-22 we can speak, without any metaphor involved, about the location of thoughts both as pieces of the ontological space and as inhabitants of mine, yours, or his/her/its mind. In fact, in the present framework we can define *a mind* either as any container of thoughts, i.e., an item locating the maximal number of co-located thoughts, or as the minimal combination of thoughts.

Instead of a conclusion

35 Combination ontology indeed offers room for advanced psychoontological research. This is not a solution but an opportunity, which can and should be taken.

References

Birkhoff G., *Lattice Theory*, Providence RI: AMS 1940.

Chisholm R. M., *Person and Object*, La Salle Ill: Open Court 1976.

Chisholm R. M., *The First Person*, Minneapolis: University of Minnesota Press 1981.

Chisholm R. M., "Thought and Its Reference", *American Philosophical Quarterly* 14 (1977), 167–172

Chisholm R. M., "Properties Intentionally Considered", in *Language and Ontology: Proceedings of the 6th International Wittgenstein Symposium*, Vienna: Hölder-Pichler-Tempsky 1982, pp. 117–121.

Grätzer G., *Lattice Theory*, San Francisco: Freeman 1971.

Perzanowski J., *Logiki modalne a filozofia (Modal Logics and Philosophy)*, Cracow: Jagiellonian University Press 1989.

Perzanowski J., "Ontologies and Ontologics", in *Logic Counts*, ed. E. Żarnecka-Biały, Kluwer 1990, pp. 23–42

Perzanowski J., "Towards Post-Tractatus Ontology", in *Wittgenstein – Towards a Revaluation: Proceedings of the 14th International Wittgenstein Symposium*, Vienna: Hölder-Pichler-Tempsky 1990, pp. 185–199

Perzanowski J., "Locative Ontology. Parts 1 - 3", *Logic and Philosophy* 1, 1993.

Perzanowski J., "The Way of Truth", in *Formal Ontology*, ed. R. Poli and P. Simons, Kluwer 1994.

Perzanowski J., *Badania Onto-Logiczne (Onto-Logical Investigations)*, work in progress.

Popper K. and Eccles J. C., *The Self and Its Brain*, Berlin: Springer 1977.

Schröder E., *Vorlesungen über die Algebra der Logik* vol. 3: *Algebra und Logik der Relative*, Leipzig: Taubner 1895.

Tarski A., "On the Calculus of Relations", *Journal of Symbolic Logic* 6 (1941), 73–89.

Wittgenstein L., *Tractatus Logico-Philosophicus*, London: Routledge and Kegan Paul 1922.

Defeasible Reasoning Based on Constructive and Cumulative Rules

Gerhard Schurz

1 Nonmonotonic Reasoning: Introduction and Problems

A great part of human reasoning is based on uncertain laws like

$$\text{Birds normally can fly,} \quad \text{formally} \quad \text{Bird}(x) \Rightarrow \text{Can-Fly}(x) \qquad (1)$$

($\Rightarrow$ for uncertain implication). Such laws admit exceptions, but are stated in a nonprobabilistic (nonnumerical) way. This was discovered by philosophers like Scriven (1959) and Rescher (1964), and became a hot topic since the 1980s in the branch of AI called nonmonotonic logic (or nonmonotonic reasoning). As argued in Schurz (1993), uncertain laws play an important role in common-sense as well as in scientific and philosophical reasoning; they also are exhibit remarkable relations to the functioning of neurons with distinct activating and inhibiting synapses.

Nonmonotonic logic starts from the observation that from the uncertain law (1) plus the instantiated antecedent Bird(tweety) ("Tweety is a bird") we are allowed to derive the instantiated consequent Can-Fly(tweety) *only* as long as *nothing else* is derivable from our knowledge base **K** which implies that Can-Fly(tweety) is false, i.e. that Tweety is an "exceptional" bird as regards flying. For instance, if we *also* know Penguin(tweety) and the strict (deterministic) law $\text{Penguin}(x) \rightarrow \text{Can't-Fly}(x)$ ($\rightarrow$ for material implication; variables are to be read as universally quantified), then the detachment from (1) is no longer allowed. We also say, that law (1) is *defeated.*

Thus, inference from defeasible laws is nonmonotonic: additional knowledge may make previously derived consequences underivable. In their seminal paper, McDermott and Doyle (1980) suggested the formalisation of uncertain laws with help of material implication and a possibility operator, in the schematic form

$$\text{Bird}(x) \wedge \Diamond\text{Can-Fly}(x) \rightarrow \text{Can-Fly}(x)$$

where $\Diamond B(x)$ is derivable from the knowledge base **K** if $\neg B$ is *not* derivable from it. Alternatively, Reiter (1980) reconstructed these inferences as *default rules*

$$\frac{\text{Bird}(x) \quad M\text{Can-Fly}(x)}{\text{Can-Fly}(x)}$$

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 297-310

with the same condition for 'M' as for '$\Diamond$'. Moore (1985) has called nonmonotonic reasoning in the McDermott/Doyle reconstruction *autoepistemic* reasoning, because the nonmonotonic possibility clause which refers to what is not derivable in **K** is part of the uncertain law itself (and hence part of the object language), while in Reiter's *default* reasoning system it is only part of the metalanguage inference rules. However, Konolige (1988) has shown that both reconstructions are intertranslatable, so this difference is not of primary importance. The version of nonmonotonic reasoning developed in this paper, *defeasible* reasoning, lies between both reconstructions: the uncertain laws without the nonmonotonic clause, like $\text{Bird}(x) \Rightarrow \text{Can-Fly}(x)$, will be part of the object language (in contrast to Reiter's defaults, which are rules), while the nonmonotonic clause will be encoded in the metalanguage inference rules.

The primary problem, rather, is how to define the set of nonmonotonic consequences of a given knowledge base **K** in a reasonable way. Let $\mathrel{|\!\sim} \in \mathcal{P}(\mathcal{L}) \times \mathcal{L}$ stand for the nonmonotonic derivability relation (where $\mathcal{L}$ denotes a fixed language). We abbreviate $\Gamma \mathrel{|\!\sim} \Delta$ iff $\Gamma \mathrel{|\!\sim} A$ for all $A \in \Delta$, and $\mathbf{C}(\Delta) := \{A \Delta \mathrel{|\!\sim} A\}$, the consequence operation. (Here A,B,... range over formulas and Γ, Δ,... over sets of them). We expect from a reasonable notion of consequence that $\Delta \subset \mathbf{C}(\Delta)$ and $\mathbf{C}(\Delta) = \mathbf{C}(\mathbf{C}(\Delta))$, i.e. that $\mathbf{C}(\Delta)$ is a *fixed point* of the operation **C**. As is well known, a necessary and sufficient condition for this is that $\mathrel{|\!\sim}$ satisfies these two axioms:

$$\frac{}{\Gamma \mathrel{|\!\sim} \Gamma} \textit{ Reflexivity} \qquad \frac{\Gamma \mathrel{|\!\sim} \Delta \quad \Gamma \cup \Delta \mathrel{|\!\sim} \Sigma}{\Gamma \mathrel{|\!\sim} \Sigma} \textit{ Cut}$$

But we want more: in the perspective of the traditional concept of proof, we want to axiomatize $\mathrel{|\!\sim}$ by a small (finite) number of *derivation rules* of the form

$$\frac{X \subseteq \Gamma \quad Y \not\subseteq \Gamma}{Z(X,Y) \subseteq \mathbf{C}(\Gamma)} \textit{ Constr}$$

where X, Y, Z are schematic expressions for *finite* sets of formulae, and $Z(X,Y)$ depends on X and Y. The point is that the preconditions of such rules do not refer to $\mathbf{C}(\Gamma)$ (i.e. to what *will* ultimately be derivable), but only to Γ (i.e. to what has been derived so far). This enables a constructive definition of $\mathbf{C}(\Delta)$ as the smallest formula set being closed under these rules, in other words, as the set theoretic limit (i.e. union) of the stepwise application of these rules to Δ, possibly infinitely many times. We call systems in which $\mathbf{C}(\Delta)$ is definable in this way *rule-constructive*. A rule-constructive definition will be *unique* only if the resulting limit is independent of the order in which the rules are applied. This implies that $\mathrel{|\!\sim}$ has to satisfy a further axiom, namely:[1]

[1] Cf. Kraus *et al.* 1990: 178. It is called "restricted monotonicity" by Gabbay 1985.

$$\frac{\Gamma \mathrel{|\!\sim} \Delta \quad \Gamma \mathrel{|\!\sim} \Sigma}{\Gamma \cup \Delta \mathrel{|\!\sim} \Sigma} \textit{ Cautious Monotonicity}$$

Together with *Cut*, this axiom implies that if $\Gamma \mathrel{|\!\sim} \Delta$ and $\Gamma \cup \Delta \mathrel{|\!\sim} \Sigma$, then also $\Gamma \mathrel{|\!\sim} \Sigma$ and hence $\Gamma \cup \Sigma \mathrel{|\!\sim} \Delta$; so the derivation process will be *cumulative*, i.e. independent from the order in which rules are applied, and thus will converge to a *unique* fixed point $\mathbf{C}(\Delta)$. For this reason, the combination of *Cut* and *Cautious Monotonicity* is also called 'cumulativity' (after Makinson 1989).

Prima facie, nonmonotonic reasoning systems are neither rule-constructive nor cumulative. For their inference rules have the general iterative form

$$\frac{\Gamma \mathrel{|\!\sim} X \quad \Gamma \not\mathrel{|\!\sim} Y}{\Gamma \mathrel{|\!\sim} Z(X,Y)}$$

The problem is the non-derivability condition (which is lacking in monotonic rule based systems). First of all, it is not rule-constructive, because its precondition refers to what is derivable in the limit. Its iterative reformulation in terms of **C** is

$$\frac{X \subseteq \Gamma \quad Y \not\subseteq \mathbf{C}(\Gamma)}{Z(X,Y) \subseteq \mathbf{C}(\Gamma)}$$

which, in distinction to *Constr*, refers in its precondition to the fixed point $\mathbf{C}(\Gamma)$.

But even if we confine our attention to rule-constructive nonmonotonic systems, based on rules of the form *Constr*, they will generally *not* be cumulative. For, even if derivability depends only on what has or has not been derived so far, the limit approximated by the iterative application of rules will be sensitive to their order. There will be several fixed points: the derivability relation will be $\mathrel{|\!\sim}$*-ambiguous*.

This is the much debated *multiple extension* problem (fixed points are also called 'extensions'). To illustrate it (in the McDermott/Doyle style), assume

$$\begin{aligned} \mathbf{K} \;=\; \{&\text{Quaker}(x) \wedge \Diamond\text{Pacifist}(x) \rightarrow \text{Pacifist}(x), \text{Quaker}(\text{nixon}), \\ &\text{Conservative}(x) \wedge \Diamond\neg\text{Pacifist}(x) \rightarrow \neg\text{Pacifist}(x), \\ &\text{Conservative}(\text{nixon})\}. \end{aligned} \tag{2}$$

Assume instantiation, MP and the $\Diamond$-rule are our only rules, and we apply them in the constructive sense. Adding first $\Diamond\text{Pacifist}(\text{nixon})$ and then $\text{Pacifist}(\text{nixon})$ gives us $\mathbf{K} \cup \{\text{Pacifist}(\text{nixon})\}$ (plus possibility clauses plus classical consequences) as one fixed point. Here, $\Diamond\neg\text{Pacifist}(\text{nixon})$, which was derivable at the start, no longer is derivable. *Vice versa*, by adding first $\Diamond\neg\text{Pacifist}(\text{nixon})$ and then $\neg\textit{Pacifist}(\textit{nixon})$, we obtain $\mathbf{K} \cup$

{¬Pacifist(nixon)} (plus possibility clauses plus classical consequences) as a second fixed point. Even worse, if we first add ◊Pacifist(nixon) and then ◊¬Pacifist(nixon), applying MP two times brings us inconsistency (i.e. $\mathcal{L}$) as a third fixed point.

To avoid inconsistency, McDermott and Doyle have defined a fixed point **F** in the following non-constructive way (that is, where '**F**' occurs in its definiens): **F** is called a fixed point of **K** iff **F** is the set of classical consequences of $\mathbf{K} \cup \{\Diamond A | \neg A \notin \mathbf{F}\}[\cup\{\Box A | A \in \mathbf{F}\}]$ (see McDermott and Doyle 1980, p. 51; the addition in square brackets is an the improvement due to Moore 1985, p. 86). Inconsistent sets are now no longer fixed points of a given *consistent* **K**, but the multiple extension problem remains: in the above example, we have now two fixed points, one entailing Pacifist(nixon) and the other ¬Pacifist(nixon). Reiter's definition of a fixed point is similar (Reiter 1980, p. 89). The alternative is to maintain a constructive understanding of the rules, and to "cross out" all inconsistent extensions in a later step. This idea is realized in the system of Poole (1988), which, in one of its versions, is equivalent to that of Reiter (1980). There are different ways to handle the multiple extension problem. Reiter (1980, p. 94) regards each fixed point as a possible consequence set of a given knowledge base; in other words, he just sticks with with the ambiguity of $\mid\!\sim$ - which is rather unsatisfactory. McDermott and Doyle define $\mathbf{C}(\mathbf{K})$ as the intersection of *all* fixed points of **K**, with the result that if **K** has two fixed points, and if A is in one of them but not in the other, and *viceversa* for B, then only the disjunction $A \vee B$ will be in $\mathbf{C}(\mathbf{K})$. As a result of the definition by means of an intersection, $\mid\!\sim$ satisfies the cumulativity axiom. Still, the definition is not rule-constructive: rather, we first have to produce all fixed points and then find out what is contained in all of them. But there is a deeper reason why this definition is unsatisfactory: it does not take into account relations of *specificity*. If we have two uncertain (defeasible) laws with instantiated antecedents and conflicting consequents, where one antecedents gives *more specific* information about the instance than the other, then we intuitively *prefer* the more specific law: it "fires", while the less specific one is defeated (or "blocked"). A well-known example is

$$\begin{aligned}\mathbf{K} \;=\; \{&\text{Mammal}(x) \Rightarrow \text{Can't-Fly}(x), \text{Bat}(x) \Rightarrow \text{Can-Fly}(x),\\ &\text{Bat}(x) \rightarrow \text{Mammal}(x), \text{Mammal(dracula)}, \text{bat(dracula)}\}.\end{aligned}$$

The multiple extension approach gives us two extensions, one in which Dracula can't and one in which he can fly, whence only the tautology that he can't *or* can fly will be derivable. Intuitively, however, we conclude that Dracula *can* fly, because being a bat is a *more specific* information than being a mammal according to the strict law $\text{Bat}(x) \rightarrow \text{Mammal}(x)$. This case, where the relation of being more specific is based on a strict law is also called *strict* specificity, in distinction to *defeasible* specificity, where the law underlying the specificity relation is defeasible. Also in this case, we prefer

the more specific law; for example (cf. Delgrande 1988, p. 80; Nute 1991, §8):

$$\begin{aligned}\mathbf{K} \;=\; \{&\text{University-Student}(x) \Rightarrow \text{Unemployed}(x),\\ &\text{Adult}(x) \Rightarrow \text{Employed}(x), \text{University-Student}(x) \Rightarrow \text{Adult}(x),\\ &\text{University-Student(peter)}, \text{Adult(peter)}\}.\end{aligned}$$

We intuitively conclude that Peter is unemployed, because being a university student is more specific than being an adult.

According to the *specificity approach*, if two laws have conflicting consequents, then the law with the more specific antecedent is always preferred. In a case like the Nixon example (2) above, where our knowledge base **K** contains no information according to which one of the antecedents is more specific than the other, the specificity approach tells us to remain sceptical, since both laws defeat each other and nothing can be concluded. Note that in our system, if two consequents are conflicting then their disjunction is always strictly (classically) derivable, so the skeptical attitude brings no disadvantage compared to the intersection approach.

The idea of more (or maximal) specificity has been discovered in philosophy of science, namely in the theory of inductive-statistical explanation of Hempel (1965) and his followers (cf. Schurz 1988, 1991). In recent years, several approaches to nonmonotonic reasoning have picked up some sort of specificity considerations. Poole's system allows for specificity priorities between conflicting defaults, but at the cost of naming defaults and adding for each priority a new "constraint" (Poole 1988, pp. 35-7). Etherington (1987, p. 49) and Brewka (1991, §5.4) suggest the introduction of a partial ordering relation among defaults. The most elegant way seems us to built the specificity conditions into the derivability rules. This is the strategy of Nute's defeasible reasoning system (Nute 1988, 1991; for similar technique cf. Delgrande 1988). The specificity approach guarantees that there will always a unique fixed point. This is a necessary but not sufficient precondition for our aim. In a preliminary version, the derivation rule of the specificity approach has to run as follows: If $A(x) \Rightarrow B(x) \in \mathbf{K}$, $\mathbf{K} \mid\!\sim A(a)$, and if, for each $C(x) \Rightarrow \neg B(x) \in \mathbf{K}$, either $\mathbf{K} \not\mid\!\sim C(a)$ or $A(x)$ is more specific than $C(x)$, then $\mathbf{K} \mid\!\sim B(a)$. This derivation rule still contains in its precondition a non-derivability clause and thus is not rule-constructive. The only approach I know which satisfies the aim of a rule-constructive cumulative derivation procedure is the defeasible reasoning framework of Nute (1991). I will base the following considerations on Nute's framework and will "build up" from that basis.

2 Defeasible Reasoning: The Basic System

The basic idea of Nute (1991) is to let the inference rules produce not only all the formulas which are derivable (from a given base **K**), but simultaneously also all those which are *not* derivable. The idea is very close to the

"negation by failure" property of PROLOG, where 'not(A)' is returned iff the search for an A-proof (in a given base $\mathbf{K}$) fails. Indeed, the predecessor of Nute's (1991) system, his (1988) system, was a PROLOG-implementation of defeasible reasoning. To implement this idea we let '$\mathbf{K} \nvdash n(A)$' stand for 'A is derivably not derivable from $\mathbf{K}$', and construct derivation rules simultaneously for formulas of A-shape and of $n(A)$-shape. (The choice of whether $n(A)$ is a formula of the object language, or a metalanguage construction, is optional.) The crucial feature of this method is that it enables to replace the non-derivability preconditions of iterative rules by derivability conditions for formulas of the form $n(A)$. This makes the system rule-constructive *and* cumulative. Our basic system will differ in some respects from Nute's. While Nute describes his system by means of proof trees, we just state the rules behind those trees. For sake of space we only present the basic part of Nute's system: we skip his "might-defeaters" (§1), and also the complications, in order to handle "preempting defeaters" (§6). Instead of Nute's operator 'd' we use a separate sign '$\vdash$' for strict derivability, i.e. derivability with help of facts and strict laws alone, in distinction from defeasible derivability $\mid\!\sim$. The derivation expressions for formulas of the form $n(a)$ must then be interpreted as follows: $\mathbf{K} \vdash n(A)$ means that is is strictly derivable from $\mathbf{K}$ that A is not strictly derivable from $\mathbf{K}$, and $\mathbf{K} \mid\!\sim n(A)$ means that it is defeasibly derivable from $\mathbf{K}$ that A is not defeasibly derivable from $\mathbf{K}$.[2] Most importantly, our basic system improves Nute's conditions for defeasible derivability in one respect explained later.

The language of our basic system is the fragment of the first order language corresponding to PROLOG but extended by a classical negation sign $\neg$ (in distinction to the operator 'n' corresponding to PROLOG's negation by failure). From now on, A, B,... denote *literals*, i.e. (open or closed) atomic formulas or their negations, and Γ, Δ,... stand for *finite* sets of them. A *knowledge base* is a pair $\mathbf{K} = \langle \mathbf{F}, \mathbf{L} \rangle$, where the set $\mathbf{F}$ of *facts* is a finite set of closed literals, and the set $\mathbf{L}$ of *laws* is a finite set of *strict* laws of the form $\Gamma \rightarrow A$ or *defeasible* laws of the form $\Gamma \Rightarrow A$, where Γ and A consist of *open* literals. As in PROLOG, sets are identified with the conjunction of their elements, so we need no extra conjunction symbol (in particular, $\bigwedge\{A\} = A$). $\Gamma \mid\!\sim \Delta$ abbreviates $\Gamma \mid\!\sim A$ for *all* $A \in \Delta$, and keep in mind that $\Gamma \mid\!\sim n(\Delta)$ abbreviates $\Gamma \mid\!\sim n(A)$ for *some* $A \in \Delta$ (similarly for $\vdash$). The domain of possible instantiations of the laws in $\mathbf{L}$ is restricted to constants occurring somewhere $\mathbf{K}$ and hence is finite. We let i, j,... range over functions which associate with each variable in $\mathbf{K}$ some constant in $\mathbf{K}$. Ai stands for the i-instantiation of A, i.e. the result of replacing each variable z in A by $i(z)$; similarly $\Gamma i := \{Bi | B \in \Gamma\}$. The first pair of rules concerns strict derivability (see Nute 1991, §3):

[2]Other "combinations", e.g. that it is defeasibly derivable from $\mathbf{K}$ that A is not strictly derivable from $\mathbf{K}$, either make no sense or are not needed.

$(M+)$	$\langle \mathbf{F}, \mathbf{L} \rangle \vdash A$	if *either* $A \in \mathbf{F}$ *or* there exists $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A = A'i$ such that $\langle \mathbf{F}, \mathbf{L} \rangle \vdash Gi$.
$(M-)$	$\langle \mathbf{F}, \mathbf{L} \rangle \vdash n(A)$	if $A \notin \mathbf{F}$ *and* for each $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A = A'(i)$ it holds that $\langle \mathbf{F}, \mathbf{L} \rangle \vdash n(Gi)$.

Observe that each rule of the form 'Z *if* Y *or* X' is equivalent to two rules 'Z if Y' and 'Z if X'. So $(M+)$ really consists of two rules, a *start rule* (with the lefthand conjunct after the 'if') and an *iterative rule* (with the righthand conjunct). This distinction will turn out to be important below. Here and in the following, each positive rule ('+') has a negative dual ('−') for deriving n-shape formulas. If the positive rule contains all preconditions for its postcondition (stuck together by disjunction instead of being split up into several rules), then the dual can always be obtained by negating the precondition of the positive rule and replacing all phrases of the form 'not $\vdash (\mid\!\sim) A$' by '$\vdash (\mid\!\sim) n(A)$' (and applying classical transformation rules).

We turn to derivability from defeasible laws. To make the rules shorter, we introduce two abbreviations. 'Morespec($\Gamma i, \Delta j/\mathbf{L}$)' stands for '$\Gamma i$ is more specific than Δj with respect to $\mathbf{L}$', and 'Conflict($Ai, Bj/\mathbf{L}$)' for 'Ai and Bj are in conflict with respect to $\mathbf{L}$'.

(Spec+)	Morespec($\Gamma i, \Delta j/\mathbf{L}$)	if $\langle \Gamma i, \mathbf{L} \rangle \mid\!\sim \Delta j$, but $\langle \Delta j, \mathbf{L} \rangle \mid\!\sim n(\Gamma i)$.
(Spec−)	notMorespec($\Gamma i, \Delta j/\mathbf{L}$)	if either $\langle \Gamma i, \mathbf{L} \rangle \mid\!\sim n(\Delta j)$ or $\langle \Delta j, \mathbf{L} \rangle \mid\!\sim \Gamma i$.

Note that the relation of specificity is relative to the given set of laws $\mathbf{L}$ and is formulated for given instantiations of two law antecedents.[3] We need a dual version of (Spec+), because in iterative rules the phrase 'is not more specific" has to refer only to positive derivability conditions.

In the next definition we take into account that law consequents do not only conflict when one is the negation of the other, but also, if from one the negation of the other is strictly derivable. Consider the following modified Nixon example, where $ND(x)$ stands for 'x is an adherent of nuclear deterrance':

$$\mathbf{K} = \{\text{Quak}(x) \Rightarrow \text{Pac}(x), \text{Cons}(x) \Rightarrow ND(x), \quad \text{Pac}(x) \rightarrow \neg\neg ND(x), \text{Quak}(\text{nixon}), \text{Pac}(\text{nixon})\}. \tag{3}$$

Here also both defeasible rules defeat each other, and so, intuitively, we conclude nothing. We define:

[3]This is because $\mathbf{L}$ may already contain some constants which may be merged by those introduced by the instantiation, whence the specificity relation between the law antecedents will not always coincide with that between their instantiations.

(Confl+) $\quad \text{Conflict}(Ai, Bj/\mathbf{L}) \quad$ if $\langle Ai, \mathbf{L}\rangle \vdash \neg Bj$
(Confl−) $\quad \text{notConflict}(Ai, Bj/\mathbf{L}) \quad$ if $\langle Ai, \mathbf{L}\rangle \vdash n(\neg Bj)$

We assume that our derivability systems masters the rules of double negation and the rule of contraposition for strict laws. With this assumption, the system also detects conflicting consequents of the defeasible laws, as in the following example:

$$\mathbf{K} \;=\; \{A \Rightarrow C, B \Rightarrow D, C \rightarrow \neg E, \neg\neg D \rightarrow E, Ai, Bi\} \qquad (4)$$

In the next two rules, set out in Table 1, '>' abbreviates strict or defeasible implication.[4]

In words, $(\Delta+)$ says that A is defeasibly derivable from $\langle \mathbf{F}, \mathbf{L}\rangle$ if either (a) A is already strictly derivable (which implies that $\vdash$ is a subrelation of $\mid\!\sim$), or else, if (b) A's negation is not strictly derivable (if otherwise, strict as well as defeasible rules with defeasibly derivable antecedents are defeated), and one of two cases obtains: (b1) either A is derivable from a strict law with defeasibly derivable antecedent, or (b2) A is derivable from a defeasible law with (b2.1) defeasibly derivable antecedent which (b2.2) is not defeated by another (strict or defeasible) law with defeasibly derivable antecedent and conflicting consequent. Note that if the other law is strict, it will be always defeating independent of specificity considerations, whence clause (b2.2.2) is restricted to defeasible conflicting laws. The dual $(D-)$ is obtained from $(D+)$ by the procedure already described below the monotonic rules $(M+/-)$.

It is in the definition (Confl) where our basic system differs from Nute's. He restricts conflicts to the case where one consequent is the negation of the other.[5] Nute's system derives from the modified Nixon example (3) Pacifist(nixon), because in his system the first defeasible law is not defeated by the second. This gives via the strict law $\neg ND(\text{nixon})$, which now defeats the second defeasible law. In example (4), Nute's system derives Ci as well as Di (and in the "strict" version of his system Ei and $\neg Ei$, and in the "semi-strict" version none of both). All this is contrary to our intuition.

It is easily demonstrably that (i) our basic system handles all the examples discussed so far successfully, (ii) if $n(A)$ is derivable, then A is not derivable. Let us finally summarize the general features of reasoning systems of the described kind. They are based on rules of the following form (X, Y and $Z(X,Y)$ are as in (Constr)):

[4] $(D+)$ corresponds to the conjunction of Nute's rules E^+ (§3), S^+ (§4) and $D^{\pm}_{\rightarrow}$ (§8), and $D-$ to the conjunction his rules E^- (§3) and $D^{-}_{\rightarrow}$ (§8), with the differences mentioned below.

[5] Nute's definition of 'contrary' in (1988: 269) corresponds to our definition of 'conflict'; I do not know why he has changed it in (1991).

$(D+)$ $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim A$ if
either (a) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash A$,
or (b) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash n(\neg A)$ and
either (b1) there exists $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$
such that $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim \Gamma j$
or (b2) there exists $\Gamma \Rightarrow A'$ and i with $A'i = A$
such that (b2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim \Gamma i$
and (b2.2) for each $\Delta > B' \in \mathbf{L}$ and j
such that $\text{Conflict}(B'j, A'i/\mathbf{L})$
either (b2.2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim n(\Delta j)$,
or (b2.2.2) $>\!\!=\!\!\Rightarrow$ and
$\text{Morespec}(\Gamma i, \Delta j/\mathbf{L})$.

$(D-)$ $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim n(A)$ if
(a) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash n(A)$ and
either (b) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash \neg A$
or (b1) for each $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$,
it holds that $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim n(\Gamma i)$ and
(b2) for each $\Gamma \Rightarrow A'$ and i with $A'i = A$ it holds that
either (b2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim n(\Gamma i)$
or (b2.2) there exists $\Delta > B' \in \mathbf{L}$ and j
such that $\text{Conflict}(B'j, A'i/\mathbf{L})$ and
(b2.2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mid\!\sim \Delta j$ *and*
(b2.2.2) either $>\!\!=\!\!\rightarrow$
or $\text{notMorespec}(\Gamma i, \Delta j/\mathbf{L})$.

Table 1: The Rules $D+$ and $D-$

$$\frac{X \subseteq \mathbf{K} \quad Y \nsubseteq \mathbf{K}}{Z(X,Y) \subseteq \mathbf{C}(\mathbf{K})}\ \textit{Start} \qquad \frac{X \subseteq \Gamma \quad Y \nsubseteq \mathbf{K}}{Z(X,Y) \subseteq \mathbf{C}(\Gamma)}\ \textit{Iteration} \qquad (5)$$

In both kinds of rules, the left or the right disjunct of the precondition may be empty, which gives four categories. If we decompose the rules with disjunctive preconditions of our basic system into several rules, then it is easy to see that they all fall in one of these four categories. The distinguished feature of these kind of rules, in contrast to the general definition of constructive rules in (Constr), is that the negative and nonmonotonic precondition applies only to the starting set $\mathbf{K}$, but not iteratively to any set Γ of formulas derived so far. This guarantees that the derivation process is not only rule-constructive but also *cumulative*. Hence, $\mathbf{C}(\mathbf{K})$ is uniquely definable as the smallest formula set which contains all formulas derivable from $\mathbf{K}$ by the start rules and is closed under the iterative rules. Let us call

(Confl+) $\quad \text{Conflict}(Ai, Bj/\mathbf{L}) \quad$ if $\langle Ai, \mathbf{L}\rangle \vdash \neg Bj$
(Confl−) $\quad \text{notConflict}(Ai, Bj/\mathbf{L}) \quad$ if $\langle Ai, \mathbf{L}\rangle \vdash n(\neg Bj)$

We assume that our derivability systems masters the rules of double negation and the rule of contraposition for strict laws. With this assumption, the system also detects conflicting consequents of the defeasible laws, as in the following example:

$$\mathbf{K} \;=\; \{A \Rightarrow C, B \Rightarrow D, C \rightarrow \neg E, \neg\neg D \rightarrow E, Ai, Bi\} \qquad (4)$$

In the next two rules, set out in Table 1, '>' abbreviates strict or defeasible implication.[4]

In words, $(\Delta+)$ says that A is defeasibly derivable from $\langle \mathbf{F}, \mathbf{L}\rangle$ if either (a) A is already strictly derivable (which implies that $\vdash$ is a subrelation of $\mathrel{|\!\sim}$), or else, if (b) A's negation is not strictly derivable (if otherwise, strict as well as defeasible rules with defeasibly derivable antecedents are defeated), and one of two cases obtains: (b1) either A is derivable from a strict law with defeasibly derivable antecedent, or (b2) A is derivable from a defeasible law with (b2.1) defeasibly derivable antecedent which (b2.2) is not defeated by another (strict or defeasible) law with defeasibly derivable antecedent and conflicting consequent. Note that if the other law is strict, it will be always defeating independent of specificity considerations, whence clause (b2.2.2) is restricted to defeasible conflicting laws. The dual $(D-)$ is obtained from $(D+)$ by the procedure already described below the monotonic rules $(M+/-)$.

It is in the definition (Confl) where our basic system differs from Nute's. He restricts conflicts to the case where one consequent is the negation of the other.[5] Nute's system derives from the modified Nixon example (3) Pacifist(nixon), because in his system the first defeasible law is not defeated by the second. This gives via the strict law $\neg ND$(nixon), which now defeats the second defeasible law. In example (4), Nute's system derives Ci as well as Di (and in the "strict" version of his system Ei and $\neg Ei$, and in the "semi-strict" version none of both). All this is contrary to our intuition.

It is easily demonstrably that (i) our basic system handles all the examples discussed so far successfully, (ii) if $n(A)$ is derivable, then A is not derivable. Let us finally summarize the general features of reasoning systems of the described kind. They are based on rules of the following form (X, Y and $Z(X,Y)$ are as in (Constr)):

[4] $(D+)$ corresponds to the conjunction of Nute's rules E^+ (§3), S^+ (§4) and $D^{\pm}_{\rightarrow}$ (§8), and $D-$ to the conjunction his rules E^- (§3) and $D^{-}_{\rightarrow}$ (§8), with the differences mentioned below.

[5] Nute's definition of 'contrary' in (1988: 269) corresponds to our definition of 'conflict'; I do not know why he has changed it in (1991).

$(D+)$ $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} A$ if
either (a) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash A$,
or (b) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash n(\neg A)$ and
either (b1) there exists $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$ such that $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} \Gamma j$
or (b2) there exists $\Gamma \Rightarrow A'$ and i with $A'i = A$ such that (b2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} \Gamma i$ and (b2.2) for each $\Delta > B' \in \mathbf{L}$ and j such that $\text{Conflict}(B'j, A'i/\mathbf{L})$ either (b2.2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} n(\Delta j)$, or (b2.2.2) $>\!\!=\!\!\Rightarrow$ and $\text{Morespec}(\Gamma i, \Delta j/\mathbf{L})$.

$(D-)$ $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} n(A)$ if
(a) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash n(A)$ and
either (b) $\langle \mathbf{F}, \mathbf{L}\rangle \vdash \neg A$
or (b1) for each $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$, it holds that $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} n(\Gamma i)$ and
(b2) for each $\Gamma \Rightarrow A'$ and i with $A'i = A$ it holds that either (b2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} n(\Gamma i)$
or (b2.2) there exists $\Delta > B' \in \mathbf{L}$ and j such that $\text{Conflict}(B'j, A'i/\mathbf{L})$ and (b2.2.1) $\langle \mathbf{F}, \mathbf{L}\rangle \mathrel{|\!\sim} \Delta j$ *and* (b2.2.2) either $>\!\!=\!\!\rightarrow$ or $\text{notMorespec}(\Gamma i, \Delta j/\mathbf{L})$.

Table 1: The Rules $D+$ and $D-$

$$\frac{X \subseteq \mathbf{K} \quad Y \nsubseteq \mathbf{K}}{Z(X,Y) \subseteq \mathbf{C}(\mathbf{K})}\ \textit{Start} \qquad \frac{X \subseteq \Gamma \quad Y \nsubseteq \mathbf{K}}{Z(X,Y) \subseteq \mathbf{C}(\Gamma)}\ \textit{Iteration} \qquad (5)$$

In both kinds of rules, the left or the right disjunct of the precondition may be empty, which gives four categories. If we decompose the rules with disjunctive preconditions of our basic system into several rules, then it is easy to see that they all fall in one of these four categories. The distinguished feature of these kind of rules, in contrast to the general definition of constructive rules in (Constr), is that the negative and nonmonotonic precondition applies only to the starting set $\mathbf{K}$, but not iteratively to any set Γ of formulas derived so far. This guarantees that the derivation process is not only rule-constructive but also *cumulative*. Hence, $\mathbf{C}(\mathbf{K})$ is uniquely definable as the smallest formula set which contains all formulas derivable from $\mathbf{K}$ by the start rules and is closed under the iterative rules. Let us call

a system axiomatisable by rules of the form (5) *grounded*. Clearly, our basic system is just one member of this large family. We do not know whether any rule-constructive and cumulative reasoning system must be grounded, but we conjecture it.

Two further important general axioms for $\mid\!\sim$ are left logical equivalence and right weakening (cf. Kraus *et al.* 1991, p. 177). Our basic system satisfies left logical equivalence but not right weakening; however its improvement in Section 4 will satisfy also right weakening. We expect from our system is that it also satisfies the axiom:

$$\text{(Negation)} \quad \mathbf{K} \mid\!\sim n(A) \quad \text{iff} \quad \text{not } \mathbf{K} \mid\!\sim A$$

Clearly, any constructive derivation system (based on decidable rules) which satisfies (Negation) must be decidable, because its theorems as well as its non-theorems will be recursively enumerable. *Vice versa*, if a nonmonotonic logic is undecidable, every rule-constructive axiomatization of it which satisfies (Negation) will be incomplete. This is not a drawback but exactly what one would expect.

3 Negation Completeness and the Problem of Circular Laws

It is easy to see that our basic system satisfies only the correctness half of the negation axiom above (from left to right), but not the completeness half (from right to left). Whenever $\mathbf{K}$ contains circular laws, as in the example

$$\mathbf{K} \;=\; \{A \Rightarrow B, B \Rightarrow A\},$$

the system will neither derive $n(A)$ nor $n(B)$ though neither A nor B is derivable. The reason is that any PROLOG program implementing the basic system will immediately run into an infinite loop: deriving $n(A)$ presupposes deriving $n(B)$ which presupposes deriving $n(A)\ldots$ This failure may lead to intuitively wrong results, as in:

$$\mathbf{K} \;=\; \{A \Rightarrow B, B \Rightarrow A, Ci, C \Rightarrow D, A \Rightarrow \neg D\}$$

Intuitively, Di should be derivable because the antecedent Ai of the conflicting rule $A \Rightarrow \neg D$ is not derivable, but since $n(Ai)$ is not derivable because of the loop, the basic system will not derive Di.

It will often happen that the laws of an expert system are circular, in particular if it contains causal as well as noncausal symptom laws. For instance, coplic spots indicate measles, which in turn causally imply the symptom coplic spots (cf. Schurz 1991). Solving the circularity problem complicates derivation rules, but whenever one intends a safe implementation, infinite loops have to be excluded anyway. Given that A is our proof goal, this can be done by keeping track of the branch $\langle A, B_1, \ldots, B_n, \rangle$ which leads from the current subgoal B_n to the goal A (the top of the proof tree

(Start) $\langle \mathbf{F}, \mathbf{L} \rangle \vdash X$ if $\langle \mathbf{F}, \mathbf{L} \rangle \vdash [X, \langle X \rangle]$, where X is of shape A or $n(A)$.

(Circ) $\langle \mathbf{F}, \mathbf{L} \rangle \vdash [n(A), S + n(A)]$ if $S + n(A)$ is circular.

(M^*+) $\langle \mathbf{F}, \mathbf{R} \rangle \vdash [A, S]$ if either $A \in \mathbf{F}$ or there exist $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A = A'(i)$ such that for each $B \in \Gamma i$, $S + B$ is n.c. and $\langle \mathbf{F}, \mathbf{L} \rangle \vdash [B, \Sigma + B]$.

(D^*+) $\langle \mathbf{F}, \mathbf{L} \rangle \mid\!\sim [A, S]$ if either (a) $\langle \mathbf{F}, \mathbf{L} \rangle \vdash [A, S]$
or (b) $\langle \mathbf{F}, \mathbf{L} \rangle \vdash [n(\neg A), S + n(\neg A)]$ and
either (b1) there exist $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$ such that for each $B \in \Gamma i$, $S + B$ is n.c. and $\langle \mathbf{F}, \mathbf{L} \rangle \mid\!\sim [B, S + B]$
or (b2) there exist $\Gamma \Rightarrow A'$ and i with $A'i = A$ such that
(b2.1) for each$B \in \Gamma i$, $S + B$ is n.c. and $\langle \mathbf{F}, \mathbf{L} \rangle \mid\!\sim [B, S + B]$, and
(b2.2) for each $\Delta > B' \in \mathbf{L}$ and j such that Conflict$(B'j, A'i/\mathbf{L})$,
either (b2.2.1) for some $C \in \Delta j$, $\langle \mathbf{F}, \mathbf{L} \rangle \mid\!\sim [n(C), S + n(C)]$,
or (b2.2.2) $>\!\!\Rightarrow$ and Morespec$(\Gamma i, \Delta j/\mathbf{L})$.

$\mathbf{K} \mid\!\sim [A, S]$ means that A is derivable from $\mathbf{K}$ as the last member of the current branch S. Formally, S is a formula sequence and $S + B$ denotes the result of attaching B to S on the right.

Table 2: The Rules M^*+ and D^*+

of A). For each new law instantiation C we check whether the extended branch $\langle A, B_1, \ldots, B_n, C \rangle$ is still non-circular, i.e. whether it does not contain the same closed literal twice. As soon as a branch becomes circular, derivability with *this* branch fails. This is reflected in new derivation rules, which are set out in Table 2. The start rule leads derivability *simpliciter* back to derivability with the top goal as the current branch. The circularity rule returns non-derivability as soon as the current branch becomes circular. The modified rules (M^*+) and (D^*+) are analogous to those before but keep track of the current branch and check whether this is still n.c. (for "non-circular"). We omit the negative duals; they can be obtained by the procedure explained in the previous section. It can be proved that this modified system is negation complete.

4 Collective Defeat and the Axiom of Right Weakening

Another problem not adequately handled by the basic system is that of *collective defeat* (so called after Pollock 1987, p. 493ff). Consider the following example:

$$\mathbf{K} = \{A \Rightarrow B, A \Rightarrow C, \{B,C\} \rightarrow D, E \Rightarrow \neg D, Ai, Ei\} \qquad (6)$$

For illustration, assume A to be uncertain evidence for being married (B) and being a priest (C), D stands for being a protestant and E for being an Italian. Here, not two but *three* defeasible laws are in conflict: the conjunction of the consequents of the first two defeasible laws ($\{Bi, Ci\}$) strictly implies the negation of the consequent of the third one ($\neg Di$). Intuitively, neither $\{Bi, Ci\}$ nor $\neg Di$ should be derivable, and since we have no reason to prefer Bi over Ci, neither Bi nor Ci should be derivable. Thus, if several laws (of same priority) collectively defeat each other, all of them are defeated – which is the solution suggested by Pollock (1987, p. 493). Our basic system, however, derives from (11) Bi, Ci and from that Di, which is inadequate.

A slight variation of (6), where the third law is not defeasible but strict is

$$\mathbf{K} = \{A \Rightarrow B, A \Rightarrow C, \{B,C\} \rightarrow D, E \rightarrow \neg D, Ai, Ei\}. \qquad (7)$$

Here our basic system derives Bi, Ci, but then $\neg Di$, which is now strictly derivable and blocks $\{B,C\} \rightarrow D$ by the first conjunct of clause (D+,b). This is again intuitively inadequate, but beyond that, it violates also the axiom of right weakening

$$\frac{\mathbf{K} \mathrel{|\!\sim} \Gamma \quad \Gamma \vdash \Delta}{\mathbf{K} \mathrel{|\!\sim} \Delta} \textit{ Right Weakening}$$

which is seen by putting $\Gamma = \{\ \{B,C\} \rightarrow D, Bi, Ci\}$ and $\Delta = \{Di\}$.

A well known example of collective defeat of the sort (6) is the *lottery paradox*, where $\mathbf{K}$ contains n defeasible laws of the form 'the kth member of the lottery L will not win', instead of the two laws $A \Rightarrow B$ and $A \Rightarrow C$ in (7). Poole (1991, p. 291) has argued that the better way to solve this paradox would be to refrain from the rule of introducing conjunctions (which trivially holds in our system since sets are treated as conjunctions) rather than to block all the collectively defeating laws. But one plausible requirement is that the conjunction of all our derived beliefs must be classically consistent. We adopt this requirement here and thus stick to Pollock's solution.

To simplify things, our new rule (D^{**}+), set out in Table 3, is formulated without keeping track of branches (as we did for the rules in Table 2).

(Confl*+)	Conflict($A1i_1, \ldots, A_ni_n/\mathbf{L}$) if, for some $k \leq n$, $\langle\{A_ki_k \mid i \neq k, i \leq n\}, \mathbf{L}\rangle \vdash \neg A_ki$
(D^{**}+)	$\langle\mathbf{F},\mathbf{L}\rangle \mathrel{|\!\sim} A$ if either (a) $\langle\mathbf{F},\mathbf{L}\rangle \vdash A$ or (b1) there exist $\Gamma \rightarrow A' \in \mathbf{L}$ and i with $A'i = A$ such that $\langle\mathbf{F},\mathbf{L}\rangle \mathrel{|\!\sim} Gj$, or (b2) there exists $\Gamma \Rightarrow A'$ and i with $A'i = A$ such that (b2.1) $\langle\mathbf{F},\mathbf{L}\rangle \mathrel{|\!\sim} \Gamma i$ and (b2.2) for each *set* of quasi-rules $\{\Delta_r \gg_r B_r \mid r \leq n\} \subseteq \mathbf{L}$ and $j_1, \ldots, j_r$ such that Conflict($B_1j_1, \ldots, B_nj_n, A/\mathbf{L}$), either (b2.2.1) for some $m \leq n$, $\langle\mathbf{F},\mathbf{L}\rangle \mathrel{|\!\sim} n(\Delta_m)$, or (b2.2.2) there exists $\Sigma \in \Pi :=$ $\{\Delta_1j_n, \ldots, \Delta_nj_n\}$ such that Σ is the instantiated antecedent of a defeasible rule, and $(P \cup \Gamma_i) - \Sigma$ is more specific than Σ.

Table 3: The Rules Confl*+ and D^{**}+

It is based on a new conflict rule telling when a set of instantiated consequents is in "collective" conflict: exactly if it is classically inconsistent. By a *quasi-rule* $\Delta \gg A$ we mean either a fact A (in which case Δ is empty) or a strict or defeasible rule $\Delta > A$. Clause (D^*+, b2.2.2) takes into account that the collective defeat is undermined if the union of some of the rule-antecedents (in the et of collectively conflicting rules) is more specific than the antecedent of another defeasible rule different from $\Gamma \Rightarrow A$ (we omit examples because of the limitations of space). The duals of (Confl*+) and (D^{**}+) are obtainable in the way described in Section 2. Note that the new rules contain the previously discussed situations where only two laws are in conflict as a special case.

We do not longer need the clause '$\langle\mathbf{F},\mathbf{L}\rangle \vdash n(\neg A)$' of ($D$+), because of the characterization in terms of quasi-rules. For, if$A \Rightarrow B \in \mathbf{L}$ and $\neg Bi$ is strictly derivable, then there will always be a set of of facts or strict rules in $\mathbf{K}$ satisfying the Conflict-clause, which will block the firing of rule $A \Rightarrow B$. Therefore, strict rules which defeasible derivable antecedents will always fire in our new system, which implies that the axiom of right weakening is satisfied. Moreover, it can be proved for the new system that if $\mathbf{K}$ is strictly consistent, also $\mathbf{C}(\mathbf{K})$ will be strictly consistent.

References

Brewka, G. 1991 *Nonmonotonic Reasoning. Logical Foundations of Commonsense*, Cambridge: Cambridge University Press.

Delgrande, J. P. 1988 "An Approach to Default Reasoning Based on a First-Order Conditional Logic: Revised Report", *Artificial Intelligence* 36, 63–90.

Etherington, D. W. 1987 "Formalizing Nonmonotonic Reasoning Systems", *Artificial Intelligence* 31, 41–85.

Gabbay, D. M. 1985 "Theoretical Foundations for Non-Monotonic Reasoning in Expert Systems", in K. R. Apt, ed., *Proceedings of the NATO Advanced Study Institute on Logics and Models of Concurrent Systems*, Berlin: Springer, pp. 439–57.

Hempel, C. G. 1965 *Aspects of Scientific Explanation*, New York: Free Press.

Konolige, K. 1988 "On the Relation between Default and Autoepistemic Logic", *Artificial Intelligence* 35, 343–382.

Kraus, S., Lehmann, D. and Magodor, M. 1990 "Nonmonotonic Reasoning, Preferential Models and Cumulative Logics", *Artificial Intelligence* 44, 167–207.

McDermott D. and Doyle, J. 1980 "Non-Monotonic Logic I", *Artificial Intelligence* 13, 41–72.

Makinson, D. 1989 "General Theory of Cumulative Inference", in *Proceedings of the 2nd International Workshop on Non-Monotonic Reasoning*, Berlin: Springer.

Moore, R. C. 1985 "Semantic Considerations on Nonmonotonic Logic", *Artificial Intelligence* 25, 75–94.

Nute, D. 1988 "Defeasible Reasoning: A Philosophical Analysis in Prolog", in J. Fetzer, ed., *Aspects of Artificial Intelligence*, Dordrecht: Kluwer.

Nute, D. 1991 "Basic Defeasible Logic", in Farinas-del-Carro and Pentonnen, eds., *Intentional Logics for Programming*, Oxford: Oxford University Press.

Pollock, J. L. 1987 "Defeasible Reasoning", *Cognitive Science* 11, 481–518.

Poole, D. 1988 "A Logical Framework for Default Reasoning", *Artificial Intelligence* 36, 27–47.

Poole, D. 1991 "The Effect of Knowledge on Belief: Conditioning, Specificity and the Lottery Paradox in Default Reasoning", *Artificial Intelligence* 49, 281–307.

Reiter, R. 1980 "A Logic for Default Reasoning", *Artificial Intelligence* 13, 1–132.

Rescher, N. 1964 *Hypothetical Reasoning*, Amsterdam: North-Holland.

Scriven, M. 1959 "Truisms as Grounds for Historical Explanations", in P. Gardiner, ed., *Theories of History*, New York: Free Press, 443–468.

Schurz, G. 1988 (ed.) *Erklären und Verstehen in der Wissenschaft*, Oldenbourg, p. 235-298.

Schurz, G. 1991 "Erklärungsmodelle in der Wissenschaftstheorie und in der Künstlichen Intelligenz", in H. Stoyan, ed., *Erklärung im Gespräch – Erklärung im Mensch-Maschine-Dialog*, Informatik Fachberichte, Berlin: Springer, 1–42.

Schurz, G. 1993 "Nonmonotonic Reasoning and Changes of Belief", to appear in P. Weingartner, ed., *Scientific Belief and Religious Belief*, Special Issue of *Philosophical Studies*.

Semantic Compositionality
The Argument from Synonymy

Francis Jeffry Pelletier

The Principle of Semantic Compositionality is the principle that the meaning of an expression is a function of, and only of, the meanings of its parts together with the method by which those parts are combined. As stated, the Principle is vague or under specified at a number of points such as "what counts as a part", "what is a meaning", "what kind of function is allowed" and the like. But this hasn't stopped some people from treating it as an obviously true principle, true almost by definition, nor has it stopped some others from attacking it both on "empirical grounds" and on theoretico-methodological grounds. It seems to me that many of these discussions fail because of a lack of precision on the above mentioned points and that other discussions are best described as "how compositionality can/cannot be accommodated within theory X" rather than whether the Principle is or is not true.

That is, the majority of the arguments against the Principle rely on having some specific grammatical framework in which to work and consist of arguing that some phenomenon cannot be compositionally treated within this theory. Some arguments rely on assumptions whose truth seems even more questionable than that of the Principle's, as for instance an argument which asserts that mental states are (or are not) compositional in nature and that since language is the mirror of our mental states therefore the Principle is (or is not) true. There are not, in fact, very many arguments in the literature in *favour* of The Principle, probably because almost all theorists – especially philosophically-oriented theorists – have a warm and fuzzy feeling about the Principle. They cannot even imagine what it might mean for it to be false and so they do not bother arguing in its favor but rather merely shake their heads sadly and knowingly at anyone whose theory supposes that the Principle is false.

So, although there may be various theory-internal grounds for adopting or denying The Principle, I claim that there are only four theory-independent arguments concerning it. There are two in its favor and two opposed. They are:

In favour	1	The argument from learnability
	2	The argument from understandability
Opposed	1	The argument from synonymy
	2	The argument from ambiguity.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 311-317

The argument from learnability is that if a language lacked compositionality it would be un*learn*able. If the meaning of the whole were not a function of the meaning of its parts, this argument says, then we would not be able to *learn* the language. The only way we can learn an infinite (or even extremely large) set of sentences would be for us to learn a finite base and learn a finite number of ways of combining items. But if the Principle weren't true then this manner of learning would not help us in knowing what the constructed sentences meant, and therefore we wouldn't really have learned the language. The argument from learnability poses a choice between the Principle on the one hand, versus being able to learn only a small portion of language.

The argument from understandability is in a way the converse of the argument from learnability. Here it is argued that the Principle is the only explanation of how a finite mechanism (such as the human brain/mind) can *understand* an infinite set of sentences.[1] How else, this argument asks us, are we to be able to figure out the meaning of an arbitrary, new, novel sentence – if it isn't by the fact that we know some finite number of parts and finite number of ways of putting them together? How is it that we can understand a novel sentence, except by predicting its meaning from our understanding of the meaning of its parts and mode of combination? And, the argument concludes, this is all that the Principle says.

The argument from synonymy is the argument that if the Principle were true generally then there could be no synonymy, whether of sentences, of phrases, or of lexical items. And since the existence of these phenomena is so clear, the argument concludes that the Principle is incorrect.

The argument from ambiguity alleges that it is impossible for there to be cases of identical surface structure but distinct meanings, if the Principle is true. There could be no such example as *Every linguist knows two languages* being simultaneously ambiguous between the two obvious meanings while at the same time that sentence having only one surface structure. (There being only one structure, the Principle would claim that this maps onto only one meaning). The argument affirms that there is certainly this type of ambiguity, where there is one surface structure and more than one meaning. Therefore the Principle is wrong.

In this paper I wish to concentrate on just one of these arguments, the argument from synonymy. All four of the arguments are interesting and I hope to investigate each one at some time or other,[2] but for now I would like to concentrate attention on this argument.[3]

[1] These arguments need not be stated in terms of an infinite set of sentences. As Grandy has pointed out, they seem to work even if we just presume a hugely large but finite set of sentences.

[2] In "The Principle of Semantic Compositionality", I make superficial remarks about all four. Although the majority of my attention there is spent looking at what I think of as the unconvincing arguments, I do spend some time discussing especially the argument from ambiguity and in making some remarks that might provide a different light on the arguments from learnability and understandability.

The argument from synonymy is *against* The Principle. Thus it needs to prove that there is no *function* from meanings of parts and modes of combination to meanings of wholes that will preserve synonymy. This means I must display some parts-and-modes meanings which are non-functional, that is, ones which do not yield at most one meaning-whole. This is the only way to show there to be no such function as required by the Principle – find an example where one and the same group of parts-and-modes yields two or more meaning-wholes. But how shall we convince ourselves that we have more than one meaning? Here is a sufficient condition:

[A_1] If Φ and Ψ have the same meaning, then they must have the same truth value.

I do not wish to enter into discussion of whether meaning is to be identified with truth conditions or truth-in-all-possible-worlds, or whether a theory of truth is *eo ipso* a theory of meaning. I only wish to insist on the very weak assumption [A_1], giving a minimal relation between meaning and truth.[4] [A_1] is called "the most certain principle" in Bäuerle & Cresswell 1988.

There are many theories of meaning – too many, some would say. Some theorists would identify meaning with a set of possible worlds, others with certain intentions of speakers, still others with a function from possible worlds and contexts to truth values, yet others with a speech act potential, and there are even some who think it is an expression of some other language. Not wishing to choose amongst such theories, I would ask only that they accept that there are some sentences – at least two – for which assumption [A_2] holds, regardless of the theory of meaning and the theory of syntax involved.

[A_2] There are syntactically distinct sentences Φ and Ψ which mean the same.

There is, in fact, at least one theory of meaning according to which this apparently innocuous assumption is false: the theory of "complete structured

[3] The argument will be recognized as being implicit in numerous discussions in semantics and in philosophy of language; but it has not, I believe, previously been explicitly deployed against The Principle. See for example Church's "translation argument" in his 1950, Bigelow's 1978, and the works of Cresswell cited in the bibliography.

[4] In fact there is a theory of meaning according to which [A_1] is false: distinguish between meaning and what is expressed in a context. According to such theories, although a sentence has a meaning in isolation, it only expresses a proposition in a context, and is therefore only true or false in a context. Thus two sentences might have the same meaning and yet not have any truth value at all... except in a context. (For example, indexical sentences like 'We are there' and 'I am here'). To such theorists, I ask that they interpret [A_1] in such a way that Φ and Ψ are in a context, and therefore able to have truth values. With such a modified understanding, I think these theorists should accept [A_1].

meanings" which has sometimes been associated with Lewis 1971 and Cresswell 1988. But even these authors believe this "complete structure" theory to be too strong. The identification of meaning with syntactic structure plus associated functions on possible worlds gives, as they say, a much too fine-grained approach to meaning. Cresswell's response is to give a theory in which the amount of syntactic structure that is part of the meaning *varies* from occasion to occasion in which it is used. One might wonder whether this, all by itself, isn't a non-compositional theory. (Cresswell says it *is* compositional.) But such discussions are beyond the scope of this paper. I would prefer here to just put it as a challenge: I believe that no theory which has the consequence of denying our [A_2] can stand much of a chance of capturing our pre-theoretic intuitions about meaning and sameness of meaning.

To find counterexamples to the Principle, we need to know when complex expressions are put together by the same method of combination. Different theories will naturally have differing ways of describing "combinations of parts". For the purposes of the present discussion we needn't choose amongst such theories; all we need to assume is that in any one theory the relevant sentences are generated or analyzed in the same way. So, the third assumption is:

[A_3] In each syntactic theory, there is only one syntactic rule (or only one sequence of syntactic rules) which creates or analyzes sentences of the form:

$$Kim + believes + that + \text{Sentence}$$

from the four component parts.

As I said before, I don't wish to take a stand on what the exact syntactic structure of such sentences is, since different theories might assign different structures and I intend my argument to hold against them all. Nor do I want to rule out the possibility that a given syntactic theory might in fact allow there to be more than one rule or rule-sequence that generates that kind of sentence. I state it this way merely for convenience of exposition. The crucial point is that I want to have to consider only one syntactic rule (or rule sequence) which creates all these types of sentences (in any one syntactic theory), so that, given two applications of this rule, if the corresponding parts mean the same then the whole will mean the same (assuming the Principle to be true). If the syntactic theory allowed more than one analysis of such sentences we could still generate the argument below, but it would introduce needless complexity into having to keep track of when a sentence is analyzed by which of the sequences of rules. Finally here I should remark that although [A_3] is stated using 'believes' (or 'believes that'), we could have used any of a number of so-called opacity-creating verbs, for example 'sincerely claims' (or 'sincerely claims that'). This point will re-emerge in the discussion of [A_4].

Finally I claim that for any pair of syntactically distinct sentences it is always possible that a person not believe they mean the same – even if the sentences really do mean the same.[5]

[A_4] If Φ and Ψ are syntactically distinct sentences, then it is possible that exactly one of (i) and (ii) is true:

(i) *Kim believes that* Φ

(ii) *Kim believes that* Ψ

Of course, there are certain philosophical theories according to which [A_4] cannot be true: if a person believes that Φ, then he must also believe that Ψ, if Φ and Ψ mean the same. One such theory would claim that *believes* is always "transparent". Another related theory would say that the evaluation of Φ picks out a certain class of possible worlds – by hypothesis, the same class that Ψ picks out. So if Kim believes Φ, Kim must also believe Ψ: after all, Φ and Ψ are the same! (Same class of possible worlds, that is.) Bäuerle & Cresswell ("Propositional Attitudes", pp. 493ff.) consider such theories, and charitably attribute to them the view that "for purposes of a viable semantics we must treat [such sentences] as if they had the same truth value". This is not the place to debate such theories, and instead I will state my case by remarking that one *could* evade the thrust of my anti-compositional argument by denying [A_4]. But notice that such counter theories have to say (implausibly) that the person "*really* believes" Φ despite his protestations to the contrary. And had I replaced *believes that* in both [A_3] and [A_4] by *sincerely claims that*, then these counter-theories would be in the unpalatable position of having to claim that the person "*really* sincerely claims that Φ", despite his (sincere!) protestations to the contrary.

Now for the (brief) argument. By [A_2] there are two syntactically distinct sentences – call them S_1 and S_2 – which mean the same thing. By [A_3] it is the same rule (or sequence of rules) which, from these identically-meaning S_1 and S_2, forms or analyzes both of

(1) Kim believes that S_1

(2) Kim believes that S_2

So according to the Principle, (1) and (2) must mean the same thing. But by [A_4] it is always possible that one of (1) and (2) is true and the other is false. And so it follows then by [A_1], that sentences (1) and (2) have different meanings. So, [A_1]–[A_4] are inconsistent with the Principle.

A variant of this argument can be wielded against anyone who admits that there are pairs of distinct phrases (as opposed to sentences) which have

[5] [A_4] needn't hold for all sentences – it need only assert that, for each syntactic theory there be some pair or other for which it is true, and that this pair satisfy [A_2] and [A_3]. (Arguably, Mary kissed John and John was kissed by Mary mean the same and that no one believes one without believing the other. If this be so, then choose some other pair of sentences which will satisfy the postulates.)

the same meaning. Consider, for example, the possibility that *a locus of all points on a plane equidistant from a given point* means the same as *a circle*. By analogy with [A_3], we would further assume that there is only one sequence of syntactic rules which forms or analyzes both of (3) and (4) from their parts.

(3) A circle is a circle

(4) A circle is a locus of all points in a plane equidistant from a given point

from its parts. Since (3) and (4) were formed by the same rule (sequence) from parts that by hypothesis have the same meaning, The Principle says that (3) and (4) mean the same. But (5) and (6) are, according to principle [A_3], formed or analyzed from (3) and (4) by using the same rule (sequence).

(5) Kim believes that a circle is a circle

(6) Kim believes that a circle is a locus of all points on a plane equidistant from a given point

Since the same rule is applied to items which mean the same, The Principle says that (5) and (6) mean the same. But [A_4] says that one of them can be true and the other false; and by [A_1] it then follows that (5) and (6) do *not* mean the same. So here the Principle is shown to be incompatible with synonymy of phrases.

The argument can be further extended to show the incompatibility of The Principle with word synonymy. This extension is of some additional interest to those theorists attracted to "structured meanings". It is tempting to adopt some variant of this "structured" theory of meaning, both for the reasons outlined in Lewis 1972 and Cresswell 1980 and for the fact that the theory of structured meanings appears to evade the preceding argument by denying identity of meaning of the sentences mentioned in [A_2] and of the phrases mentioned in (3) and (4). I think it will appear to be much less tempting when it is pointed out that the theory must also deny the possibility of lexical synonymy. For, if there were two words which meant the same, for example maybe *attorney* and *lawyer*, then we could mimic the preceding argument. By analogy with [A_3], we assume that there is just one sequence of syntactic rules which forms or analyzes both of

(7) All lawyers are scoundrels

(8) All attorneys are scoundrels

By hypothesis this one sequence is operating on parts (ultimately, meanings of words) which are identical. Therefore, according to the Principle, (7) and (8) mean the same thing. But once again [A_3] analyzes both (9) and (10) by the same rule sequence:

(9) Kim believes that all lawyers are scoundrels

(10) Kim believes that all attorneys are scoundrels

Again by hypothesis, (9) and (10) have parts that mean the same, namely, (7) and (8), and therefore (9) and (10) must mean the same. Yet according to [A_4] one of them can be true and the other false; and by [A_1] it follows that (9) and (10) must mean different things – thereby generating the contradiction once again.

Thus the Principle is incompatible with [A_1] – [A_4], and also with the extended notions of phrasal and lexical synonymy. Which should be rejected? I think there is no question that [A_1] and [A_3] must be maintained. So the choice is amongst: (a) the Principle, (b) that different strings of symbols can mean the same ([A_2]), and (c) that a person might believe something yet reject a synonymous claim ([A_4]). Of these three, surely the Principle is the most suspect. Who would ever wish to deny that there is *any* synonymy – neither word nor phrasal nor sentence synonymy (i.e., deny [A_2])? And who really thinks that the manifest facts concerning belief should be overthrown by something so theoretically-motivated as the Principle?[6]

References

Bäuerle, R. & M. Cresswell 1988 "Propositional Attitudes", in D. Gabbay & F. Guenthner *Handbook of Philosophical logic* 4, Dordrecht: Kluwer, pp. 491–512.

Bigelow, J. 1978 "Believing in Semantics" *Journal of Philosophical Logic* 9, 101–144.

Church, A. 1950 "On Carnap's Analysis of Statements of Assertion and Belief", *Analysis* 10, 97–99.

Cresswell, M. 1980 "Quotational Theories of Propositional Attitudes", *Journal of Philosophical Logic* 9, 17–40.

Cresswell, M. 1985 *Structured Meanings*, Cambridge, MA: MIT Press.

Lewis, D. 1972 "General Semantics", in D. Davidson & G. Harman (eds.), *Semantics of Natural Language*, Dordrecht: Reidel, pp. 169–218.

Pelletier, F.J. 1994 "The Principle of Semantic Compositionality", *Topoi* 13, 11–24.

[6]Thanks go to David Braun, Sandro Zucchi, and Manfred Krifka for discussions and comments. I'm afraid that Sandro and Manfred believe that the Principle is correct and that [A4] is false... they call it "biting the bullet", I'd call it something else. David disbelieves The Principle, but also disbelieves [A_1] and [A_2]. I'd call this a case of overkill.

Stage Setting in Intentional Discourse

Andrew Woodfield

1 The Context and the Co-text of an Utterance

The stage-setting in which a speaker utters a propositional attitude sentence can profoundly affect the way that the audience interprets the utterance. In many cases, the stage-setting helps the audience to determine which mental content the speaker was ascribing.

In section 257 of the *Philosophical Investigations*, Wittgenstein says "a great deal of stage-setting in the language is presupposed if the mere act of naming is to make sense". He was talking about a special type of speech act, the act of giving an invented name to a sensation. In this paper, the things that will be said to have "stage-settings" are *particular* linguistic acts performed at particular times and places, and *particular* utterances, i.e. the meaningful products of those acts. I shall not make much of the act/product distinction, but I do emphasise the *locatedness*. The stage-setting of a past utterance is the set of circumstances which surrounded its production, including the utterances (if any) which preceded it. Setting the stage for a future utterance consists in creating the circumstances that will surround it when it occurs.

In the *Investigations*, Wittgenstein also claimed that mental states themselves depend upon background conditions. Certain thoughts are unavailable to dogs, because dogs lack the requisite "form of life", which is a background condition. Since we shall be considering the discourse produced by a reporter R about the mind of a subject S, we need to distinguish very clearly between two sorts of backgrounds belonging to two different things: one background comprises the *circumstances of the utterance* made by *R*, the other is the *situation* that *S* was in at the time of his reported mental state. The term 'stage-setting' covers only the former.

When Wittgenstein spoke of stage-setting *in the language*, he had in mind constant or semi-permanent conditions, such as the existence of a grammar, a lexicon, etc. Call such general conditions the *deep* stage-setting. Every particular utterance takes place also within a *local* stage-setting. The local stage-setting of an utterance may be divided into two components. These are the *co-text* and the *context*. The co-text comprises the utterances which preceded the given utterance within the same conversation or discourse. The context is the physical environment of the speaker and the audience, plus their knowledge and attitudes.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 319-330

Contextual information plays an important part in communication. Many assumptions are left unspoken. If the speaker explicitly tells the hearer something about the context instead of leaving it to be contextually understood, the uttering of this information becomes part of the co-text for the utterances that come after. Occasionally, certain bits of contextual information must *not* be made explicit, because to do so would alter the content or force of a key utterance. For example, spelling out an assumption can sometimes spoil a joke.

Narratives often exemplify a simple pattern: the speaker utters several sentences describing a scene, then delivers a *punch-line*. The punch-line describes a key incident in the scene. There is often a structural similarity between the scene and the story. For example, on Monday a chronologically ordered sequence of events e_1, e_2,... e_m occurs as the run-up to the target incident e_n. On Tuesday, a story is told about this incident. The discourse contains an ordered sequence of utterances such that the first utterance describes e_1, the second utterance describes e_2, and so on, and the last utterance describes the incident e_n. Each utterance is set in the stage provided by the preceding utterances; thus the discursive stage-setting is cumulative. It will be assumed that the punch-line is always the last utterance. When the punch-line is a propositional attitude sentence of the form "S ψ-ed that z", R *imputes* an attitude to S and *ascribes a content* to S's attitude. Ascribing content is a speech-act performed by a *person*, who uses the resources of language as an aid. Ascribing content is not something done by the sentence itself. The semantic properties of the sentence constrain the speaker's message, but do not fully determine it. In particular, the literal meaning of the embedded sentence 'z' sets limits to the range of contents which R may be taken to have ascribed. However, the stage-setting imposes constraints too.

The stage-setting often affects what R *said*, when he uttered the punch-line. Furthermore, the stage-setting may provide clues as to what R *meant* when he uttered the punch-line. This becomes important in cases where R meant something more than, or different from, what he said. In this paper I look at three kinds of phenomena: reference assignment, sense assignment, and metaphor. But there are other phenomena (e.g. ellipsis) which substantiate the general thesis that content-ascriptions are sensitive both to context and to co-text; space-limitations prevent me from discussing any of these others.

2 Reference-Assignment

Consider sentence (1), taken from a paper by John Haugeland 1979:

(1) I left my raincoat in the bathtub, because it was still wet.

The anaphoric pronoun 'it' harks back to an antecedent reference-marker. Within (1) there are two possible antecedents: 'my raincoat' and 'the bathtub'. But 'my raincoat' makes more sense.

Haugeland's purpose was to illustrate a fact about human communication, namely that we infer anaphoric connections with the help of common sense knowledge about life situations. Knowledge of language is not enough, because either of the two terms could be the antecedent of 'it' as far as syntax and semantics are concerned. Haugeland cites this as one of the many obstacles that face AI workers in their attempts to build a language-understanding machine.

What if sentence (1) had occurred within a larger text? Suppose it was uttered as the last sentence of a discourse (D1) where the preceding sentences were (0) and (0*).

(D1) (0) I went for a walk in the rain, and my raincoat got wet.
(0*) When I got home, the rain had stopped.
(1) I left my raincoat in the bathtub, because it was still wet.

Haugeland's main point continues to hold. To understand this three-sentence text, you need knowledge of the world and of human practices. Yet given the co-text, a computer could be programmed to select the antecedent 'my raincoat' as the reference-marker for 'it' in response to *textual* clues. The fact that the term 'my raincoat' is tokened in the co-text of (1) gives it some salience over the term 'the bathtub', which occurs only in (1). A simple heuristic such as 'Pick the term which has occurred twice' would not always get the right answer, but it would work in (D1). A more complex heuristic might exploit the fact that 'still wet' entails 'was wet'. Sentence (0) explicitly links the predicate 'wet' to the term 'my raincoat', and refers to an earlier time. So a condition for the antecedent's being 'my raincoat' is manifestly satisfied.

Leaving aside issues about machine-"understanding", it is surely true that the co-text of (1) helps a human interpreter to work out what 'it' refers to. The reason is that the preceding sentences explicitly describe the scene of a familiar type of incident. The hearer who has been told about the scene does not need to construct so much, or work so hard, as a hearer who confronts (1) on its own. His prior knowledge of a type of life situation has already been activated by the co-text; (1) gives him not just information, but an increment of information, about a particular situation of that type. The hearer who is exposed to the co-text finds one reference-assignment more available than the other, because he has been primed about the situation. He can guess rapidly and accurately what the target-incident was. His hypothesis about the narrated incident exerts a top-down influence in favour of assigning one anaphoric structure to (1).

Discourse Representation Theory has studied these phenomena in detail (see Heim 1983, Kamp 1984/5, Asher 1986, Spencer-Smith 1987). Sometimes the reference of an anaphoric expression in a discourse is uniquely determined by co-text. What this example shows is that co-text may contain clues which enable the hearer to assign a referent to 'it' more *confidently* and *quickly* than he would have been able to do, had there been no co-text.

The same holds when anaphoric expressions occur in the 'that' clauses of propositional attitude sentences. Suppose that R utters (0), (0*) and then (2):

(2) I left my raincoat in the bathtub, because I thought it was still wet.

The hearer can use the information gleaned from (0) and (0*) to strengthen his estimate that R is referring to the raincoat. A person in that situation is likely to have thought that the raincoat was still wet. Therefore it is most plausible that R, qua speaker, is now *reporting* that S (who happens to be himself) thought that the raincoat was still wet. The discourse-setting helps the hearer to identify with confidence which content-ascription the speaker is making.

Every textbook recognises that what is said can depend on co-text and context. When the utterance is a report in *oratio obliqua*, part of it consists in the uttering of sub-sentence '*z*'. The fact that '*z*' is embedded in a 'that' clause does not seal it off from the stage-setting. The speaker's choice of words in '*z*', and the intonation and stress with which he articulates '*z*', are influenced by his knowledge of context and co-text.

When '*z*' contains an anaphoric expression, the stage-setting not only helps to fix the reference, it may also carry a message about the way in which S conceives the referent. Let me display an example of this.

On Thursday July 22nd, an extraordinary event occurred in the British Parliament during the debate about the ratification of the Maastricht Treaty. On a Government motion the result was 317 votes in favour, 317 against. To the Prime Minister's disgust, many members of the Conservative party voted against the Government. In fact, the anti-Maastricht rebels in the Conservative party would have been sufficient in number to defeat the Government, had they all voted against. But three of them were Ministers in the Government, and hence they were obliged to vote in favour. The doctrine of collective responsibility by the Cabinet was the only thing that saved John Major on this occasion. Later, in an off-the-record conversation, Major referred to certain people as "those bastards". His unguarded remark provoked some controversy in the Press.

Now that I have described the situation, you are in a position to understand me when I give you the punch-line:

(3) John Major believed that it was good to have those bastards in the Cabinet.

To whom am I referring? Not to all the anti-Maastricht faction, nor to all the Conservatives in that faction who voted against. Neither class satisfies the condition of being in the Cabinet. Since Major presumably knows who is in his own Cabinet, it would be absurd to ascribe to him a belief, concerning many people who were not in his Cabinet, that it was good to have *them* in the Cabinet. When I uttered the anaphoric expression 'those bastards', I was referring to a certain set of men which my text had

previously introduced by means of the term 'three of them'. I was ascribing to Major a belief about the three anti-Maastricht *Ministers*.

The expression is not only anaphoric, it also contains the pejorative descriptive term 'bastard'. Yet it is not the case that the reference was *determined* descriptively in virtue of that term's descriptive meaning. The reference was determined by selecting a previous reference marker whose reference satisfied a different descriptive condition, namely, *being in the Cabinet*. So the function of the pejorative term is to supply an additional comment about certain men who have been independently identified. Who is the author of this unfavourable characterization? I was the producer of sentence (3); I characterized them as 'bastards'. But you understood that I was imputing this mode of characterisation to John Major. My preamble gave evidence that Major thought of his three colleagues in this way.

It is a difficult theoretical matter whether I, the speaker, strictly and literally *said* that John Major conceived of the three men as bastards. But given the co-text, it was clear that I *meant* that Major conceived of the three men as bastards. If content-ascribing is a speech act at the same level as speaker-meaning and communicating, then I ascribed a structured double content to Major's cognitive state. And you, the audience, understood that I was ascribing a complex content, because the previous discourse had primed you.

3 Sense-Assignment

Many word-types in English are lexically ambiguous. Usually, tokens of ambiguous words uttered in discourse are easy to disambiguate, because the surroundings often favour one sense assignment. As with reference assignment for anaphoric expressions, the preceding discourse sometimes makes one interpretation more plausible than others. The more clues there are which point to a particular reading, the more secure that reading becomes.

In (D2) there are conflicting clues as to the intended interpretation of the ambiguous word 'bank':

(D2) John walked into town to cash a cheque. As he crossed the bridge over the river, a sight confronted him. The bank had been fenced off.

Any competent narrator will be aware that the audience does not know at this stage whether he means *bank building* or *riverside*. So he will normally supply a piece of clarifying information.

(D2) contains no propositional attitude constructions. Let us look at a similar discourse (D3), whose last-sentence contains 'He thought that...'.

(D3) John walked into town to cash a cheque. As he crossed the bridge over the river, he thought that the bank had been fenced off.

The ambiguity carries over into *oratio obliqua* to produce an ambiguous report. *R* is presumably not ascribing an ambiguous content to the thought

itself. John would have been thinking about either the bank building or the river bank; his thought would not have been indeterminate between these two, nor would it have a thought about both. The punch-line itself is unclear, in so far as it fails to settle which thought R ascribed to John. And the preceding text offers conflicting clues.

If R had provided more information in the co-text, he could have tipped the balance in favour of one reading. It is interesting to note that there are two different routes which R could use. Both exploit certain inferences that the hearer might be expected to make, drawing upon prior knowledge. But the inferences proceed in different ways, depending on the sort of information that R provides. The hearer's inference could pass *from* an independently supported assumption about R's word meanings *to* a conclusion about the thought which R is imputing to S in the punch-line. Call this the first route. Or the inference could be *from* assumptions about S and S's situation, *to* the conclusion that it is most plausible that R is imputing such and such a thought to S. This is the second route.

The first route is followed if R provides some co-textual or contextual clue that he is going to mean, say, *river bank*. R could indicate this in an aside to the audience, explicitly telling them that he will be using the word 'bank' in that sense. The hearer is then primed to interpret every occurrence of 'bank' in this way, including those occurrences which occur in the complement clauses of propositional attitude sentences. The hearer presumes, unless there is reason not to, that the speaker means what he says. He then infers, upon hearing the punch-line, that the speaker means that John thought that the river bank was fenced off.

The second route is more complex. The hearer approaches the interpretation problem having already constructed a partial representation of the scene which R is describing. The hearer marshals a set of premises about John and John's situation in an environment, and possibly also about John's interests and attitudes. These assumptions may be based on what the narrator has said in the co-text, or they may be drawn from other sources. The hearer then infers that John would most likely be thinking about, say, the river bank, at the time of the incident. The hearer then moves on to consider the intentions of the narrator. He reasons thus: "I assume that R is in a position to know which thought John had. R presumably knows, then, that John had a thought about the river bank and not about the financial institution. R wants communicate this. Therefore, when R said 'bank' in the punch-line, R must have meant *river bank*'. This strategy is appropriate when there is no "story-independent" part of the stage-setting which indicates that R intends his own utterances of 'bank' to be read in the sense of *river bank*. This second route involves top-down inference, like the kind mentioned earlier in connection with anaphoric reference assignment.

I agree with Spencer-Smith (1987:18), when he claims that "the interpretation of an attitude report may require the attitude to be located within a larger network of the subject's attitudes". But it might be objected

that this claim involves a circularity. According to the claim, the interpreter follows a procedure in which one step must be carried out *before* another, because one step is *required* in order to carry out the other. The earlier step is to locate a certain attitude – the attitude which the speaker is imputing to S – within a larger network of S's attitudes. The subsequent step is to interpret the attitude report – that is, to identify which attitude the speaker is imputing to S. According to the objection, such a procedure is impossible because the first step logically requires that the attitude be already identified. The interpreter is supposed to put it in its proper location within a network *at the early stage*. He can hardly be expected to do this unless he knows which attitude it is. But if the interpreter knows which attitude it is, he has already achieved the final stage. For the goal of interpretation is precisely to identify *which* attitude the speaker imputed to S.

The correct response to this objection is: there is no circularity, because identifying an attitude is not necessarily the same thing as ascribing content to it. The interpreter's first step, of discovering which attitude R imputed to S, can be a matter of identifying it as the attitude that occupies a certain functional position in a network (e.g. as causally or inferentially related to certain other attitudes held by S). R may have described some of its relational properties in the co-text, and other such properties may be gleaned from the context. The second step, of interpreting R's ascription of content, can then be modulated by a presumption that the content ascribed by R is consonant with the attitude's functional role as described and/or presupposed by R. Route two is certainly possible, though the steps need to be carefully specified.

Incidentally, it is not necessary to present the interpreter as literally carrying out a sequence of steps one after another. This is merely a useful expository device. The real-time task of interpreting what a speaker means might be a process of multiple-constraint satisfaction.

Let us now consider the use of dialect words. The problem of sense-assignment arises when a single word form has two or more meanings. Some words have a standard meaning and a dialect meaning. A speaker or writer of standard English may switch into non-standard English within a discourse. Dialect words may be quoted, of course, but also R can use them without inverted commas and mean them in their dialect sense. Of course, the audience needs to know that he is doing this. There are two risks of communication failure. The audience might not recognise that the word is being used as a dialect word at all, and might assume that it bears its standard meaning. Or the audience might know that the word is being used as a dialect word, but not know what it means in the dialect. When successful, however, the use of words in non-standard senses can achieve a variety of cognitive effects. For example, if the audience is familiar with the dialect, the author's use evokes associated ideas and imagery, the flavour and accent of a region.

An author can convey extra information about a character's cognitive state, by using a dialect word in *oratio obliqua*. In 1987, the distinguished

novelist V.S. Naipaul published a semi-autobiographical novel entitled *The Enigma of Arrival.* It describes the period in Naipaul's life when he rented a cottage in rural Wiltshire. Naipaul's cottage was the property of a gentleman who owned a large house with an overgrown garden. A local man named Pitton was employed to look after this garden. Discourse (D4) is a passage in which Naipaul describes a pile of garden-waste which Pitton has collected.

(D4) This vegetable graveyard or rubbish-dump Pitton described as a "garden refuge", and a certain amount of ingenuity went into finding or creating these hidden but accessible "refuges". That was how Pitton used the word: I believe he had two or three such refuges at different places. Refuse, refuge: two separate, unrelated words. But "refuge", which Pitton used for "refuse", did in the most remarkable way contain both words. Pitton's "refuge" not only stood for "refuse", but had the additional idea or association, not at all inappropriate, of asylum, sanctuary, hiding, almost of hide-and-seek, of things kept decently out of sight and mind. He might say, of a fallen beech branch on the lawn, or a heap of grass clippings: "That'll be going to the refuge." Or: "I'll take it down to the refuge presently."

I thought at first that it was only Pitton's way with the word. But then I discovered that it was more or less common usage in the valley.[1]

In this passage the author does not use *oratio obliqua* to report Pitton's thoughts or sayings. But he could have done. Let me take the liberty of imagining a possible continuation of the narrative. Imagine a possible book in which an extra paragraph (D4*) is appended to (D4):

(D4*) One morning, I saw Pitton carrying the body of a dog into the garden. The dog was Pitton's. A passing car had run into the animal and killed it. Pitton deposited the corpse next to the refuge, and returned to fetch his spade. It did not surprise me that he intended to bury the body there. Pitton would have thought that the refuge was a suitable burial-ground.

In the last sentence, the imagined author speculates about what Pitton would have thought. Given that the co-text is (D4), the ascribed content is complex; the author manages to ascribe two thoughts for the price of one. It is clear that the token of the term 'the refuge' refers to the waste-heap; the author is *using* the word 'refuge" in the dialect sense. But speakers of Wiltshire dialect also know the standard sense, through listening to the radio and so on. It is possible that they do not regard the word as having two distinct and unrelated senses. They may see it as a having a family of related senses. Naipaul himself toys with the idea of connecting the standard sense with the dialect sense. Maybe the concept which the natives express by the word "refuge" is a fusion of two concepts.

[1] *The Enigma of Arrival*, p 182. Reproduced by permission of V.S. Naipaul ©.

Why would Pitton have thought that the refuge was a suitable burial-ground? First, he would have thought that the waste-heap was suitable from a practical point of view, as the body would decompose quickly and turn into compost. Secondly, it may be that he believes the place to be suitable from a spiritual point of view, because in his mind it presents itself as a refuge, a haven of rest. The second reason is perhaps a sentimental conjecture. But it is certainly an angle that the author might mean to communicate to the reader. An author who placed (D4*) against the stage-setting established by (D4) would almost certainly intend to ascribe a mental state flavoured by a special concept of *refuge.* Since any sensitive reader will pick up on that, the intention will be successful.

Here, then, is a third route by which R can exploit co-text so as to influence the interpretation of the punch-line. The first route was where R gave out information about his own use of language. The second route was where R let the hearer's assumptions about S and S's situation do some of the work. In the Naipaul-inspired example, R provides information in the co-text about the *language used by* S. This third route, like the second, requires the interpreter to take overall plausibility into account, but the inference proceeds in a significantly different way.

4 Metaphor in Content-Ascriptions

According to the pragmatic account which I favour, a metaphorical utterance is a communication in which the speaker means something different from what the sentence means. The interpreter's task is to *infer* a likely speaker-meaning (or a *range* of possible speaker-meanings), on the basis of knowledge of the literal meaning of the sentence plus the stage-setting. Such an approach conforms loosely to Dr Johnson's famous dictum that metaphor "gives you two ideas for one", because the interpreter has to retrieve the idea evoked by the literal meaning in order to discover the speaker's root idea.

Literature is full of examples where metaphors and other figures of speech help to convey what a person thinks. I want to focus upon metaphors occurring within *oratio obliqua.* Sometimes metaphors figure in indirect reports of what someone *said.* If a subject S spoke metaphorically, a reporter can, under suitable conditions, accurately report what S said by repeating S's original metaphor. Any adequate theory of metaphor must explain how it is possible for a metaphor to be "carried over" into a report.[2]

However, "carrying over" is a special case. Metaphors in "that" clauses are not always echoes. A speaker may choose to ascribe content to S's attitude by putting his own metaphor into sentence 'z'. In some cases S never expressed his attitude in words. In other cases, S expressed his attitude, but

[2] Cohen 1979 argues that "carrying over" is a problem for pragmatic theories of metaphor such as Searle's, but not a problem for a semantic feature-cancellation theory. (Cohen calls Searle's a "speech-act theory".) I believe that pragmatic theories *can* handle the "carry over" phenomenon; e.g. the relevance-theoretic approach can (see Sperber and Wilson 1986). But I do not argue for this here.

not in a metaphor. In those special cases where S expressed himself by means of a metaphor, R might use a *different* metaphor. Carry-over does not occur in any of these cases. Moreover, there is a fourth possible case. S spoke in metaphor, R reported S using a metaphor, and R chanced upon the same metaphor as S, without knowing that S had used it. Double employment of the same metaphor may seem far-fetched, but in fact it can be readily explained on the "Great minds think alike" principle. R and S might be psychologically and linguistically similar.

Suppose that R writes a biography of the English novelist D.H. Lawrence, intended for literary readers who know Lawrence's work. The biography contains a stretch of text (D5):

(D5) The prevailing literary culture in European countries of the early 1930's was hostile to Lawrence's philosophy. He found life in Europe intolerable, stifling, claustrophobic. Undoubtedly, this dissatisfaction was one of the motives which led to his decision to travel to the new world. At the time he set sail for Mexico, Lawrence believed that Europe was over-upholstered.

The punch-line contains a metaphor. The reader, aware that Lawrence did not believe that Europe literally had too many padded armchairs and sofas, will understand that the author is not imputing *that* belief. R is employing a metaphor to ascribe content to another belief, a belief which makes sense in the light of Lawrence's other attitudes. R expects the reader to gather this by following route two.

Which belief is it? A fair paraphrase of its content might be: "people in Europe were cossetted, hidebound by bourgeois conventions, excessively complacent, too interested in their own comfort, disposed to spend their lives in cosy drawing-rooms". The metaphorical ascription gets this message across more economically. The responsive reader first entertains a certain vision of Europe (by interpreting the metaphor), and then entertains the thought that Lawrence saw Europe in that way.

Since the biography is intended for literati, the ideal reader is someone familiar with Lawrence's style. Such a reader will recognise that, although the report was produced by the biographer, the metaphor in it is typically *Lawrentian*. "Overupholstered" is the sort of metaphor that Lawrence would have used, had he expressed his belief in words. There is nothing in (D5) which says explicitly that Lawrence had a penchant for metaphors like this. But suppose that other sections of the biography do discuss Lawrence's metaphors. These surrounding chapters of the book constitute the wider co-text of (D5). The biographer might hope and intend that the reader, having been supplied elsewhere with information about Lawrence's use of language, will squeeze out an extra message via route three. The biographer hints that the belief is one which, had Lawrence expressed it, Lawrence might well have expressed in that metaphorical way.

5 Conclusion

This paper has tried to show that the stage-setting of a content-ascribing utterance helps the hearer to work out which content the speaker ascribed. Two theses need to be highlighted.

(i) In order to determine which content a speaker ascribed, you have to look at what the speaker meant. In straight talk, the speaker means what he says and says what he means. But a lot of talk is loose. Speaker-meaning does not always coincide with sentence-meaning. Consequently, a theory of content-ascriptions cannot just analyse the semantic properties of propositional attitude sentence-types. Sentence semantics *underdetermines* the content which the speaker ascribes when he utters a sentence. Which content the speaker ascribes depends also on the speaker's situated intentions. Clues to these are often furnished by the stage-setting.

Some philosophers (e.g. Stich 1983) emphasize, rightly, that people's intuitions about the *truth or acceptability* of a content-ascription are context-sensitive. But people can't begin to evoke intuitions about whether a given ascription is true or acceptable, unless they first identify which content it ascribes. Their interpretation is partly determined by what they know of the particular stage-setting in which the act of ascribing occurred. Yet most theories of content-ascription (including Stich's) abstract from the contextuality of the act of ascribing.[3] This is a mistake.

(ii) Stage-settings tune the process of interpretation in many different ways. I distinguished three routes by which information that is explicit or implicit in the co-text can affect the hearer's interpretation of the content-ascribing punch-line.

In the first route, the co-text says or implies something about the language which the speaker is employing. This includes indications about the senses and references of the speaker's terms. When the speaker uses these terms in the punch-line, the hearer is ready to assign the intended senses and references.

The second route is where the co-text, through describing the scene and the subject, gives the hearer a reason to favour a certain construal. The hearer takes the punch-line to be ascribing a thought-content which fits in well with the rest of the story, provided that the semantics of the sentence allow (but do not determine) such a construal.

The third route is where the co-text informs the hearer about the subject's language (*langue* and/or *parole*). The hearer learns that the subject

[3] An exception is Rumfitt. He correctly notes that "in thinking about content, we need to take into account the context of the content-specifying report as carefully as we do the context of the saying which is being reported" (Rumfitt 1993: 445). The point holds for reports of propositional attitudes as well as reports of sayings.

might well use such and such words to express his thought, if he were to express it. This is a functional property of S's thought. The content ascribed by the speaker is a content constrained by that functional property.

References

Asher, N. 1986 "Belief in Discourse Representation Theory", *Journal of Philosophical Logic* 15, 127–189.

Cohen, L.J. 1979 "The Semantics of Metaphor", in A. Ortony (ed.), *Metaphor and Thought*, Cambridge: Cambridge University Press, pp. 64–77.

Haugeland, J. 1979 "Understanding Natural Language", *The Journal of Philosophy* 76, 619–632.

Heim, I. 1983 "File Change Semantics and the Familiarity Theory of Definiteness", in R. Bauerle, C. Schwarze and A. von Stechow, (eds.), *Meaning, Use and Interpretation of Language*, Berlin: de Gruyter.

Kamp, H. 1985 "Context, Thought and Communication", *Proceedings of the Aristotelian Society* LXXXV, 239–261.

Naipaul, V.S. 1987 *The Enigma of Arrival*, London: Penguin.

Rumfitt, I. 1993 "Content and Context: The Paratactic Theory Revisited and Revised", *Mind* 102, 429–454.

Spencer-Smith, R. 1987 "Survey: Semantics and Discourse Representation", *Mind and Language* 2, 1–26.

Sperber, D., Wilson, D. 1986 "Loose Talk", *Proceedings of the Aristotelian Society* LXXXVI, 153–171.

Stich, S. 1983 *From Folk Psychology to Cognitive Science*, Cambridge MA: MIT Press.

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

Semantic Localism: Who Needs a Principled Basis?

Michael Devitt

1 Introduction

Holism is usually accompanied by a "no-principled-basis" consideration along the following lines:

> There is no principled basis for the molecular localist's distinction between the few inferential properties of a token alleged to constitute its meaning and all its other inferential properties. Only a token that shared all the inferential properties of the original token would really share a meaning with it.

Many who are not sympathetic to holism are impressed by the no-principled-basis consideration. Fodor and Lepore (1992) are striking examples. I have argued elsewhere (1993b) that no good case has been made *for* the consideration. Even if this is right, it will seem to many that the localist who wishes to allow inferential properties a role in constituting meanings still faces a challenge: she must produce a case *against* the consideration. My purpose is to consider whether this is true. My approach will be naturalistic.

2 What Are Meanings? (I)

To fulfil this purpose, we must situate the issue properly. The dispute between holists and localists is over the nature of meanings. But what *are* meanings? The answer to this question is far from clear.[1] Indeed, the intractable nature of semantic disputes in general largely stems from differing opinions about what counts as a meaning. Bill Lycan has brought out the problem wittily with his "Double Indexical Theory of Meaning":

MEANING $=_{def}$ Whatever aspect of linguistic activity happens to interest me now. (Lycan 1984: 272)

We start semantics in the unusual position of having to specify a subject matter. We should not insist on great precision about this in advance of theory, but we do need some explication of our vague talk of "meanings". And to avoid Lycan's mockery, we must specify a subject matter worthy

[1] "The chief problem about semantics comes at the beginning. What is the theory of meaning a theory of?" (Higginbotham 1991: 271)

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 331-341

of investigation; we need an explication that is not *ad hoc*. To meet these needs, I propose to address three questions. First, what might plausibly be seen as our ordinary way of ascribing meanings? Second, what do we ascribe them to? Or, putting this another way, what are the phenomena that concern semantics? Third, what is our semantic interest in these phenomena? Or, putting this another way, what purposes are we serving in ascribing meanings?

3 Ascriptions of Meaning: Semantic Phenomena

What might plausibly be seen as our ordinary way of ascribing meanings? Consider the following: 'Ruth believes that Gorbachev has fallen' and 'Adam said that Yeltsin has risen'. Such "propositional attitude ascriptions" mostly use no semantic words but nevertheless seem partly to ascribe meanings. My working assumption for this section and the next is that they do indeed partly ascribe meanings: we specify meanings by the 'that' clauses in attitude ascriptions (also by their 'to' clauses that follow 'want,' etc., and by clauses that follow 'wonder whether'; briefly, by "t-clauses").

The meanings specified by t-clauses are complex. They are composed of simpler meanings like those specified by 'Gorbachev' and 'fallen'.

Some remarks about usage. I shall continue to talk of the "meanings" of thoughts where some would talk of their "contents". And I shall apply linguistic terms like 'word' to mental vehicles of meaning. Finally, I shall adopt a usage according to which meanings are *properties*.

We ascribe meanings to thoughts and utterances. So, thoughts and utterances, are the immediate phenomena of semantics. But thoughts, at least, are not so immediate to observation. What phenomena lead us to the view that an object has thoughts? Partly, the utterances of the object make us think this. Utterances are linguistic behavior. Clearly, the object's nonlinguistic behavior may also lead us to the view that it has thoughts.

4 Semantic Purposes

What is our semantic interest in these phenomena? What significant purposes - explanatory, practical, or whatever - are served by the ascription of meanings which, according to our working assumption, is part of the ascription of thoughts and utterances like those to Ruth and Adam?[2] My hypothesis is that people have two distinct purposes for these ascriptions: first, to explain and predict behavior, which I shall abbreviate "to explain behavior"; and, second, to form thoughts of their own using the thoughts and utterances of others as indicators. I shall consider these in turn. Our interest in thoughts is primary, our interest in utterances, secondary. So I shall start with thoughts.

[2]I have made several attempts to answer this question in the past, most recently in Devitt 1991, ch. 6. These answers all have similar elements but I think that none of them have been quite right.

1 Consider this explanation of nonlinguistic behavior:

> Why did Granny board the bus? She wants to buy a bottle of scotch. She believes that she has her pension check in her pocket. She believes that the bus goes to her favorite liquor store.

Such explanations of behavior are familiar and central parts of ordinary life, of history, of economics, and of the social sciences in general. They all ascribe thoughts with meanings specified by t-clauses.

Consider next this explanation of linguistic behavior:

> Why did Granny produce the sound /I need a drink/? She believes that she needs a drink. She wants to express her belief to her audience. She believes that /I need a drink/ expresses that belief.

Again we ascribe thoughts to explain behavior. Such explanations are not so common because they are so obvious. They are implicit in our responses to communications.

2 Ascribing thoughts serves another remarkably valuable purpose. These ascriptions provide information about the way a person believes the world to be, desires the world to be, hopes the world to be, and so on. As a result, we not only attempt to explain her behavior but also are influenced in our own thoughts about the world. Thus, if she desires a certain situation we may, for a variety of reasons (love, fear, etc.) come to desire it too and so try to bring it about. Another person's beliefs can be particularly useful as guides to external reality: if she believes that the world is such and such, and is reliable, then we have good reason to believe that the world is such and such. Thus, attributing the property of meaning that it is raining to Mark's belief not only helps to explain his rain-avoidance behavior but also gives us evidence about the weather.

Turn now to the ascription of *utterances*. Granny's behavior leads us to remark, 'Granny says that she needs a drink'. If the English-speaking Mark produces /It is raining/, we may respond, 'Mark says that it is raining'. If we think that Granny's and Mark's utterances are sincere expressions of beliefs, we will ascribe the same meanings to their beliefs as we do to their utterances. Utterances are indicative of thoughts. Indeed, it is because of our interest in thoughts that we are interested in utterances. Thus, it is because we want to use Granny's thoughts to explain her behavior that we are interested in her utterance. And it is because we want to use Mark's thoughts to explain his behavior and inform us about the weather, that we are interested in his utterance.[3]

[3]The story is a bit more complicated because of the Gricean distinction between the speaker meaning of an utterance and its conventional meaning on the occasion. It seems that we ascribe the latter with the 'says that' construction, yet we identify the former with the thought meaning if the utterance is sincere. We are interested in the conventional meaning as a guide to the speaker meaning with which it usually coincides. If we want

This is how things seem to us. But maybe we are all wrong in attempting to explain behavior and learn from each other in this way. Perhaps we are totally mistaken and should become behaviorists or eliminativists of some other sort. Or perhaps we should be less radically revisionist, ascribing only "narrow contents" or syntactic properties to serve our purposes. Still, my hypothesis is that, rightly or wrongly, we do ascribe meaningful thoughts and utterances to serve these purposes.

In this section, I have described two significant purposes: explaining behavior and forming thoughts by using indicators. Doubtless these purposes can be served in many ways. Our interest in meanings, together with the working assumption, have led us to two particular ways of serving them: directly, by ascribing to a token that is thought a property of the sort specified by t-clauses; indirectly, by ascribing such a property to a token that is uttered. Let us say that in attempting to serve the purposes in this way, our purposes are "semantic". And let us say that a property plays a "semantic role" if and only if it is a property of the sort specified by t-clauses and, if it were the case that a token thought had the property, it would be in virtue of this fact that the token can explain the behavior of the thinker or be used as an indicator by others.[4]

5 What Are Meanings? (II)

The discussion of the last two sections was prompted by the need for an explication of our talk of meanings that was not *ad hoc*; we need to specify a subject matter worthy of study. To meet these needs I adopted a working assumption: we specify meanings by the t-clauses of attitude ascriptions. Guided by this I have described purposes and roles which I have called "semantic". Now it is certainly worthwhile to study a property that plays a semantic role, that plays one or both of the roles just outlined in the explanation of behavior and the formation of thoughts. So, I propose to add the following explication to the statement of the basic task: a property is a meaning if and only if it plays a semantic role. We can now drop our working assumption: what we specify by a t-clause will count as a meaning if and only if it does indeed have a semantic role. Any other property will count as a meaning if and only if it has such a role.

It might be objected that this explication is too liberal because not just *anything* that had a semantic role would be a meaning. It is usual to think that meanings are constituted by relational properties: "internal" ones involving inferential relations among words and/or "external" ones involving certain direct causal relations to the world. So it might be claimed that we should modify our explication so that only properties with this "appropriate" sort of constitution count as meanings.

to draw attention to a difference between the two meanings it seems that we use the construction 'says that... but...means that...'

[4]Perhaps this should be broadened to allow tokens in, say, a visual module – hence not objects of thought – to have a semantic role, but I shall not attempt to do so.

Nothing in my argument hinges on my not adopting this modification, but I shall not adopt it. Note that my definition of "semantic role" slipped in the following constraint: a property has a semantic role only if it is "of the sort specified by t-clauses". So the "liberal" explication already places *this* constraint on meanings. In effect, the modification makes this somewhat vague constraint more precise by feeding in some of our theory of meanings. Yet that degree of precision seems undesirable in an initial attempt to specify the subject matter.[5]

This completes my explication of the talk of meanings: being a meaning is a property of certain properties, the ones that play a semantic role and hence that it serves our semantic purposes to ascribe. We have avoided the "*ad hoc*" charge.

6 Semantic Tasks

I want now to distinguish three semantic tasks. First, there is the basic task: to explain the natures of meanings. My explication relates this task closely to another that I will call the normative task. This is the task of explaining the natures of the properties we ought to ascribe for semantic purposes. Clearly, for an ascription to serve those purposes, the property ascribed must have one or both of the semantic roles. So, according to the explication, the property we ought to ascribe must be a meaning. On the other hand, *prima facie*, if something is a meaning then we ought to ascribe it for semantic purposes. Perhaps the reasons for thinking that a property plays a semantic role and hence counting it a meaning are not always sufficient to make it something we should ascribe for semantic purposes. However, if the property is really not one we ought to ascribe, the question of whether or not to count it a meaning becomes uninteresting: interesting meanings are ones we ought to ascribe. So, it will do no harm to adopt the *prima facie* view. So, we can assume that a property is a meaning if and only if it is one we ought to ascribe for semantic purposes.

In the light of this, I propose that we should tackle the basic task by tackling the normative one.

We are already familiar with tasks of saying what we ought to ascribe for certain purposes. So one advantage of this methodological proposal is that it relates the basic task to these other ones. Another advantage is the obvious contrast between the normative task and what I will call the descriptive task. This is the task of explaining the natures of the properties we do ascribe in attitude ascriptions for semantic purposes; the task of explaining the semantic *status quo*. It is what most people working in semantics – philosophers, linguists and psychologists – are in effect doing. Yet it is very different from the normative and basic tasks. Note particularly that

[5]It may be an advantage of the liberalism that it makes eliminativist about meaning harder: in general, the less that is essential to being an F the harder it is to show that there aren't any F's.

the properties it investigates are meanings only if ascribing them really does serve our semantic purposes; only if they really play semantic roles.

The above definition of the semantic tasks may be novel and is certainly not generally agreed. It would be interesting to contrast the view with others, but that must be left to another time (Devitt forthcoming a, b).

7 Does a Token Have More Than One Meaning?

Does a token have more than one meaning? Note that this question does not concern the familiar matter of ambiguity, which is a property of types.

Ordinary talk of "*the* meaning" of a word encourages the view that a token does not have, indeed cannot have, more than one meaning. This view is taken over in our semantic theories and disputes; for example, the dispute over semantic holism seems to be over whether the one and only meaning a token has is holistic or localistic. Yet the view seems to be nothing but a prejudice.[6]

I have noted that meanings are usually thought to be constituted by inferential relations among words and/or direct causal relations to the world. A token has many such relational properties and clearly different sets of them could constitute indefinitely many candidates to be meanings. To tell whether one of these really is a meaning we have to see whether it has one of the semantic roles. Prior to arguments one way or the other on this, we ought to take the property's candidature seriously. And we ought to take the possibility of a token having more than one meaning very seriously.

Given the situation in other areas, we should expect to find that a token does indeed have more than one. Thus, a particular person may have a variety of economic properties: being a capitalist, a landowner, a banker, and so on. Each of these properties plays a role that economic properties are supposed to: for example, explaining economic behavior. We should expect that a particular expression token will have a similar variety of semantic properties. Each of these will play a semantic role: one might explain one bit of behavior, another, another, and a third might serve as an indicator for the formation of thoughts.[7]

If the early Quine is anywhere near right in his view of ordinary attitude ascriptions then the folk ascribe more than one property to a token for semantic purposes. Quine distinguishes between "transparent" ascriptions, where only the reference of a word is of interest, and "opaque" ascriptions, where some finer-grained meaning of the word is of interest.[8]

[6]There is one respect in which some theories do not seem to follow the prejudice: under the influence of Grice they allow that a token has both a speaker meaning and a literal or conventional meaning, although these meanings are usually the same. What theories do not contemplate is that a token might have more than one speaker meaning, or more than one conventional meaning. That is the prejudice that concerns me here.

[7]The view that a token has more than one meaning seems to be an idea whose time has come. It has been independently suggested by Akeel Bilgrami (1992) and Eric Lormand (unpublished).

[8]And I suspect that the folk would be prepared to call either of these properties "meanings". If not, this would be an example of the way that the ordinary use of 'meaning' was not perfectly suited to theoretical semantics. What we are considering here is the ordinary application of 'meaning' to tokens. It is probably more often applied to types, to what convention makes common to tokens, the sort of thing that is to be found in a dictionary. So, the meaning of a type that has tokens that do not conventionally have the same referent – for example, the type 'she' – is not thought to include a referent. In contrast, if the tokens do conventionally have the same referent – as, for example, tokens of 'echidna' do – ordinary usage may often treat that referent as the meaning of the type.

Consider also the popular two-factor theories of meaning. These theories often describe purposes for ascribing meanings that are similar to the two I have outlined. They then assign two relatively independent meaning factors to a token to serve those respective purposes, a conceptual-role factor to explain behavior and a truth-referential factor to serve as an indicator for the formation of thoughts. They prefer to say that these factors jointly constitute *the* meaning of a token rather than that each severally constitutes a meaning of the token, but this preference seems to reflect only the prejudice that a token must have just one meaning.

In sum, I think that we should leave open the possibility that we ascribe more than one property to a token for semantic purposes (descriptive), that we ought to ascribe more than one (normative), and hence that a token has more than one meaning (basic).

8 The Descriptive No-Principled-Basis Consideration

What are we to make of the no-principled-basis consideration in the light of this discussion? I take it that the consideration is not making an *epistemological* point about how we tell what meanings are. Rather it is making a *metaphysical* point about how things are in reality. And the point seems to be a basic one about what meaning a token has. But perhaps the consideration is making a *descriptive* point about what we take a token's meaning to be; about the property we do, as a matter of fact, ascribe for semantic purposes; about its "putative meaning". I shall start with the descriptive point.

The point is that there is no principled basis in reality for the distinction between the few inferential properties of a token alleged to constitute the property that we ascribe to the token for semantic purposes and all its other inferential properties. Only a token that shared all the inferential properties of the original token would really share the ascribed property with it.

From the naturalistic perspective, there is nothing special about semantics. If the semantic localist really has an onus to produce a basis for this distinction then localists in other areas ought to have similar ones. So, we can assess the consideration in semantics by comparing it with analogous considerations elsewhere.

Take astronomy as an example. We ascribe the property being a planet to Mars. A brief investigation would show that some of Mars' properties constitute its being a planet but many do not. The distinction between these

two sorts of property needs no principled basis. No more does a similar distinction for the biological property being an echidna, the psychological property being a pain, the economic property being a capitalist, and the artifactual property being a hammer. These properties, like many others, are constituted localistically out of some properties and not out of others. That's the way the world is and nothing more needs to be said. Venus does not share all of Mars' properties and yet it still shares Mars' property of being a planet.

The same is true in semantics. We ascribe a putative meaning, LM, to a linguistic token for semantic purposes. An investigation might lead us to claim that some of the token's inferential properties constitute its having LM but many do not. We would need no principled basis for this distinction. The world may simply be such that LM is constituted localistically in this way and nothing more would need to be said. Another token that does not share all of the original token's inferential properties may still have LM.

9 The Basic No-Principled-Basis Consideration

It is more likely that the no-principled-basis consideration is making a basic point. I have just dismissed the need for a basis for the distinction between the few inferential properties of a token alleged to constitute what we take its meaning to be in our ascriptions and all its other inferential properties. The basic point concerns the meaning it has: there is no principled basis in reality for the distinction between the few inferential properties of a token alleged to constitute the meaning of the token and all its other inferential properties. Suppose, for example, that LM is indeed constituted by only a few of the inferential properties of the token. The point is that there is no principled basis for distinguishing those few as meaning constituters because there is no basis for distinguishing LM from any other set of the token's inferential properties as its meaning. Although all of its inferential properties are not equally LM constituters, they are equally meaning constituters. So its meaning is really HM, the set of them all. Only a token that had HM would really share a meaning with the original token.

We do need a principled basis here but it is not to distinguish one property as *the* meaning of the token: perhaps the token has several meanings. We need a basis to distinguish any meanings it has from its other properties; to distinguish the properties that have the second level property of being a meaning from those that do not. The framework I have developed provides one. A property is a meaning if and only if it plays a semantic role; i.e. if and only if it plays the role outlined in the explanation of behavior or the formation of thoughts; i.e. if and only if it is a property we ought to ascribe for semantic purposes. One inferential property may be distinguished from another in that the one but not the other constitutes a property that plays a semantic role and should be ascribed for semantic purposes; i.e that is a meaning. The basis for being localistic and not holistic in semantics is

that properties like LM, constituted by only a few inferential properties, are meanings and that holistic properties like HM, constituted by many, are not. Tokens that share such localistic properties really do share meanings. This is all the basis we need.

I have dismissed the descriptive point and found an easy answer to the basic point by drawing on our earlier discussion. The no-principled-basis consideration does, of course, leave us with an epistemic problem (although there is no reason to think that this was what prompted the consideration). The problem is that of showing that reality is as I have just claimed it to be: that localistic properties not holistic ones are meanings, and that one localistic property and not another is a meaning. This has become the problem of showing that our semantic purposes are served by ascribing certain localistic properties.

I summarize three arguments I urge elsewhere to show this (Devitt forthcoming b). First, all the properties we do ascribe for semantic purposes are in fact localistic. So, given the success of our current ascriptions in serving those purposes, we have good reason to suppose that the properties we ought to ascribe are localistic. Second, in general, whether our purposes are explanatory, practical, perhaps even frivolous, we tend to ascribe properties that are localistic because only localistic properties have the sort of generality we are interested in; localistic properties are likely to be shared by many things. This yields the simplest, least theory-laden, argument against semantic holism: we ought to ascribe localistic properties because only such properties have the generality that will serve our semantic purposes. Hence, only localistic properties play semantic roles. Hence, all meanings are localistic. Third, the popular overarching theory that word meanings are entirely constituted by referential properties provides a further argument, for no such meaning is holistic.[9]

10 Why is the No-Principled-Basis Consideration Appealing?

If I am right the no-principled-basis consideration that has caused so much concern in semantics turns out to be a damp squib. Why then is the consideration so appealing? The appeal is particularly striking in that it precedes any tackling of the epistemic problem.

A theory in any area must show that its purposes are served by ascribing the properties it does. I suggest that it would be surprising to discover that ascribing a holistic property *ever* served those purposes. It would certainly be surprising if there were some *general* reason why ascribing localistic properties *never* did so. If we could show that ascribing localistic properties did so in semantics, we would have shown that we have the only principled basis we need. Why then are so many philosophers convinced, in advance of

[9] In Devitt 1993a, I used this overarching theory to provide the required principled basis also. I still think that the theory provides a basis but the argument in the text is preferably because it is less theory-laden.

investigation, that we cannot provide a principled basis?

Part of the answer must be the appeal of the arguments for holism that I have considered elsewhere (Devitt 1993b). Beyond that, I think that the answer is, at bottom, the sheer difficulty of semantics: this surely is an area in which we do not know our way about. However, saying exactly where we are lost is hard.

My view of the no-principled-basis consideration reflects the framework set out in the early sections. My best attempt at saying where we are lost must do so also. I think that the trouble starts with the failure to explicate talk of meanings in terms of semantic roles and purposes. As a result the close relationship between the basic task of explaining what meanings are and the normative one of explaining what we should ascribe for semantic purposes is missed. The basic task is not then sharply distinguished from the descriptive task of explaining what we do ascribe for semantic purposes. And, finally, the basic no-principled-basis consideration, which the framework should be used to answer, is not sharply distinguished from the descriptive one, which should be dismissed.

I suspect that the failure to note that a token may have several meanings is important. It encourages a confusion between the set of a token's meanings and one particular meaning of it; hence, between the properties that are candidates to be constituters of some meaning or other of a token and the actual constituters of a particular meaning of it; hence, between all of its inferential properties and a few of them.

Finally, I think that an accident of history may have contributed to the appeal. Historically, semantic theories have presupposed the Cartesian idea that meanings are things that competent speakers must know about. The combination of this idea with molecular localism's distinction between the inferential properties that are constitutive of a meaning and the ones that are not has been thought to yield a distinction between what is known a priori and what is known empirically. This epistemic distinction was then taken as the criterion for the localist's distinction. Quine has shown that there is no such epistemic distinction and hence no such criterion. So it was thought that we must find another criterion (see, for example, Fodor and Lepore 1992). But there is nothing in this story to show that we needed a criterion for this purpose in the first place. And we do not need one as the discussion of the descriptive no-principled-basis consideration showed: we do not need a principled basis for distinguishing what constitutes a particular property from what does not constitute it. The Cartesian idea, which should have no place in semantics, presented us with a criterion that we did not need.[10]

[10]I am grateful for comments on earlier versions of this paper delivered at the First European Congress of Analytic Philosophy in Aix-en-Provence, France, April 1993, and in the "Language and Cognition" section of the 16th International Ludwig Wittgenstein Colloquium in Kirchberg, Austria, August 1993.

References

Bilgrami, A. 1992 *Belief and Meaning: The Unity and Locality of Mental Content*, Oxford: Basil Blackwell.

Devitt, M. 1991 *Realism and Truth*, Oxford: Basil Blackwell, 2nd edn revised (1st edn, 1984).

Devitt, M. 1993a "Localism and Analyticity", *Philosophy and Phenomenological Research* 53, 641–646.

Devitt, M. 1993b "A Critique of the Case for Semantic Holism", *Philosophical Perspectives* 7: Language and Logic, James E. Tomberlin (ed.), Atascadero: Ridgeview 1993, pp. 281–306. To be reprinted with a new Postscript in a special issue of *Grazer Philosophica*, Jerry Fodor and Ernest Lepore eds

Devitt, M. Forthcoming a "The Methodology of Natualistic Semantics", *The Journal of Philosophy.*

Devitt, M. Forthcoming b *Coming to Our Senses: A Program for Semantic Localism*, Cambridge: Cambridge University Press.

Fodor, J. A., and Lepore, Ernest. 1992 *Holism: A Shopper's Guide*, Oxford: Basil Blackwell.

Higginbotham, J. 1991 "Truth and Understanding", *Iyyun, The Jerusalem Quarterly*, 40, 271–88.

Lormand, E. Unpublished "A Marriage of Holism and Atomism".

Lycan, W. G. 1984 *Logical Form in Natural Language*, Cambridge, MA: MIT Press.

Processing Models for Non-Literal Discourse

François Récanati

1 Context and Comprehension

Comprehension often involves contextually assigning a particular value to some constituent of the uttered sentence. When the semantic value to be contextually assigned is one of the conventional "readings" of the expression, the contextual process of value assignment is called disambiguation or *sense selection*: An expression has more than one conventional sense, and comprehension involves selecting one of the senses as contextually appropriate.

There are many cases in which the semantic value contextually assigned to a constituent of the sentence is not linguistically conventional, however. For example there are cases of *contextual sense construction.* Sometimes, the conventional sense of the constituents of a complex phrase and the way they are grammatically combined is not sufficient to determine the semantic value of the complex phrase involving those constituents; the overall meaning of the complex phrase must be constructed in context. Thus 'he finished the book' can mean that he finished reading the book, writing it, binding it, tearing it into pieces, burning it, and so forth (Pustejovsky 1991; Cohen 1985); 'finger cup' will mean either 'cup having the shape of a finger' or 'cup containing a finger of whisky' or 'cup which one holds with one finger', or whatever (Bühler 1934, Kay and Zimmer 1967); 'John's book' can mean 'the book that John owns, wrote, gave, received', or whatever (Kempson 1977, Récanati 1989). In all such cases there is not a 'selection' from a limited range of preexisting interpretations for the complex phrase. Rather, an indefinite number of possible interpretations can be constructed in a creative manner.

Very similar to sense construction is the process of *sense specification*, by which a term which has a general meaning can be contextually interpreted in a much more specific sense (Atlas 1989; Bach forthcoming). For example the mass term 'rabbit' will be preferentially interpreted as meaning 'rabbit fur' in the context of 'He wears rabbit' and as meaning 'rabbit meat' in

The ideas in this paper were first sketched in London during an informal workshop on my book *Direct Reference* (June 1993); I thank the organizers of the meeting, Dan Sperber, Deirdre Wilson, and Robyn Carston, and the participants, especially Dan Sperber and Barry Smith, whose remarks contributed to shaping the present paper. Thanks are due also to Roberto Casati and the other organizers of the 16th International Wittgenstein Symposium on "Philosophy and the Cognitive Sciences" (August 1993), where I read the paper. Last but not least, I am grateful to Kent Bach and Geoff Nunberg (and also to my student Sylvie Lachize) for extensive discussion of the issues dealt with here.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 343-356

the context of 'He eats rabbit' (Nunberg and Zaenen forthcoming).[1] Again it's not a matter of selecting a particular value in a finite set; with a little imagination, one can think of dozens of possible interpretations for 'rabbit' by manipulating the stipulated context of utterance, and there is no limit to the number of interpretations one can imagine in such a way.

Sense construction can be analysed as involving a hidden variable (Partee 1984, Récanati 1989). For example, 'John's book' can be said to mean, 'the book that bears relation r to John', where the variable r must be contextually instantiated for the complex phrase to possess a definite semantic value. This analysis stresses the analogy between sense construction and another process to be contrasted with sense selection, namely *reference assignment.*

The semantic value of a referential expression is not the linguistic meaning of that expression (its "character", in Kaplan's terminology) but the reference which must be contextually assigned to the expression by virtue of the semantics of the language (Kaplan 1989, Récanati 1993). The linguistic meaning of the expression only helps to contextually identify the reference. It follows that the semantic value or "content" of the expression (viz. its reference) is contextually variable and does not possess the constancy of linguistic meaning. 'He's in Paris' can mean that John, Bob or *virtually anybody* is in Paris. Here also, we have a variable which must be contextually instantiated for the utterance to possess a definite semantic value. A reference has to be found (and can be freely sought) in context for the pronoun 'he', in the same way in which a particular relation has to be found (and can be freely sought) in context in order to construct the sense of the complex expression 'John's book'.

There is much to be said about each of these processes, and especially about the general picture of interpretation which emerges when they are taken seriously. In this paper, however, I will use them only to shed light on a further contextual process, by which a constituent of a sentence is contextually assigned an interpretation distinct from its literal interpretation, e.g. a metaphorical one or a metonymical one. I think it is worth comparing this process of *non-literal interpretation* to those listed above: sense selection, contextual sense construction, reference assignment and sense specification. When one uses the latter to shed light on the former, it turns out that the standard theory of discourse interpretation, inspired by Grice (1989), a theory which is accepted by a vast majority of researchers in semantics and pragmatics, is unfounded and deserves questioning.

According to the standard theory, the interpreter computes the proposition literally expressed by an utterance, and on the basis of this proposition and general conversational principles infers what the speaker means (which may be distinct from what is said, i.e. from the proposition literally

[1]The term 'specification' comes from Bach. Nunberg & Zaenen use 'precision' in the same sense.

expressed). In contrast to this dogma, I shall argue that the so-called "proposition literally expressed by an utterance" – whatever it is exactly – need not be computed in the process of interpreting a non-literal utterance.

2 Serial versus Parallel Processes

There has been a good deal of psycholinguistic research on the processing of ambiguity, out of which two basic models have emerged:

- According to the *serial model*, the most accessible candidate for semantic value is tried first; if there is a problem some backtracking takes place.
- According to the *parallel model*, all candidates – or at least, all candidates which reach a certain level of accessibility – are tried in parallel; the first candidate whose processing yields satisfactory results in the broader context of discourse is retained, while the others are suppressed.

Before considering whether and how these two basic models apply to phenomena other than ambiguity, a few clarificatory comments are in order.

(1) By a "candidate", I mean one of the semantic values which can be assigned to an expression in context. In the case of ambiguous words, the relevant semantic values – the relevant "candidates" – are the (multiple) conventional readings of the expression. For example the candidates for semantic value in the case of an ambiguous word such as 'bank' are the two readings 'river edge' and 'financial institution'. But the notion of a candidate for semantic value extends beyond the case of ambiguous words; it applies whenever several distinct values can be contextually assigned to an expression. Thus the literal and the metaphorical interpretation of an expression are two "candidates"; different possible referents for a pronoun in a particular context are "candidates", and so forth.

(2) Accessibility of a mental representation is an intuitive notion which can be made precise in many ways. According to Barsalou and Billman (1989), accessibility is a multi-factor phenomenon which involves frequency of processing, recency of processing and contextual relevance. Sperber and Wilson (1986) characterize accessibility in terms of recency of processing and/or associative links to recently processed information; they don't invoke relevance as they use accessibility in *defining* relevance. Following these authors I will assume that accessibility, for a mental representation, involves the following factors (including a recursive one): recentness of processing, close associative links to accessible representations, and frequency. As accessibility so characterized corresponds to the intuitive notion of salience, I will use both terms indifferently. Another expression I will use as equivalent to 'accessible' is 'activated'. Indeed, the three factors in terms of which I

have characterized accessibility can easily be rephrased in the vocabulary of activation and activation spreading: Some representations are activated because they are being currently processed or have been recently processed (first factor); they spread their activation to associatively related representations (second factor); and frequency of processing can be conceived of as lowering the activation treshold of the representation (third factor).

(3) The opposition between serial and parallel models should not be taken too strictly; it is, to a large extent, a matter of degree. If parallel processing comes with a bias, the parallel model may come close to the serial model (Garman 1990:360). Be that as it may I will not be concerned with the details here, but only with the general tendencies embodied in the two basic models.

The two models apply neatly in the case of reference assignment. On the parallel model, all (sufficiently accessible) candidates for the status of referent are tried in parallel in such a way that the first one which yields satisfactory results in the broader context of discourse is retained, the others being suppressed. On the serial model, the most accessible candidate is tried first and backtracking occurs only if there is a problem. Consider, for example, utterance (1) (from Récanati 1993:265):

(1) John was arrested by a policeman yesterday. He had just stolen a wallet.

In order to interpret the utterance one must assign a reference to the pronoun 'he' in the second sentence. The two persons who were mentioned in the first sentence, namely John and the policeman, are obvious candidates, and we may suppose that there is no other (sufficiently accessible) candidate. (There would be one if, for example, the speaker pointed to someone while uttering the pronoun.) According to the serial model, if one of the two candidates is more accessible than the other, it is processed first, and the resulting interpretation of the utterance is retained unless it turns out to be unsatisfactory. According to the parallel model, both candidates are processed in parallel, i.e. both interpretations ('the policeman had just stolen a wallet' and 'John had just stolen a wallet') are constructed and entertained until one is found satisfactory and the other is suppressed.

What about non-literal interpretations? Can we apply the two models in this domain also? In the literature only one model is considered seriously: the serial model. The parallel model is taken to be ruled out by the particular structure of non-literality, namely the fact that there is an *asymmetry* between the literal and the non-literal interpretation: the latter presupposes the former. This is so because the process of non-literal interpretation consists in inferentially deriving one interpretation from the other (Grice 1989). Now this means that non-literal interpretation proceeds *serially*. On the

inferential model of non-literality, we process first the literal interpretation, and go on to the non-literal interpretation when this is required to make sense of the speaker's utterance.

I agree that the inferential model of non-literality, which is accepted by almost everybody today, entails rejecting the parallel model of non-literal interpretation. For the inferential model of non-literality embodies a particular version of the serial model, which I call the LS model (the "literality-based serial model", according to which the literal interpretation is processed first). Still, I think that the parallel model is *not* ruled out: there is still the possibility of rejecting the inferential model of non-literality. This is the line I will take in this paper.

3 A Parallel Model for Non-Literal Discourse

What makes the serial model so attractive is the fact that the non-literal interpretation is derived from, and presupposes, the literal one. This I do not wish to deny – I admit that the literal interpretation comes first. Still, I want to resist the conclusion that the literal interpretation must be 'processed' first. To say that the literal interpretation is processed first is to say that we have a two-step procedure: (i) the interpreter accesses the literal interpretations of all constituents in the sentence and uses them to compute the proposition literally expressed; (ii) on the basis of this proposition and general conversational principles he or she infers what the speaker means (which may be distinct from what is said, i.e. from the proposition literally expressed). What I am disputing is not the claim that the literal interpetation of the constituent is accessed before the non-literal interpretation - that I take to be obvious - but the claim that the process of semantic composition which consists in putting together the semantic values of the parts to determine the semantic value of the whole begins by paying attention only to literal semantic values, and turns to non-literal values only after the literal semantic value of the whole (the proposition literally expressed) has been computed. It is this picture which I think is unwarranted.

If I am right, the asymmetric structure of non-literality (the fact that non-literal interpretations presuppose literal interpretations) does not rule out a parallel model such as the following:

Parallel Model of Non-Literal Interpretation The literal meaning of the relevant constituent is activated, and this activation automatically (i.e. non-inferentially) spreads to various associatively related representations; the latter are potential candidates for the status of semantic value. All candidates, whether literal or non-literal, are processed in parallel and compete. When an interpretation which fits the broader context of discourse is found, the others are suppressed.

On this view, non-literal interpretations still proceed (associatively) *from* literal interpretations, which they presuppose; but, although gener-

ated serially, they are processed in parallel. The literal interpretation has no privilege over non-literal interpretations; they compete and it is possible for some non-literal interpretation to be retained (if it fits the broader context of discourse) while the literal interpretation is suppressed. In other words, the non-literal interpretation is *associatively* derived from the literal interpretation, but not *inferentially* derived. Inferential derivation entails computation of the literal value of the global sentence (which serves as input to the inference), while associative derivation is a "local" process (Récanati 1993) which does not require prior computation of the proposition literally expressed.

Consider, as an example, Geoff Nunberg's famous ham sandwich. The waiter says 'The ham sandwich has left without paying'. On the LS model the interpreter computes the proposition literally expressed by the sentence – namely the absurd proposition that the sandwich itself has left without paying – and from its absurdity infers that the speaker means something different from what she says. On the parallel model the description 'the ham sandwich' first receives its literal interpretation, in such a way that a representation of the ham sandwich is activated; activation is then spread to related representations, including a representation of the man who ordered the ham sandwich. *All* these representations activated by the description 'the ham sandwich' are potential candidates for the status of semantic value of the expression; all are equally susceptible of going into the interpretation of the global utterance. Now the ham sandwich orderer is a better candidate than the ham sandwich itself for the status of argument for '... has left without paying'. It is therefore the derived, non-literal candidate which is retained, while the literal interpretation is discarded.

An important difference between the LS model (according to which the literal interpretation is processed first) and the parallel model just outlined is this: on the parallel model is it possible for an utterance to receive a non-literal interpretation without the literal interpretation of that utterance being ever computed. The non-literal interpretation of the global sentence does not presuppose its literal interpretation, contrary to what happens at the constituent level. If the non-literal interpretation of some constituent fits the context especially well it may be retained (and the other interpretations suppressed) *before* the literal interpretation of the sentence has been computed. Whether or not this sort of thing actually happens, this is at least *conceivable*, on the parallel model. (On the LS model this situation is ruled out in principle.)

Of course I am not saying that the LS model never fits the facts – that we never compute the proposition literally expressed and infer what the speaker means. I fully admit that there are conversational implicatures which work exactly that way (even though not everything which has been treated as conversational implicature in the literature on these subjects belong to the same category [Carston 1988, Récanati 1989, Levinson forthcoming]). I am only saying that there is no conceptual necessity why we should

(always) have to wait until the literal interpretation of the sentence is computed and contextually tested before trying non-literal interpretations. In particular, the fact that the non-literal interpretation of a constituent presupposes the literal interpretation of that constituent does not entail that the non-literal interpretation of the whole sentence also presupposes the literal interpretation of that sentence.

4 The Accessibility-Based Serial Model

So far I have mentioned two models for non-literality, the literality-based serial model (LS model) on the one hand and the parallel model. But another model deserves to be considered.

In the case of disambiguation and reference assignment also I made a distinction between two models, a serial one and a parallel one; but if one looks closely one sees that the two models distinguished for reference assignment and disambiguation do not coincide with the two models distinguished for non-literal discourse. The parallel model is in all cases the same, but the serial model envisaged for dealing with non-literality is quite different from the serial model envisaged for dealing with reference assignment and disambiguation. The former is a literality-based serial model, while the latter is an accessibility-based serial model.

Recall what was said concerning reference assignment. Whereas, on the parallel model of reference assignment, all (sufficiently accessible) candidates for the status of referent are processed in such a way that the first one to yield satisfactory results is retained, on the serial model the most *accessible* candidate is tried first. It is accessibility, not literality, which matters for reference assignment, for one very good reason: the various candidates involved are on the same footing and cannot be discriminated in terms of literality. In example (1) above, John and the policeman can both serve as referent for the pronoun 'he' without either of them being more "literal" than the other. It follows that, if one of the candidates is processed before the others, that cannot be because it is literal while the others aren't.

The same thing holds of ambiguity. In a process of sense selection the candidates are all literal interpretations of the expression – none is *more* literal than the others. It follows that if one is processed first, this can't be because it's the literal interpretation. In all those cases (reference assignment, sense selection, sense construction) we find the same pattern: the candidates are all on the same footing with respect to literality, so, if a serial model applies and one of the candidates is processed first, this cannot be because that candidate is literal (while the others are not). What determines that one candidate is processed before the others can only be accessibility – the fact that one candidate comes to mind more readily than the others.

To sum up, whenever the candidates are on the same footing with respect to literality, the only serial model available is the accessibility-based serial model (AS model), according to which the most accessible candidate is

processed first. But when it is possible to distinguish one of the candidates as being literal, another serial model is available, namely the literality-based serial model (LS model), according to which the literal candidate is tried first. In such cases we have three models to consider: the parallel model, the LS model and the AS model.

Note that sense specification and sense construction do not fall into the same category here. The candidates for the status of semantic value in a case of sense construction are on the same footing with respect to literality; whichever relation holds between John and the book in a contextual interpretation of 'John's book', that relation will not be more literal than another relation which would be relevant in a different context. The situation is very similar to that of reference assignment. But in a case of sense specification, there is one particular interpretation which is more literal than the others, namely the 'general' reading. The latter is more literal in the sense that it corresponds to what is linguistically encoded, while the contextually specified readings go beyond what is encoded. Thus the mass term 'rabbit' literally means 'rabbit stuff'. The other, more specific interpretations ('rabbit meat', 'rabbit fur', and so forth), are in a sense less literal, whence it follows that a literality-based serial model ought to be available for sense specification as it is for non-literal interpretation.[2]

There is more to be said in this connection, but I cannot elaborate. The only thing that matters is that for non literal discourse as well as for sense specification there are, in principle, three models to consider: the parallel model and *two* serial models, viz. the AS model and the LS model. All candidates may be processed in parallel, or the most accessible may be processed first, or the literal candidate may be processed first.

The parallel model and the AS model share a common feature as opposed to the LS model. On the LS model the literal interpretation of the sentence must be computed, whether or not it is the ultimately correct interpretation of the utterance. On both the parallel model and the AS model the non-literal interpretation of the global sentence does *not* presuppose its literal interpretation, contrary to what happens at the constituent level. In this paper I am not arguing specifically in favour of either of these two models; rather, I am arguing against the LS model, and especially against the inferential picture which supports it, by drawing attention to several plausible alternatives.

5 Accessibility-Ranking and Multiple Activation

At this point, an advocate of the LS model may be tempted to deny that there are three distinct options in play. The two serial models, she may argue, are equivalent, as it is not possible for some interpretation distinct

[2]Indeed I think sense specification must be considered as a particular variety of non-literal interpretation based on "enrichment" rather than "transfer" (Récanati 1993). (According to Dan Sperber there is another sort of non-literality based on impoverishment rather than enrichment. A similar idea has been put forward by Kent Bach.)

from the literal interpretation to be more accessible than the latter. After all, non-literal interpretations on the picture I have sketched are accessed *through* literal interpretations. Now whatever activation is spread from the representation which constitutes the literal interpretation of the expression to other, associatively related representations cannot exceed the amount of activation present in the literal representation where the process of activation spreading originates. It follows that a non-literal interpretation cannot be more accessible than the literal interpretation from which it proceeds. Thus, even if we accept the AS model, according to which the most accessible interpretation is processed first, we need not reject the LS model, according to which the literal interpretation is processed first; for the literal interpretation *is* the most accessible one. If this is true we gain some generality; we can have a single serial model for all cases (ambiguity, sense construction, reference assignment, sense specification and non-literal interpretation), namely the AS model, and consider the LS model as a particular version of the AS model (appropriate for cases where one candidate is distinguished as "literal").

Although I find it highly desirable to have a single model for all cases, I strongly reject the conclusion that the literal candidate *is* the most accessible one. The reasoning which leads to that conclusion involves a fallacy very much like that which I exposed earlier, to the effect that only a serial model is available for non-literal discourse. The earlier reasoning used the asymmetric nature of non-literal meaning (its "asymmetric dependence" on literal meaning) as evidence that the literal interpretation must be processed first. The fallacy lied in the fact that 'the literal interpretation is processed first' is ambiguous: there is a distinction between processing the literal interpretation of the constituent (accessing it, as it were) and processing the literal interpretation of the *sentence* where it occurs. The former process is local, the latter global. The only thing which the asymmetric nature of non-literal interpretation entails is the priority of literal interpretations at the constituent level, but this is consistent with the absence of priority of literal interpretation at the sentence level. The parallel model and the AS model both incorporate this absence of priority.

The new fallacy, to the effect that there is no genuine distinction between the LS model and the AS model, also uses the asymmetric nature of non-literal interpretation as premiss. This asymmetry, in conjunction with the principle that a representation cannot transmit more activation than it possesses, is said to entail the following conclusion: that non-literal interpretations, insofar as they proceed from literal interpretations which spread activation to their neighbours, cannot be more accessible than the literal interpretations from which they get their activation. This is a fallacy, for, as Freud noticed long ago (Freud 1900, ch. 6), activation can be *over*-determined – it may come from different sources. Owing to this possibility of multiple activation, a representation may have a higher degree of activation (hence be more salient or accessible) than one of the representations from

which it gets (part of) its activation, even if we accept that a representation cannot transmit more activation to associatively related representations than it possesses.

Let me say more precisely where I think the fallacy in the second literalist reasoning lies. The reasoning involves two basic principles as premisses, namely the Conservative Principle (CP) and the Principle of Asymmetric Dependence (PAD):

(CP) A representation cannot transmit more activation than it possesses.

(PAD) Non-literal interpretations are accessed through literal interpretations.

The two principles together are said to entail that non-literal interpretations cannot be more accessible than literal interpretations (since they proceed from the latter, and a representation cannot transmit more activation than it possesses). But to get this result, we need something stronger than the PAD, and this is where the fallacy lies. In order to conclude (from the Conservative Principle) that non-literal interpretations cannot be more accessible than literal interpretations, we need what I call the *strengthened* PAD:

(Strengthened PAD) Non-literal interpretations are accessed through, and *only* through, literal interpretations.

If non-literal intepretations are accessed through, and only through, literal interpretations, then whatever activation they have come from the literal interpretations; it follows that they can't be more accessible than the latter (in virtue of the Conservative Principle). But is it true that non-literal interpretations are accessed *only* through literal interpretations? No. The literal representation may be only one among many factors which contribute to the activation of the associatively related representation. As a result the latter may get a higher degree of activation than the literal representation itself possesses. It follows that the non-literal interpretation *can* be more accessible than the literal interpretation from which it proceeds, as a result of the simultaneous influence of *several* representations.

Consider the case of sense specification I used as example. When someone talks of 'wearing rabbit', the representation 'rabbit fur' is activated as the result of (i) transmission of activation from the representation 'rabbit' linguistically activated (*qua* literal semantic value of the term 'rabbit'), and (ii) transmission of activation from the associatively related representation 'wearing' which is also linguistically activated. It is therefore theoretically possible for the representation 'rabbit fur' to be more activated, in this context, than the more general representation 'rabbit'.[3] Whatever we think

[3]This shows that one cannot consider (the mental representation) 'rabbit fur' as being compositionally constructed out of the representations 'rabbit' and 'fur'. It has to be considered as an independent concept – perhaps a *de re* concept (Récanati 1993).

of this particular example, it seems to me that the following situation can arise: An expression linguistically encodes a certain representation; the latter is accessed when the expression is uttered, but another, associatively related representation is also activated (in part – but in part only – through the encoded representation) and turns out to be more salient in that context than the original representation from which (in part) it derives. On both the parallel model and the AS model, the derived representation goes directly into the overall interpretation of the utterance, without depending on the prior computation of the literal interpretation of the utterance involving the original representation. On the AS model *only* the derived representation goes into the overall interpretation of the utterance; on the parallel model both the derived representation and the literal representation go "directly" into (distinct and parallel) interpretations of the utterance, without one interpretation being asymmetrically dependent on the other.

Multiple activation need not be simultaneous. Some representation may be accessed through its links to the representation linguistically encoded and *become* more accessible than the latter as a result of the coming into play of further linguistic material. Consider sentence (1) again, or rather a variant of (1) in which the indefinite description 'a policeman' has been replaced by a name of the policeman to simplify matters:

(2) John was arrested by Mike yesterday; he had just stolen a wallet.

We may suppose that when the pronoun 'he' is uttered Mike is slightly more accessible than John as a candidate for semantic value, for Mike was mentioned last. But as the other constituents of the sentence are processed the accessibility-ranking of the candidates changes. When the predicate is uttered, John becomes more accessible than Mike as a candidate for the status of referent of 'he'. The explanation for this fact is very simple. John is the subject of 'was arrested' and therefore occupies the role of the person being arrested; now that role is linked to the role of the person doing the stealing, in some relevant frame (on frames, see for example Fillmore and Atkins 1992). Because of this link, the representation of the referent of 'he' as the person doing the stealing contributes some activation to the role of the person being arrested and therefore to John *qua* occupier of this role. John thus gains some extra activation and becomes the most accessible candidate.

John and Mike are on the same footing as far as literality is concerned (Section 4), but if we turn to a case in which some candidate is distinguished from the others by its being literal, we see that the same sort of temporal shift in accessibility can occur. When the words 'the ham sandwich' are uttered, we may consider that the representation of the ham sandwich is more accessible than other, related representations which are activated through their links to that representation. (This is a straightforward application of the CP.) Thus we may suppose that the representation of the ham sandwich is more accessible (more salient) than the "derived" representation of the ham sandwich orderer. This is similar to the fact that Mike is more access-

ible than John when the pronoun 'he' in (1) is uttered, except that Mike is more accessible because he was mentioned last, while the representation of the ham sandwich is more accessible because it is linguistically encoded. In both cases, the initial ranking is reversed when further linguistic material comes into play. After the predicate in the sentence 'the ham sandwich has left without paying' has been processed, the ham sandwich is no longer a more accessible candidate than the ham sandwich orderer – the order of accessibility is reversed. The explanation, again, is very simple and does not appeal to inference on the hearer's part. The predicate 'has left without paying' demands a person as argument; this raises the accessibility of all candidates who are (represented as) persons. In this way the ham sandwich orderer gains some extra activation which makes him more accessible than the ham sandwich, after the predicate has been processed.

6 Conclusion

In this paper I have argued against a widely accepted model of utterance interpretation, namely the LS model, according to which the literal interpretation of an utterance (the proposition literally expressed by that utterance) must be computed before non-literal interpretations can be entertained. I have tried to make room for the possibility that an utterance receives a non-literal interpretation (i) as a result of a *local* process concerning only a particular constituent of the sentence, and (ii) without the literal interpretation *of the utterance* necessarily being computed. On my view, the literal interpretation of the constituent must be accessed, but there is no need to take it into the global interpretation of the utterance; the latter may recruit only the non-literal interpretation of the constituent, in such a way that the literal interpretation of the global utterance is not computed, or at least need not be computed for the non-literal interpretation to be entertained. (It may be computed in parallel to the non-literal interpretation, but at least the latter is not asymmetrically dependent on the former, contrary to what the LS model claims.)

Two models were presented as alternatives to the LS model. On the parallel model, literal and non-literal interpretations of the utterance are simultaneously processed. On the AS model, only the most accessible candidate for the status of semantic value of the relevant constituent goes into the overall interpretation of the utterance; if the most accessible candidate is not the literal semantic value of the constituent, the literal interpretation of the sentence is not computed – only the non-literal interpretation of the sentence (involving the highly accessible but non-literal interpretation of the constituent) is delivered.

I have considered two arguments in favour of the LS model. The first argument purports to show that the parallel model is ruled out in principle, because of the asymmetric nature of non-literality (the fact that non-literal meaning asymmetrically depends on literal meaning). The second argument also appeals to the asymmetric nature of non-literality to rule out the AS model construed as an alternative to the LS model. Both arguments have been shown to be fallacious: the asymmetric dependence of non-literal meaning on literal meaning does *not* support the LS model as opposed to the parallel model or the AS model. If I am right the LS model is baseless: there is no particular reason why the process of non-literal interpretation should proceed serially through computing the literal interpretation of the sentence. At the very least the LS model is far from obvious and cannot be simply taken for granted (as it often is).

The most radical claim contained in this paper is the suggestion that the proposition literally expressed need not be computed in interpreting an utterance; this is the gist of the AS model. How radical is that claim? Advocates of the traditional view will point out that it is not so radical as it may seem, for it concerns only *processing*. Granted; but I think it is possible to go much further and deny that there (always) *is* such a thing as the proposition literally expressed by an utterance, whether computed or not. We can imagine cases in which a contextual process of sense construction for a complex phrase (requiring contextual instantiation of a free variable) appeals to the non-literal sense of a constituent of the complex phrase; in such a case the utterance would be semantically indeterminate at the purely literal level as the instantiation of the free variable, without which the proposition literally expressed cannot be computed, would itself depend on some feature of the non-literal interpretation of the utterance and therefore could not be achieved at the literal level. My final claim, which goes beyond the scope of the present paper, is that there would be *no* "proposition literally expressed" in such a case.

References

J. Atlas 1989, *Philosophy without Ambiguity*, Oxford: Clarendon.

K. Bach (forthcoming), "Semantic Slack".

L.Barsalou & D. Billman 1989, "Systematicity and Semantic Ambiguity", in D. Gorfein (ed.), *Resolving Semantic Ambiguity*, New York: Springer.

K. Bühler 1934, *Sprachtheorie*, Jena: Gustav Fischer. English Translation by D. F. Goodwin, *Theory of Language*, Amsterdam: Benjamins 1990.

R. Carston 1988, "Implicature, Explicature, and Truth-Theoretic Semantics", in R. Kempson (ed.), *Mental Representations: The Interface between Language and Reality*, Cambridge: Cambridge University Press.

L. J. Cohen 1985, "How is Conceptual Innovation Possible", *Erkenntnis* 25, 221–38.

C. Fillmore & B. Atkins 1992, "Toward a Frame-Based Lexicon: The Semantics of RISK and its Neighbors", in A. Lehrer & E. F. Kittay (eds.), *Frames, Fields and Contrasts: New Essays in Semantic and Lexical Organization*, Hillsdale: Lawrence Erlbaum.

S. Freud 1900, *Die Traumdeutung*, Leipzig: F. Deuticke.

M. Garman 1990, *Psycholinguistics*, Cambridge: Cambridge University Press.

P. Grice 1989, *Studies in the Way of Words*, Cambridge, MA: Harvard University Press.

D. Kaplan 1989, "Demonstratives", in J. Almog, H. Wettstein and J. Perry (eds.), *Themes from Kaplan*, New York: Oxford University Press.

P. Kay & K. Zimmer 1976. "On the Semantics of Compound and Genitives in English" *Proceedings of the Sixth California Linguistics Association Conference*, San Diego: Campanile.

R. Kempson 1977, *Semantic Theory*, Cambridge: Cambridge University Press.

S. Levinson (forthcoming), "Generalized Conversational Implicatures and the Semantics/Pragmatics Interface".

G. Nunberg & A. Zaenen (forthcoming), "Systematic Polysemy in Lexicology and Lexicography".

B. Partee (1984), "Compositionality", in F. Landman & F. Veltman (eds.), *Varieties of Formal Semantics*, Dordrecht: Foris.

J. Pustejovsky 1991, "The Generative Lexicon", *Computational Linguistics* 17:4, 409–41.

F. Récanati 1989, "The Pragmatics of What is Said", *Mind and Language* 4, 295–329. Reprinted in S. Davis ed., *Pragmatics: A Reader*, Oxford: Oxford University Press 1991.

F. Récanati 1993, *Direct Reference: From Language to Thought*, Oxford: Blackwell.

D. Sperber & D. Wilson 1986), *Relevance: Communication and Cognition*, Oxford: Blackwell.

Constraints on Universals

Alberto Peruzzi

1 Introduction

Philosophy is again at a turning point. After the linguistic turn and the cognitive turn, it has become progressively clear that neither a view of semantics centred on the purely logical structure of language, nor one based on computational models of mind (and its "language"), can explain the actual interface of the linguistic capacities with the systems of vision and bodily motion, as activated in a natural environment. Whereas it is precisely through that interface that meaning flows. Nonetheless we are finally in the position of having rigorous theoretical tools, capable of capturing a broad range of results of cognitive science, in such a way as to provide an empirically significant solution to a classical problem: that of identifying the universals that underlie intentionality.

To this end, it has to be acknowledged that the alternative in front of us no longer consists either in the recourse to some formalism passing over the complexity of phenomenological structures, or the renunciation of any attempt to construct a unified theory – in the conviction that, at best, one could achieve pragmatic generalisations about the use of language in context, while supposing the notion of context itself is not principled. Rather, the decisive progress of semantic theory stems from grasping the *depth of surface*, in other words the fact that the roots of semantics lie in the features of macro-objects described by algebraic and differential topology, and in related dynamical schemes of interaction.

It thence follows that we have to develop a systematic analysis of the constraints on the constitution of the possible objects of reference in common experience. From this perspective, the identification of semantic universals rests on recent research concerning

i the mental structures involved in making possible our talk of what we see and our seeing what we talk about, and

ii the bodily structures involved in factorising the continuum of macrospace into mereological entities.

Of course, both kinds of structures are closely related to the cosmological "window" whithin which the genesis and the stability of our natural environment are made possible. Yet, in this paper I shall not enter into the philosophical questions concerning "naturalism", a by now over-extended

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 357-369

concept, that covers the most diverse views – for this reason I have used elsewhere[1] the term 'entwined naturalism' in order to characterise the dialectical unity of subject and object (metatheoretically speaking, of epistemology and ontology). In this connection, I wish to emphasise that a precise mathematical description of such a unity is now at hand through the development of category theory, and it is from a categorical point of view that semantics is reconceived in the present paper.

2 The "Logik der Abbildung" and the Dogmas.

Proposition 4.015 of the *Tractatus* stated that "Die Möglichkeit aller Gleichnisse, der ganzen Bildhaftigkeit unserer Ausdruckweise, ruht in der Logik der Abbildung". Only today are we in a position to do justice to the depth of Wittgenstein's claim, through the recognition that there are "Abbildung"-schemes of an imaginative character. Such schemes are essentially grafted onto a topological basis and play a fundamental rôle in extracting semantic structure out of perceptual and environmental constraints. The way this graft works can be described in categorical terms (fibrations and adjoint functors). For category theory makes room for a non-punctiform view of the structure of objects and actions inhabiting the base space (generalised and enriched in structure) of macrophysical bodies and their mutual interactions; and moreover, it allows a rigorous description of the lifting of schemes from the base space to the whole of cognition (logical syntax included), as represented in language.

The idea underlying this application of category-theoretic concepts and techniques in the domain of semantics of natural language is that a *real* phenomenology of the emergence of meaning has a principled basis, and such a basis is directly related to the foundations of mathematics. So, there is no ontological gap between formal structures, e. g. of a computational sort, and their material "implementation", for the simple reason that there is nothing implemented that wasn't already built into the architecture of dynamical functions proper to a natural system, governed by physical and morphogenetic laws, both of these essentially paired to structures of real space. Thus, separation of the mind from the body is itself only an operation of the mind. In fact, even our understanding of syntax is "space-laden". And yet, the reduction of intentionality to the physiology of inner "processing" units is prevented, see Peruzzi 1993a.

Analytic philosophy, as generally conceived and practiced in the twentieth century, didn't provide room for a theory of the basic schemes mentioned above, nor was this aim achieved by mainstream cognitive science, inspired by digital computer design. Thought was taken by analytic philosophers as represented in language, and susceptible to being investigated only in this medium. The structure of language could and had to be formalised by means of logic, and logic itself was given a semantics within the universe of

[1] Peruzzi 1993.

set theory. (As a byproduct, all non-linguistic components of thought were progressively marginalised.)

Whithin analytic philosophy, the intensionalist tradition, centred on the notion of meaning, met insurmountable difficulties in being faithful to the dogma that semantic competence can be fully characterised in logico-linguistic terms. The intension of any designator δ is defined as a function that assigns to each index $\langle \mathbf{i} \rangle$ (the n-tuple formed by possible world, time and contextual parameters) the extension of δ at $\langle \mathbf{i} \rangle$. It follows that any intension is just a set, as long as functions are sets of ordered pairs. So, the constructive, conceptual, aspect of meaning is lost, as is any connection with mental capacities. Our intuitive understanding of what a function is, also vanishes.

On the other hand, the extensionalist tradition, centred on the primacy of the notions of truth and reference, met analogous difficulties. The gap between language understanding and actual features of the macro-world is indeed hard to bridge if one preserves the autonomy of semantics from the sensory-motor abilities involved in determining the location, shape, solidity, colour, etc. of the objects we refer to (all the more hard, if appeal to such referential resources as manifested in human beings is doomed to commit the psychologistic "fallacy"). Nor can one bridge the gap while being faithful to the dogma of a set-theoretic framework for semantics.

It follows that set theory cannot provide the right framework for a faithful analysis of thought. So, some (or all) of the suppositions made by analytical philosophers have to be drastically revised, or abandoned. An alternative that has been widely canvassed consists in recasting the linguistic turn in terms of cognitive science, in such a way as to maintain the first supposition: the analysis of thought becomes the analysis of one or more programming languages, in which to express the various cognitive operations, while reducing the abstract generality of logic through a formalisation completely in terms of computability theory. But this theory, as presented in terms of λ-calculus, cannot have a semantics in the category of sets (objects=sets, maps=functions). If the operations of thought can always be objectified and self-applied, they have to form a set X of symbolic manipulations, closed with respect to exponentiation, so that $X^X \cong X$, whereas the only set with this property is a singleton (up to bijection), and this evidently fails to be a serious candidate for modelling thought. True, one can pass to a typed language, but as soon as this move is made, a connection emerges with the *continuous* variation of semantic entities (as is clear from the work of D. Scott), and a further connection emerges when typed λ-calculus is expanded into higher-order logic (with bounded quantification), revealing the strict link between continuity and constructivity requirements, as is manifest in topos theory.

Thus, even pursuing the computational version of the linguistic turn leads us out of classical set-theoretic semantics. More specifically, the twofold connection above imposes constraints on the underlying ontology and they

are independent from the point-like notion of space inherited by set theory. A situation of this kind occurs already with Grothendieck toposes, and a further departure from the usual semantics is obtained when λ-abstraction applies not only to terms but also to types, in a way that reduces the gap between the proofs of any proposition and its truth-value (a gap which still persists in toposes). Once again, even granted the adequacy of any one computational model of the mind, such a model would require a framework entirely different from that furnished by the usual pairing of first order logic and set-theoretic semantics, still (more or less tacitly) exploited in current discussions about the relationship between the computational architecture of mind and its intentional resources. A category-theoretic perspective on the linguistic and non-linguistic ingredients of real semantics points rather towards what I have called the "sheaf model of mind" (see Peruzzi 1993), thus supporting a modularity thesis that, through a reduction of the "modules" involved, applies to central processes of cognition as much as to input-systems.

So far, it might appear that the main advantage of a category-theoretic approach is that thought, at least in its computational aspects, finds a much more suitable logical and algebraic expression. But there is more than logical and algebraic structure in thought. If that were the whole story, to replace set theory with category theory would still be unsatisfactory (even apart from foundational questions). What is still lacking is the essential rôle of topology (not necessarily defined in terms of sets of points) and, epistemically speaking, of what can be briefly termed as "intuition of space". The point is that, in their very inception, categorical constructions are deeply related to the links between algebra and topology. This same point can be fruitfully pursued in approaching the study of the bases of meaning.

3 Dynamic Cohesiveness

On one side, the development of cognitive science, with the enormous increase of knowledge of the structures of mind, and specifically concerning the interface of language with the perceptual origins of categorization (in the cognitive sense), forces our philosophical doctrines to face the tribunal of experience (and experimentation). On the other side, the development of a new conception of language, logic and mathematics – that provided by category theory – allows us to lay the ground for overcoming the dogmas that see thought as essentially dependent on language and semantics as essentially dependent on set theory. What is suggested is *not* that, by means of more sophisticated mathematical tools, one can refine the (idealistic) understanding of an order of dependence: from language to categorization, from categorization to perception; nor that the order of dependence has simply to be inverted. For this project too would entail a seriously vitiated program, unless categorization were itself restricted to a basic level of complexity.

Such a level is in fact that of a particular collection of kinds, those

that are ontogenetically primary. Here, we face a fecund synthesis of empirical knowledge about the way ontogenesis of language is possible with theoretical knowledge about the way a proper representation of such genesis is possible.Within such a synthesis, Wittgenstein's claim, at 4.015, can find concrete justification, although the price to pay is the abandonment of linguistic universalism, with the ineffability of semantics involved in it. To this end, the modular view of mind and the relevant contribution of systems other than language in moulding intentionality are decisive. Moreover, the escape from ineffability is paralleled by a denial of the classical contraposition between semantic holism and atomism, as argued in Peruzzi 1994. At one extreme, there are pointlike objects, at the other, globally holistic objects; but only in between do we find cohesive objects, ones that conform to gestaltic structures and correspond to the actual functions exercised by symbol-makers such as we are.

If it is clear that the presence of *Gestalten* signals a primary rôle for topological aspects, it has to be emphasized that the way cohesive objects are related to pointlike and holistic ones can best be captured by means of category-theoretic constructions that make essential use of adjoints.[2]

To give a simple yet paradigmatic example, consider the category **TOP** of topological spaces and continuous functions. The forgetful functor from **TOP** to **SETS** that assigns each space X its set of points $\mathrm{pt}(X)$ has both a left and a right adjoint. The former assigns each set A the discrete space $D(A)$, that is, the most disconnected spacegenerated by A, the latter assigns A the co-discrete space $C(A)$, where the only opens are $\emptyset$ and A, hence the most connected space generated by A. There are obvious natural isomorphisms $Hom_{\mathbf{TOP}}(D(A), X) \cong Hom_{\mathbf{SETS}}(A, \mathrm{pt}(X))$ and $Hom_{\mathbf{SETS}}(\mathrm{pt}(X), A) \cong Hom_{\mathbf{TOP}}(X, C(A))$. Notice also that D has a left adjoint, that assigns a space the set of its "components". In between $D(A)$ and $C(A)$ we have the range of degrees of cohesiveness a generic space based on A can possess.

The situation becomes much more interesting when, following Lawvere 1989, **TOP** is replaced by a topos **M** of cohesive *Mengen* that are invariant under a monoid or group action, so that in general the points of a non-zero object X are no longer enough to distinguish any two morphisms from X having the same codomain, i.e. the common form of the axiom of extensionality that holds in **SETS** cannot be lifted to **M** (see Peruzzi 1989). In fact, kinds of macrophysical objects of reference are generated by "figures" that don't reduce to only one point.

All this suggests that semantic universals can be described not as a priori concepts, or conceptual relations, of a *static* character, but rather as uniform principles of conceptual *construction*, that underlie the categorization of experience, so that both global holism and atomism are excluded

[2]For a self-contained exposition of the rôle of adjoints and their applications in logic, see Peruzzi 1991.

from "real" semantics. The former entails the non-local dependence of what is to be constructed on any component of the "background": every concept is at the same time determined by the whole, and intrinsically unstable inasmuch as it is arrived at via a constructive process. The latter entails the arbitrary parsing of naked data into a hierarchy of types that are independent of any perceptual constraint. Non-trivial epistemological cohesiveness (what I elsewhere called "local holism") corresponds to a bounded form of impredicativity.

The emergence of cohesive objects as invariants with respect to well-defined types of actions points to an ontology of natural language in which spatial aspects, long investigated whithin algebraic and differential topology, have a central place. But once paths and singularities are recognised as natural ingredients of semantics, traditional theories of meaning and reference of purely logical character are seen to be flawed. They loose track of the phenomenological constitution of reference, and they fail to make contact with the mathematical framework which most fruitfully describes that constitution; for this very reason "formal semantics" has amounted, up to now, to a sort of "abstract nonsense", while the "abstract nonsense" alleged to be the defect of category theory allows us to recover exactly the depth of surface, and thus to be more faithful to the *gros grain* of the world, as represented in language.

The proper consideration of paths and singularities also imposes strict constraints on the semantic universals previously proposed by logicians, philosophers and linguists. Other constraints are due to the foundational rôle (in Lawvere's sense) played by adjoint functors.

From this perspective, common theories of mereology are also defective: insofar as semantic entities are the correlate of conceptual constructions, they are not to be viewed as sets, be they punctate or pointless. This remark would itself be pointless, were mereology intended as a theory of parts and wholes *independent* of the phenomenological aspects of semantics. But then its interest as a "formal ontology" would be reduced, and lack of consideration for the process aspects (the "maps") would mean it served to describe only a static world. Since the relationships between mereology and topology are of great importance, simple constructions such as the quotient space, the compactification, the connected sum of manifolds, etc., should serve as a reminder that topology is more than a taxonomy of different kinds of spaces (or different regions in space). It is also a theory of how different kinds of continuous variation affect the properties of any given space (and its regions) and how these kinds of variation are algebraically expressible. It is precisely because of this, that topology is relevant to semantics. So, rather than some weak form of topology, a richer theory has to be sought for, in which different notions of space can be compared.

Moreover, this perspective leads one to doubt the existence of "objects in general", as intended referents of a formal ontology, independent of any material domain (this doubt was already raised by Carl Stumpf). True, a

similar doubt can also apply to abstract sets and arbitrary categories: for this very reason the present approach sees the foundations of semantics and the foundations of mathematics as intimately linked. Typing is relevant here, yet it does not solve the philosophical problem concerning the way any notion of object is *possible*, as rooted in features (not point-like but still nomological) of the "common-sense" world. For instance, in natural language even the notion of "set" is expressed in a typed manner, one sensitive not only to the "homogeneity" of the entities collected, but also to the mereological modality of their grouping (team, sheaf, fleet, swarm, series, bunch, cluster, etc.). The claim is that set theory results from a (usually implicit) abstraction from the type and the modality whithin which entities are collected – modalities are not simply *order* types. That contemporary mathematics succeds in recovering such modalities (e. g. consider the topos $\mathbf{G\text{-}Sh}(X)$ of sheaves over a space X with a group G acting on them) is a fact of crucial philosophical importance. It also suggests that mathematics is directly involved with structures of the material world and does not need logic or language as intermediacy agencies – though, evidently, mathematics needs language and logic in order to be communicated and organised in a stable way.

A theory investigating the relationships of part and whole remains inadequate until the stability of the objects involved with respect to (actual or imagined) actions is taken into account. There have been attempts at axiomatising mereology in (classical) first order logic, but they either presuppose (classical) set theory or occur in a semantic vacuum. Phenomenological wholes are rich in structure, so they have rather to be investigated by starting with a category of G-spaces (e. g. the category of differentiable manifolds with the action of a Lie group), then passing to sheaves over such spaces. Likewise, the problem of constraining the formation of unions of parts (of a given whole) also has a direct solution: parts and their unions are different from arbitrary subsets, being closed with respect to G-action. (In order to be applied in cognitive science, such a model will have to be greatly refined, taking into account edges, junctures, etc.) Of course, the difficulty lies in dealing with singularities. The approach by Petitot (1994) is a step in this direction.

As this line of investigation is developed, it becomes apparent that cognition is *inseparable* from perception and sensory-motor abilities. Concepts result from schemes of objects and processes. In its generality, semantics corresponds to a variety of fibrations (each related to a given ontological region) amalgamated by coherence principles. For our present aims, it is sufficient to use the notion of discrete fibration; it is given by the functorial assignment of a family $\mathbf{X}$ of categories $\mathbf{X}_C$, indexed by the objects C of a base category $\mathbf{B}$, in such a way that for any map f from C' to C in $\mathbf{B}$ and for any X in $\mathbf{X}_C$, there exists exactly one object X' in $\mathbf{X}_{C'}$ and one f^* from $\mathbf{X}_{C'}$ to $\mathbf{X}_C$, such that $f*(X') = X$ – i.e. the assignment of f^* to f is contravariant. We can think of$\mathbf{B}$ as a category of concreta (perceptually salient, basic kinds of objects and processes) and of $\mathbf{X}$ as the abstracta, to which $\mathbf{B}$-

features are lifted. The origin of such a construction lies in homotopy theory for covering spaces, and one can already find it instantiated in the case of sheaves over a topological space: both the notion of *site* and the notion of *locale* represent categorical renderings of the rôle assigned here to the (base) space. But homology also suggests a direct presentation of extended bodies in terms of (basic) *figures*, acting as generators.

Since concreta are extended, **B** can be regarded as spatial in character, in accordance with one of the above notions of space. In this way, one can tackle the problem of change of base, from **B** to **B'**, finally dealing with the case in which the family of **B**,**B'**, ..., can be "pasted" together, forming a universal base space. However, in this paper, I shall not enter into the details of this view. It should already be apparent how great a philosophical significance categorical constructions possess for semantics. Let me only add that the existence of an adjunction $F \dashv G$, with $F : \mathbf{C} \to \mathbf{D}$ and $G : \mathbf{D} \to \mathbf{C}$, determines a monad T (in **C**), namely $T = GF$, for which algebras can be introduced, so that we can raise the problem of the initial algebra for T and thus face the fix point problem for computations relative to **C**. Since we started from adjoints related to topological notions, this paves the way to the recognition that, essentially, the manipulation of symbols, as algebraically described, is itself rooted in the structure of space.

4 Types, *Gestalten* and Schemes

Categorization is one of the fundamental cognitive activities. Though, in itself, it does not require language, the linguistic counterpart of categorization is *typing*. Now the human ability to identify new tokens as instances of a given type cannot be represented set-theoretically. If it could be so represented, it would reduce to a search and find strategy, thus presupposing the totality of instances of any type as given, which is almost never the case in ordinary experience. At the same time, new concepts of kind, and corresponding new types, can be created, dependent on types already formed, with a feedback on previously categorised tokens, so that what was taken as a – i.e. an instance of kind – K turns out to be a K'. The compenetration of constructivity and sensitivity to the ambient structure, as exemplified in the above process relating tokens and types, has begun to find mathematical formulation.

Various type theories have been elaborated in recent years, especially in connection with programming languages, and significantly, their semantics is categorial, rather than set-theoretic. For instance, starting from typed λ-calculus, which is interpreted in cartesian closed categories, we have arrived at polymorphic λ-calculus and other related systems, which in view of their intended categorial semantics have been exploited in the very design of programming languages (such is the case for ML). Second-order λ-calculus (J. I. Girard) and some versions of P. Martin-Löf's theories of dependent types have also received categorial models.

If we stick to toposes, that is cartesian closed categories with a sub-object-classifier, their higher-order "internal language" is an intuitionistic type theory, in which terms, identity and quantification are (strictly) typed. The semantic counterpart of types are the objects of such a category, just as the counterpart of terms are maps. Moreover, a particularly powerful typed system such as the theory of constructions of T. Coquand and G. Huet (which extends both second-order λ-calculus and a fragment of Martin-Löf type theory) has been semantically interpreted by means of suitable fibrations over a base category, one no longer supposed to be cartesian closed, but with cartesian closed fibers richer than poset-categories (see Hyland, Pitts 1989).

In either case (toposes and general fibrations), objects and processes are on the same footing and much more strictly linked than in previous set-theoretic semantics. This has direct philosophical consequences for the development of the notion of constructive truth, in view of the Curry-Howard correspondence between types and formulae. (Incidentally, this point is further positive evidence for the inseparability of philosophy of language and philosophy of mathematics, which was already urged above, through the suggestion that Abstracta should be regarded as fibered over Concreta.)

In many languages, concepts of object and concepts of process are respectively manifested by (count or mass) nouns and predicables (adjectival forms and verbs); their semantic counterparts are respectively kinds of object and kinds of state or action. Once we have sub-types, and types can be joined to form other types, we have a hierarchy of kinds. Moreover, kinds are organized into levels. Although the notion of "level" is used intuitively here, it can be rendered mathematically precise by following the standard procedure of quotienting a poset into a linear order of equivalence classes. The problem then lies in finding the suitable quotient that matches categorization levels, as described by psycholinguists in terms of perceptual saliency.

For any individual x there is exactly one basic level kind to which x "belongs"; at such a level the world is partitioned into disjoint equivalent "classes" that have no common subkinds. However, one could regard the basic level as privileged only perceptually, not cognitively (and logically). It seems to me that, in so doing, one would be forgetting that at the basic level we meet not just kinds of objects but also kinds of maps (relations and actions), and that these too result from a privileged stabilisation of dynamical *Ur-Gestalten* – it is exactly such stabilisation that first gets lexicalised. Moreover, as each basic level kind has a prototype, this applies also to maps, hence we have *dynamic* saliency: it is in virtue of the perceptual saliency of *go*, *open*, *put together*, etc., that their schemes are successively transposed to the whole of cognition (on this subject, a comparison with P. Kitcher's suggestions regarding a naturalistic semantics for set theory would be useful). It is clear, then, why category theory is required for dealing with a "base space" **B** of perceptually salient entities as well as with the variety of liftings

from this space that constitute human knowledge.

The main hypothesis advanced here is that ground types correspond to basic level kinds. So, the criterion for "belonging" to such kinds is ultimately topological, being traced back to the existence of gestaltic features that define prototypes and satisfy a minimum informational condition, see Rosch *et al.* 1976. Whereas, in general, subordinate kinds don't satisfy the minimum condition and super-ordinate kinds lack prototypes. Even in the latter case, however, we have schemes (think, for instance, of "quadruped", "plant", or "metal"), though with indeterminate parameters.

Thus, it is still profitable to investigate the adjunction between the category **I** of individuals (with localized actions) and the category **K** of kinds, in such a way that the induced equivalence between the subcategories $\mathbf{I}_0$ and $\mathbf{K}_0$ corresponds to the equivalence of generic individuals (with localized action-schemes) and basic level kinds (with basic kinds of action). Because, then, we can formalise the dialectical character of categorization: individuals trace the identity of kinds, while kinds make access to individuals possible. In consequence, both the principle that intension determines extension and its converse turn out to be hypostatic, one-sided versions of such a dialectics: the loop becomes stable precisely in the case of basic kinds and basic schemes.

However, as I have argued elsewhere, prototypes should not be confused with stereotypes (sets of beliefs). In this regard, the category-theoretic approach to semantics of natural language avoids the defects of previous attempts at connecting the notions of prototype and fuzziness. These defects are either inherent in fuzzy set theory, see Barr (1986), or the byproduct of interpreting the prototype/kind relation in terms of membership.

By the use of categorial methods in model theory, it can be shown that "geometric" theories (forming a particularly important class from the logical point of view) have "generic" models, through which any other model of such a theory factorises (in a canonical way). These models represent also a solution to the classical problem of generic objects (insofar as these are described by geometric formulae). For, the objects in the generic model for a geometric theory T satisfy all and only those properties that are true at every model of T. So, for example, such objects can function as "generic triangles" and the like. (Of course, terms for these objects are not in the same range of variables for particular instances of the kind "Triangle"; still they work as they should, through the action of suitable functors, which collectively codify the plasticity of each given notion.) Similarly, we can model the transfer of schemes for basic kinds to any other cognitive domain. All these schematic objects that are defined by a geometric theory can then be extracted from their associated generic models and treated in one and the same "ambient" setting (still avoiding collision of variables for generic and non-generic objects, because the internal language of categories is typed). In this ambient setting, the view proposed by Ellerman 1988 of predication as "participation" μ (for $\mu\varepsilon\tau\varepsilon\xi\iota\varsigma$) in a universal can be exploited, recovering membership (ϵ, for $\varepsilon\sigma\tau\iota$) as the static trace of participation.

The point is that basic level kinds are themselves the supervenient outcome of more fundamental topological building blocks, that have, in general, no independent presentation. I have termedthese building blocks "*Ur-Gestalten*", for they correspond to our intuitions of curvature, texture, singularity of a surface, hole, etc. Yet there is no texture, etc. that isn't the texture, etc., of some*thing*. All these aspects, through the information extracted out of motion, contact, reflection of light on physical bodies, etc., *constitute* reference. Algebraic aspects are mainly related to various kinds of action performed (or simply imagined as performed) on objects, firstly through one's own body.

These actions, in turn, contribute to stabilising reference, so that we reach an internalised image of both discrete kinds of continua and discrete kinds of continuous actions. Therefore, basic kinds are not primitive: they can and have to be subjected to further analysis in mathematical terms. The entire epistemic *Aufbau der Welt* is possible only because the natural world is sufficiently rich in information that we can "paste" together different sources of structure, and we can identify ourselves as single persons through this process. The quotientings of data we perform are not predetermined by physics – different species can well use quotients different from ours – but the range of possible quotients is constrained by the mathematical laws of nature.

Significantly, the types of quotients on which human cognition is grounded are few. Kinds not only form a hierarchy rooted in a basic level, characterized by perceptual saliency; they are also grouped along ontological/cognitive axes that give rise to the fundamental articulation of thought. Through the linguistic investigations of C. Fillmore, J. Gruber, L. Talmy, R. Jackendoff and G. Lakoff, a vast collection of data has revealed the pervasive presence of spatial structures in the semantics of natural language. Let me refer only to Jackendoff 1987 and Johnson 1987. The former lists the primitive "categories" and gives an account of the evidence supporting the central rôle of thematic relations of a spatial character; the latter provides a thorough, philosophically oriented, appreciation of image schemes.

From a categorial point of view, the data accumulated by linguists find their proper presentation in a rigorous theory. The main point to be emphasized is that not only certain kinds of object, but also certain kinds of action, are referentially privileged, and it is exactly the fundamental spatial schemes of the latter kind that are systematically transposed to any area of human experience. This transposition is categorially interpreted as lifting. Without the joint consideration of both aspects (the objectual and the processual), the status of image-schemes would remain static, passive and confined to the original context (the base space). Thus, the problem of how these schemes undergo variations, dislocations to other semantic domains, and finally abstract transposition in our very talk about symbols, would remain problematic and, in the absence of any constraint, unable to be given precise mathematical form.

But, we have seen that constraints on universals can be given such a mathematical form by means of a fibered semantics, where the base space is a category of basic level kinds, organized in conformity to gestaltic primitives and satisfying stability principles with respect to basic kinds of actions, principles topological in content. Furthermore, we have seen that the variety of cognitive constructions, partially, but significantly, revealed by analysis of language, is the result of lifting the patterns instantiated in the base space (of concreta) to the global space (of abstracta), giving rise even to "manipulation" of symbols. Finally, when this lifting involves an adjointness situation, it is also endowed with computational relevance, for a monad can be derived from it, as sketched above.

So, since the overall picture is governed by functorial principles, significant constraints are imposed on the way the process of abstraction can take place. At the same time, other constraints on concept formation (and transformation) derive from the gestaltic origin of properties characterising invariants within the base space. Accordingly, there is a twofold source of constraints, which contribute to determining the meaning of any sentence of natural language: we identify as semantic universals those schemes of object and action that are preserved through their lifting, and can thus be doubly globalised: horizontally (by change of base) and vertically (by change of lifting).

This all imposes severe bounds to the range of possible universals, so that such bounds can be conceived of as the "pairing constant" of mind and nature. Moreover, the constraints in question can also be connected with a new approach to the foundations of mathematics, from a perspective which takes account of the richness of basic structures pertaining to the "subject" of mathematics, as a system embedded in the natural world.

References

M. Barr 1986 "Fuzzy set theory and topos theory", *Canadian Mathematical Bulletin* 29, 500–508.

D. Ellerman 1988 "Category theory and concrete universals", *Erkentnnis* 28, 409–429.

J. Gray, A. Scedrov (eds.) 1989 *Categories in Computer Science and Logic*, Providence RI: American Mathematical Society.

M. Hyland, A. Pitts 1989 "The Theory of Constructions: Categorical Semantics and Topos-Theoretic Models", in J. Gray, A. Scedrov (eds.) 1989, 137–200.

R. Jackendoff 1987 *Consciousness and the Computational Mind* Cambridge MA: MIT Press.

M. Johnson 1987 *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*, Chicago: University of Chicago Press.

F. W. Lawvere 1989 "Qualitative Distinctions Between Some Toposes of Generalized Graphs", in J. Gray, A. Scedrov (eds.) 1989, 261–300.

J. Macnamara, G. Reyes (eds.) 1993 *Logical Foundations of Cognition*, Oxford: Oxford University Press.

J. Macnamara, G. Reyes, M. Reyes 1993 "Foundational Issues in the Learning of Proper Names, Count Nouns and Mass Nouns", in J. Macnamara, G. Reyes (eds.) 1993.

A. Peruzzi 1988, "Forms of Extensionality in Topos Theory", in C. Cellucci, G. Sambin (eds.), *Temi e Prospettive della Logica e della Filosofia della Scienza contemporanee* vol. I, Bologna: Clueb, 223–226.

A. Peruzzi 1991, "Categories and Logic", in G. Usberti (ed.), *Problemi Fondazionali nella Teoria del Significato*, Firenze: Olschki 137–211.

A. Peruzzi 1993, *From Kant to Entwined Naturalism*, Biblioteca del Dipartimento di Filosofia, Università di Firenze, Firenze: Olschki.

A. Peruzzi 1993a, "Prolegomena to a Theory of Kinds", in J. Macnamara, G. Reyes (eds). 1993.

A. Peruzzi 1994 "Holism: The Polarised Spectrum", *Grazer Philosophische Studien.*

J. Petitot 1994 "Phenomenology of Perception, Qualitative Physics, and Sheaf Mereology", this volume.

E. Rosch *et al.* 1976, "Basic objects in natural categories", *Cognitive Psychology* 8, 382–439.

On Some Implications from Linguistics for Theories of Mind

Robert D. Van Valin, Jr.

1 Introduction

The investigation of human language, be it by philosophers or linguists, has long been held to have important implications for theories of mind. For the past three decades, Chomsky has addressed what he calls "Plato's problem", that is, how can we know so much on the basis of so little experience? (see Chomsky 1986), and his view of the nature of linguistic structure has led him to propose that we know so much about our language because its essential features are given in advance in the form of autonomous modules in the mind. Moreover, within his conception of language, the major components are each encapsulated in autonomous modules, with the syntactic module being the central component connecting all of the others. This view entails some very strong claims about both the organization of our mental faculties and their content.

The purpose of this paper is to explore the primary argument that is given in support of Chomsky's solution to Plato's problem, namely the argument from the poverty of the stimulus, and the conception of language acquisition in which it is situated, namely the logical problem of language acquisition. The issue may be formulated as follows: given an account of adult grammatical competence (what Chomsky calls the "final state" of the organism), we may deduce the initial state of the language acquirer by factoring out what is supplied by experience. This may be represented graphically as in Table 1.

If there is some element of the final knowledge state which is not attributable to experience, then it must be part of the initial knowledge state or language acquisition device; this is the argument from the poverty of the stimulus. Since a child can learn any human language, the language

	Final knowledge state	(= Adult grammatical competence)
−	Input from experience	
=	Initial knowledge state	(= Language Acquisition Device)

Table 1

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 371-385

acquisition device is a theory of universal grammar.

The argument from the poverty of the stimulus is an argument to the effect that a particular rule, principle or constraint is "psychologically real", because, it argues, the rule, etc. is part of a speaker's innate grammar. There are two crucial presuppositions inherent in Table 1 which I wish to explore. First, the logical problem of language acquisition in Table 1 presupposes an accepted, widely agreed-upon account of the final knowledge state. This is important because the account of the linguistic phenomena in question determines both the nature of the cognitive constructs to be posited and the nature of the evidence that can bear on their potential learnability. The theory underlying the account specifies the nature of linguistic knowledge, and different theories make very different claims about how that knowledge is instantiated formally. Many theories assume a constituent-structure model for clause structure (i.e. traditional S → NP+VP tree structures), e.g. Chomsky's Government and Binding theory (Chomsky 1986) and Phrase Structure Grammar (Gazdar et al. 1985), some assume representations based on grammatical relations like 'subject' and 'direct object', (e.g. Lexical-Functional Grammar (Bresnan 1982) and Relational Grammar (Perlmutter 1980)), and some assume a more semantically-based representation, e.g. Role and Reference Grammar (Van Valin 1993a). These contrasts in turn define the type of evidence that could be relevant to the child for acquiring some aspect of grammar. If some phenomenon is analyzed strictly in terms of constituent-structure configurations, then evidence regarding constituent structure will be relevant to its acquisition; if, on the other hand, the phenomena are analyzed strictly in terms of grammatical relations, then evidence relating to grammatical relations will be crucial. The main point is that given competing characterizations of grammatical competence, the argument from the poverty of the stimulus cannot decide between them; rather, it can only provide an argument as to the psychological status of some construct within a particular scheme.

The second presupposition in Table 1 concerns the nature of the learning theory that is assumed in the "input from experience". A critical assumption is the no-negative-evidence hypothesis, i.e. the claim that children are not exposed to any ill-formed strings labelled as such and are forced to generalize only from positive tokens. That is, it is assumed that adults don't produce ill-formed utterances and then tell the child that they are ill-formed, and they don't correct the child's ungrammatical utterances. This is a strong restriction on the language-learning environment. The actual learning theory that is tacitly assumed is quasi-behaviorist; that is, learning is the result of induction from repeated exposure to phenomena. For example, children acquiring English learn the argument structures of English verbs by hearing verbs in sentences together with nouns functioning as subject, direct object, and indirect object (however these may be conceived in a given theory) and come to associate the patterns of nominal complements with particular verbs. It is generally assumed that whatever

is actually learned in acquisition is learned through direct exposure to the relevant tokens. These assumptions lead inevitably to the characterization of the language-learning environment as "severely impoverished", and this in turn leads to the standard conclusion that the initial knowledge state must be very rich and complex.

2 Restrictions on Question Formation

I would now like to explore a detailed example of the argument from the poverty of the stimulus and the implications of competing theoretical analyses for it. These phenomena are widely considered to be one of the most compelling arguments for Chomsky's position: namely, the constraints on the formation of WH-questions (e.g. *What did Fred see?*) and related constructions known as *island constraints*. I will first present a Chomskyan analysis of them, looking at data from two languages. I will then sketch an alternative analysis and examine its implications for language acquisition. I will then return to the logical problem of language acquisition in (1) and explore the implications of each analysis for it. In particular, I will argue that the alternative analysis has very different implications for acquisition and modularity from the Chomskyan account.

The restrictions in question are well established from Ross' seminal work in this area (Ross 1967). Due to space limitations, I will deal only with the phenomena subsumed by Ross under the "Complex Noun Phrase Constraint". The basic facts from English are presented in Table 2; '*' indicates an ungrammatical sentence.[1]

a. Max believes [$_S$ that Susan lost her wallet].
b. What does Max believe [$_S$ that Susan lost ___]?
c. Max believes [$_{NP}$ the claim [$_S$ that Susan lost her wallet]].
d. *What does Max believe [$_{NP}$ the claim [$_S$ that Susan lost ___]]?
e. Fred talked to [$_{NP}$ the man [$_S$ who bought the house down the street]].
f. *What did Fred talk to [$_{NP}$ the man [$_S$ who bought ___]]?

Table 2

In (b) the direct object of the verb in the complement clause appears as a WH-word at the beginning of the sentence; the result is a grammatical question. In (d), on the other hand, the WH-word is the direct object of the verb in a sentence which is a complement to a nominal head (*the claim*), and the result in this instance is an ungrammatical question, despite the alleged near synonymy of (a) and (c). This semantic similarity, it is argued, shows that the ill-formedness of (d) cannot be attributed to semantic or other non-structural factors. Ross proposed the Complex Noun Phrase Constraint,

[1]The square brackets indicate a constituent boundary: with subscript 'S' they signal the boundary of an embedded clause, with subscript 'NP' they signal the boundary of a noun phrase.

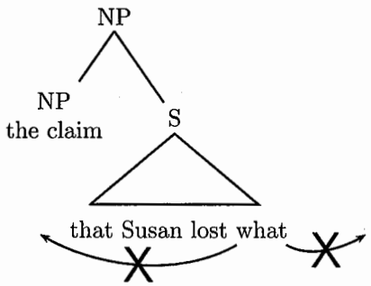


Figure 1: Complex Noun Phrase Constraint (Ross 1967)

which states, roughly, that a WH-word like *what* or *who* cannot be moved out of a sentence which is part of a noun phrase; the relevant structural configuration is represented in Figure 1.

Since Chomsky (1973), the Complex Noun Phrase Constraint has been subsumed under a more general principle called *subjacency*, which has gone through a number of reformulations. In essence, it states that no element can move across more than two NP and S nodes, which are called "bounding nodes". Given this characterization of the phenomena, the existence of subjacency effects in a language has been taken as a diagnostic for WH-movement in a language. That is, since subjacency violations are caused by movement across bounding nodes, the existence of subjacency effects in a language is evidence for the existence of a movement rule in the grammar of the language.

It is difficult to imagine how an abstract restriction like subjacency could be learned. First, it is argued to be a purely structural restriction with no semantic or other non-structural basis, as noted above. Second, it is an instance of a *systematic non-occurrence* of a logically possible phenomenon rather than a systematic occurrence of some phenomenon, and therefore this otherwise allegedly unmotivated restriction could not be induced by the child from the data to which he or she is exposed. Moreover, adults do not produce utterances like (d,f) of Table 2 and then tell children that they are ill-formed and that one doesn't produce sentences like them. The apparent impossibility of learning subjacency, combined with the virtual universality of the restrictions, seems to point unequivocally to the conclusion in terms of the schema in Table 1 that subjacency must be part of the language acquisition device, following the argument from the poverty of the stimulus. In addition, it is difficult to imagine how such a restriction could be applicable to other areas of cognition, and consequently this principle appears to be uniquely linguistic and not derivative of a general cognitive principle of any kind. This supports Chomsky's modular view of the mind.

a. *Šųkmánitu-thąka ob wachí Nąpé nąžį-wį*
coyote-big with dance hand stand-FEM
čąxíye yelo.
love DEC(Male spkr.)
"Dances-with-wolves loves Stands-with-a-fist."

b. *Šųkmánituthąka ob wachí Nąpé nąžį́wį čąxíyą he?*
Dances-with wolves Stands-with-a-fist loves Q
"Does Dances-with-wolves love Stands-with-a-fist?"

c. *Šųkmánituthąka ob wachí tuwá čąxíye yele.*
Dances-with-wolves who/someone love DEC(Fem. spkr.)
"Dances-with-wolves loves someone."

d. *Šųkmánituthąka ob wachí tuwá čąxíyą he?*
Dances-with-wolves who/someone love Q
"Who does Dances-with-wolves love?", or "Does Dances-with-wolves love someone?"

e. *Tuwá Šųkmánituthąka ob wachí čąxíyą he?*
who/someone Dances-with-wolves love Q
"Who loves Dances-with-wolves?", or "Does someone love Dances-with-wolves?"
"*Who does Dances-with-wolves love?"

Table 3

This conclusion apparently receives further support when we look at languages of a different structural type. In Lakhota (Teton Dakota, a Siouan language of North America) questions are not formed syntactically; that is, there is no subject-auxiliary inversion or movement of WH-words to sentence-initial position, as in English. The basic Lakhota facts are illustrated in Table 3.

Basic word order is Subject-Object-Verb, as illustrated in (a). In order to form a yes-no question, the question particle he is added at the end of the sentence, as in (b); no other change is made in the structure of the sentence. WH-words double as indefinite-specific pronouns; when one occurs in a sentence without a question particle, as in (c), it means 'someone' (*tuwá*) rather than 'who'. When there is a WH-word and a question particle, as in (d), the result is ambiguous: if the WH-word is the focus of the question, then the sentence is interpreted as a WH-question, whereas if the focus falls elsewhere in the sentence, then the sentence is interpreted as a yes-no question with an indefinite-specific pronoun like 'someone' or 'something'.[2] An important feature of Lakhota WH-questions is that the question word does not move to initial position but rather remains *in situ*. This can be

[2]Ultimately, the determination of which NP is in focus is contextually based. It appears that the primary signal of focus is prosodic, as in many other languages.

a. *Wičháša ki* [$_{NP}$[$_{S}$ *šų́ka wą igmú óta wičhayáxtake*] *ki le*]
man the [$_{NP}$[$_{S}$ dog a cat many bite] the this]
wąyą́ke.
saw
"The man saw the dog which bit many cats."

b. *Wičháša ki* [$_{NP}$[$_{S}$*šų́ka wą táku yaxtáke*] *ki le*]
man the [$_{NP}$[$_{S}$dog a *what/something bite] the this]
wąyą́ka he?
saw Q
"*What did the man see the dog which bit ____?"
"Did the man see the dog which bit something?"

Table 4

seen clearly in (d,e); in (d) *tuwá* does not move from the preverbal object position, while in (e) the sentence-initial *tuwá* can only be interpreted as the subject, not as the object.

Since there is no movement of WH-words in Lakhota, it might be expected that the language would not show subjacency effects, if in fact these effects are caused by the movement of elements across the specified structural configurations as represented in Figure 1. Yet this is not the case, as the sentences in Table 4 show.

In a definite restrictive relative clause like Table 4 (b)[3] the object NP in (a), *igmú óta* 'many cats', is replaced by *táku* 'what' or 'something', and the question particle *he* is added at the end of the sentence to make the utterance a question. The result is a well-formed question, but it can *only* be interpreted as a yes-no question; it cannot be interpreted as a WH-question, as the translations indicate. The impossibility of the starred reading is a subjacency effect, because it means that a word like *táku* cannot be interpreted as a question word when it is an argument of a verb in a relative clause; thus, to generalize across both English and Lakhota, it is impossible to form a WH-question when the potential question word functions as an element of a definite restrictive relative clause. This shows that subjacency operates in the grammar of languages in which WH-words do not move to the beginning of the sentence in a question.

At first glance, this would seem to be a serious problem for the theory: how can there be subjacency effects if the WH-word does not move across the specified structural configurations? The answer adopted by Chomsky, originally proposed in Huang (1981), is that there *is* movement in languages like Lakhota, but only at an abstract, non-overt level; that is, the movement

[3]Lakhota relative clauses have no external head noun, unlike their English counterparts; the NP which is interpreted as the head must be indefinite within the embedded clause, its true definiteness status being indicated by the article + demonstrative combination at the end of the whole construction.

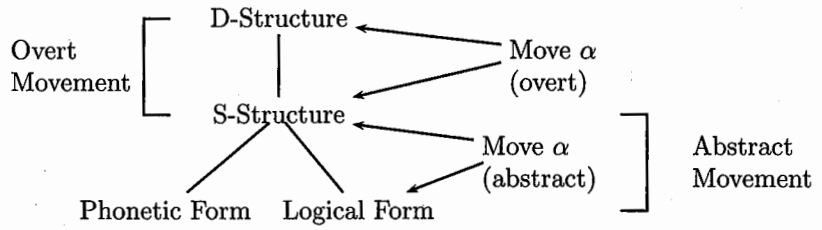


Figure 2: Application of Movement Rules in Chomskyan Theory

rule (Move α) applies not between D-structure (formerly "deep structure") and S-structure (formerly "surface structure"), as in English, but between S-structure and the abstract level of L[ogical] F[orm] (the level of semantic interpretation), as illustrated in Figure 2.

Recall that subjacency is a diagnostic for movement, and the existence of subjacency effects in a language is taken as evidence that there must be movement rules applying in the grammar. Thus in Lakhota subjacency applies only to abstract movement, while in English it applies to overt movement (cf. Chomsky 1986).[4]

This analysis of Table 4(b) has profound implications for the logical problem of language acquisition: it is simply not possible for a child to learn a constraint on movement in a language in which there is no overt evidence of movement in the first place. Hence the existence of subjacency effects in a language like Lakhota seems to provide very strong support for the conclusion that subjacency must be a principle of the language acquisition device, and it is a textbook case of the argument from the poverty of the stimulus.

3 An Alternative Account

A crucial part of the logic underlying the analysis of extraction restrictions in a language like Lakhota is the assumption that the existence of subjacency effects in a language is a diagnostic for movement. The reasoning goes as follows: the theory states that subjacency effects are caused by the movement of elements across a proscribed number of bounding nodes in languages like English, and therefore if these effects are found in a language, then there must be movement rules in the grammar. This account is clearly derived from the analysis of languages like English with overt movement in

[4]It must be emphasized that this situation is not unique to Lakhota. The statistically most frequent word order type in human language is Subject-Object-Verb; in virtually all of the languages of this type there is no obligatory WH-word movement, and in most it is not allowed under any circumstances, as in Lakhota. Moreover, there are Subject-Verb-Object languages, e.g. Mandarin and the other varieties of Chinese, which lack WH-movement. Thus, the method for forming WH-questions in Lakhota is more common in the world's languages than the method employed in English.

the grammar and extended to languages without overt movement. Let us look at this issue from a slightly different perspective. Both English and Lakhota show subjacency effects; English has overt movement of question words, while Lakhota does not. One could conclude, then, that movement is in fact *irrelevant* to the phenomena in question and that the source of the restrictions lies elsewhere. What I wish to explore now is this alternative interpretation and what the source of the restriction could be, if we ignore movement rules. It is necessary to look at Lakhota again, since it lacks overt movement rules. It is then necessary to see if the analysis can be extended naturally to languages with overt movement.

The place to begin the investigation is yes-no questions, a phenomenon which has not been thought to be relevant to the issue of extraction restrictions, since it involves no movement of a question word or the like. Are there constraints on the interpretation of yes-no questions comparable to the one exemplified in Table 4(b)? In Table 5, the possibilities for interpreting yes-no questions containing complex embedded structures is illustrated.

The issue in these examples is what can be in scope of the interrogative illocutionary force operator he in the sentence. The part of the sentence in its scope is called the potential focus domain of the question operator. It was noted with respect to Table 3 (d) that different elements could be in its scope in a simple sentence, i.e. either of the NPs or the verb. Thus the entire simple sentence is within the potential focus domain in Lakhota.[5] In the sentences in Table 5, the distribution of the question focus in complex sentences is shown by means of determining what a potentially felicitous answer to the question could be.[6] In (a) the embedded clause is an object complement, and as the potentially appropriate response shows, it is possible for the embedded clause to be in the scope of *he*, since it is felicitous to deny the subject NP of the complement in the response. In (b) the embedded clause is a relative clause, and here the situation is somewhat more complicated. It is possible for all of the main clause elements to be in its scope, including the NP interpreted as being modified by the relative clause, but as the last response shows, it is not possible for an element within the relative clause to be in its scope. The last example, (c), contains an adverbial subordinate clause, and as the range of potential responses indicates, the constituents of the subordinate clause cannot be within the scope of *he*. Thus while any major element of the main clause can be within the scope of the illocutionary force operator, a relative clause or adverbial subordinate clause cannot be. This means that in some constructions part of the sentence is outside the potential focus domain. This is summarized in Table 6.

It is clear from Table 5 (b) that it is impossible to form a yes-no question in Lakhota in which the focus of the question is a non-head element

[5]This is not true in all languages; see Van Valin 1993a: §2.4 for detailed discussion.

[6]There is another, more technical way of determining the distribution of focus in complex sentences in Lakhota; see Van Valin 1993a: §7.3.1 for detailed presentation and exemplification.

a. *[s Hokšíla etą́ it thaló ki manúpi] iyúkča he?*
boy some meat the steal think Q
"Does he think some boys stole the meat?"
___ *Hiyá, wičhį́čala eyá.*
no girl some
"No, some girls."

b. *Wičháša ki [NP[s šų́ka wą igmú eyá wičháyaxtake] ki le]*
man the dog a cat some bite the this
wąyą́ka he?
see Q
"Did the man see the dog which bit some cats?"
___ *Hiyá, wąyą́kešni.*
no, see.NEG
"No, he didn't see it."
___*Hiyá, wį́yą ki (wąyąke).*
woman the (see)
"No, the woman (saw it)."
___*Hiyá, mathó wą (wąyąke).*
bear a (see)
"No, (he saw) a bear."
___**Hiyá, magá eyá (wičháyaxtake).*
duck some (bite)
*"No, (it bit) some ducks."

c. *[s Wičháša ki wóte] ečhúhą, tha-wį́čų ki mní ikíču he?*
man the eat while his-wife the water bring.for Q
"While the man was eating, did his wife bring him water?"
___*Hiyá, Fred (mní ikíču/*wóte).*
'No, Fred (brought it to him)', or 'No, she brought it to Fred')
*'No, Fred was eating.'

Table 5

a. **[Hokšíla etą́ thalo ki manúpi] iyúkča** he?
= Table 5 (a)

b. **Wičháša ki [[šų́ka wą** igmú eyá wičháyaxtake] **ki le] wąyą́ka** he?
= Table 5 (b)

c. [Wičháša ki wóte] **ečhúhą, tha-wíču ki mní ikíču** he?
= Table 5 (c)[7]

Table 6: Summary of Potential Scope of *he*: potential focus domain in boldface.

a. [$_S$ *Tuwá thaló ki manú*] *iyúkča he?* (cf. Table 5 a)
who meat the steal think Q
"Who does he think stole the meat?",
or "Does he think someone stole the meat?"

b. [$_S$ *Wičháša ki táku yúte*] *ečhúhą,*
man the *what/something eat while
tha-wíču ki mní ikíču he?
his-wife the water bring.for Q
"While the man was eating something, did his wife bring him water?"
"*What did his wife bring him water, while the man was eating ___?"
(cf. Table 5 c)

Table 7

in the relative clause, and this is exactly parallel to the situation found in (4b), in which it was impossible to interpret *táku* 'what/something' as a question word when it was inside a definite restrictive relative clause. It appears, then, that yes-no and WH-questions are subject to the same restriction in terms of the potential scope of *he*. This restriction is formulated in the

General Restriction on Question Formation: The element questioned (the focus NP in a simple, direct yes-no question, or the WH-word in a simple, direct WH-question) must be in a clause within the potential focus domain of the question operator of the utterance.

If the element is in a clause within the potential focus domain, then the focus can fall on it, otherwise, not. A definite restrictive relative clause is outside the potential focus domain, and therefore it is outside the scope of *he*.[8] This explains the possible interpretations of the questions in Table 4 (b) and Table 5 (b). It also predicts that a WH-word/indefinite pronoun like *táku* could be interpreted as a question word in a complement clause but not in an adverbial subordinate clause, and this is correct, as the sentences in Table 7 show. *Tuwá* in Table 7 (a) can be interpreted as either 'who' or 'someone', depending on context, whereas *táku* in Table 7 (b) can only be construed as 'something', following the General Restriction on question

[8] An absolutely essential component of any explanatory account of these restrictions is an independently motivated determination of the potential focus domain in a complex sentence. See Van Valin 1993a: §§6.6, 7.3.1 for an independently motivated account of the potential focus domain in complex sentences within Role and Reference Grammar based on the interaction of clause structure, lexical semantics, and pragmatic functions. It is on this point that other functional accounts have foundered; for example, Erteschik-Shir & Lappin (1979) argue that extraction is only possible out of dominant constituents, but they provide no independent explanation for the distribution of dominant constituents, thereby severely limiting the explanatory potential of their account.

a. After you left the party, did you take Mary to the movies?
b. Yes.
No. (= didn't take Mary, ≠ didn't leave the party)
No, Bill did. (= Bill took Mary, ≠ Bill left the party)
No, Susan.
No, the park. (= went to the park, ≠ after you left the park)

Table 8

a. Did Max return the papers which the secretary photocopied
to the lawyer?
b. Yes.
No. (= Max didn't return the papers,
≠ the secretary didn't photocopy)
No, Bill did. (= returned the papers, ≠ photocopied the papers)
No, the envelopes.
No, the IRS agent.(= to the IRS agent,
≠ which the IRS agent photocopied)

Table 9

formation. Thus the General Restriction provides the basis for an account of extraction restrictions in Lakhota which makes no reference to any kind of syntactic movement, either overt or covert.

Can this account be extended to languages like English which have WH-movement and no overt question operator akin to *he*? The first step in answering this question is to recognize that all languages have grammatical or intonational means for indicating the illocutionary force of an utterance, and therefore it is appropriate to posit an illocutionary force operator for a language like English, even though it is not an overt morpheme as in Lakhota. It is also necessary to recognize that not every part of a sentence can be questioned, asserted or denied, just as in Lakhota; in other words, parts of an English complex sentence may be within the scope of the illocutionary force operator and other parts may not be. This can be seen clearly by noting that the possible interpretations of the English translations of the questions in Table 5 seem to be subject to the same restrictions as their Lakhota counterparts. This can also be seen in the examples in Tables 8 and 9. In these two examples involving an adverbial subordinate clause and a relative clause, the range of possible felicitous responses is restricted in the same way as in the Lakhota yes-no questions. Thus it is clear that there are parts of each sentence which are not in the potential focus domain. The General Restriction predicts that WH-question extraction should be impossible out of these structures, and this is indeed the case, as one can easily verify by

looking at the English translations of Table 4(b) and Table 7(b).

The General Restriction was formulated for Lakhota and refers to the position of the WH-word in a clause within the potential focus domain. For a language like English, however, this wording will not do, because the WH-word uniformly appears in sentence-initial position in a simple, direct WH-question, and therefore it is not the position of the WH-word which is relevant to distinguishing (2b) and (2d). Rather, it is the location of the gap left by the displaced WH-word that is crucial. Hence the restriction must be reformulated as the

General Restriction on Question Formation (revised): The element questioned (the focus NP in a simple, direct yes-no question, or the WH-word or the gap left by a displaced WH-word in a simple, direct WH-question[9]) must be in a clause within the potential focus domain of the illocutionary force operator of the utterance.

Thus, we arrive at a tentative analysis of these extraction restrictions which is as applicable to languages with overt WH-movement as to those without it.[10]

4 Implications for Acquisition

This account of extraction restrictions has interesting implications for language acquisition. In Van Valin (1986, 1993a) it is argued that the revised General Restriction is ultimately derivable from Grice's Cooperative Principle and the Maxim of Quantity. This Gricean foundation is very important, since these principles are considered to be general principles of rational behavior and therefore are not strictly linguistic in nature (see Kasher 1976). In terms of the phenomena under discussion, it has never been claimed that constraints on the interpretation of yes-no questions are innate or are even part of grammatical competence; they could be part of what Chomsky calls 'pragmatic competence', which he characterizes as follows:

> [Pragmatic competence] may include what Paul Grice has called a "logic of conversation". We might say that pragmatic competence places language in the institutional setting of its use, relating intentions and purposes to the linguistic means at hand. (Chomsky 1980:224-5)

If the constraints on yes-no questions are not innate, then where do they come from? There would appear to be abundant evidence relating to them available to the child through everyday conversation in which children

[9]Note that the qualification "simple, direct WH-question" eliminates echo questions, since they are not subject to this restriction (nor to subjacency). For a detailed presentation of the application of this analysis to English, see Van Valin (1993a,b).

[10]The technical Role and Reference Grammar analysis underlying this informal account is presented in Van Valin 1993a,b.

a.	Caregiver:	What did you eat? Eggs and ...	
	Child:		Mbacon.
b.	Caregiver:	Oh, that's a ...	
	Child:		Aleph.
	Caregiver:	That's a aleph.	

Table 10

a. What are you cookin' on a hot ___? [Answer: "Stove"]
b. What are we gonna go at [to] Auntie and ___?
c. What are we gonna look for some ___ with Johnnie?

Table 11

ask and answer questions constantly with peers and caregivers. It seems entirely reasonable that pragmatic knowledge of this kind would arise through conversational interactions. But how does this relate to constraints on WH-questions? As we have seen, yes-no and WH-questions are subject basically to the same constraint, and the hypothesis is that children learn the conditions on yes-no questions and extend them to the corresponding type of WH-question. WH-questions appear after yes-no questions, and WH-questions out of complex sentences would be last type of question to develop, given its complexity.

Is there any kind of evidence that this transfer of restrictions could occur? Wilson & Peters (1988) document an interesting set of deviant WH-questions produced by a three-year-old. While the child was involved in typical question-answer interactions and could produce 'normal' yes-no and WH-questions, he also learned the special question-answer routine exemplified in Table 10. In this routine, the father, the primary caregiver, would ask the child a question by leaving a gap in a statement, which the child would fill in. The child, having learned the pragmatic restrictions on the caregiver's questions, created analogous WH-questions in which the question word filled a gap in the same positions left vacant in routines like those in Table 10. The result is WH-questions like those in Table 11.

Wilson & Peters argue explicitly that the type of routine as in Table 10 is the source of the questions in Table 11. The child learned the 'question rule' from the game in Table 10, and when he began to make WH-questions with displacement of the WH-word, he produced the questions in Table 11. It appears, then, that the kind of transfer of restrictions posited above occurred in this instance.

5 Conclusion

If the analysis proposed here (or something like it) is correct, then there *is* evidence regarding the constraints on WH-question formation available to the child in the language-learning environment. Hence in terms of Table 1 one could conclude that these constraints are NOT part of the initial knowledge state of the child. This conclusion highlights the crucial importance of the theory which characterizes the final knowledge state that is assumed in the schema in Table 1. The analysis sketched in this paper potentially yields a diametrically opposed result to the conclusion drawn from a Chomsky-type analysis. This point does not depend upon the ultimate correctness of the proposed alternative account. This discussion involves a rather striking contrast between the accounts, but the same point could also be made regarding competing accounts of any other grammatical phenomenon. Therefore the conclusion drawn by the argument from the poverty of the stimulus based on the Chomskyan account cannot be considered valid, given the different conclusion that can be drawn from a competing analysis. Thus as long as there are competing accounts of the phenomena in question, the argument from the poverty of the stimulus can tell us little about the language acquisition device.

Moreover, the alternative analysis also highlights the second presupposition in Table 1 regarding the nature of the evidence available to the child. The theory underlying the account of the phenomenon in question defines the nature of the possible evidence relevant to its acquisition. A Chomskyan analysis by its very nature precludes the possibility of there being any evidence regarding subjacency available to the child, especially in a language like Lakhota. The alternative analysis, on the other hand, makes a very specific claim about what the evidence could be, and it suggests that available evidence relevant to one aspect of grammar (namely, yes-no questions) may be extended to inform the acquisition of a different, albeit related part of the grammar (WH-questions). In addition, this claim can be tested empirically through the study of the acquisition of yes-no and WH-questions in English, Lakhota, and other languages. This account also suggests that information used in the formulation of a constraint or principle by the child could have a very indirect source. In this case, evidence regarding restrictions on yes-no questions is applied to a distinct but related grammatical phenomenon, WH-questions. Thus, these indirect sources of information about grammar make the language-learning environment richer than is standardly supposed. Furthermore, the Gricean nature of an important constraint like the revised General Restriction has significant implications for the question of modularity, as argued in Van Valin (1986, 1991). Alternative analyses in which semantics and pragmatics play central roles will inevitably have very different implications for the organization and content of our mental faculties from the autonomous syntax analyses proposed by Chomsky.

In conclusion, the existence of competing analyses of a particular grammatical phenomenon renders standard argument-from-the-poverty-of-the-stimulus conclusions regarding the psychological reality of linguistic constructs highly questionable in the absence of any experimental, observational or other empirical corroboration. Dan Slobin refers to the argument from the poverty of the stimulus as the "argument from the poverty of the imagination", i.e. "I can't imagine how anyone could learn this, so it must be innate". Dawkins (1986) refers to it as the "argument from personal incredulity". As there will always be competing analyses in different theories, the argument from the poverty of the stimulus by itself cannot be taken seriously as an argument regarding the initial state of the language learner and human cognitive organization.

References

Bresnan, J. 1982 "Control and Complementation", in J. Bresnan (ed.), *The Mental Representation of Grammatical Relations*, Cambridge MA: MIT Press, pp. 282–390.

Chomsky, N. 1973 "Conditions on Transformations", in N. Chomsky, *Essays on Form and Interpretation*, New York: North-Holland 1977, pp. 81–160.

Chomsky, N. 1980 *Rules and Representations*, New York: Columbia University Press.

Chomsky, N. 1986 *Knowledge of Language*, New York: Praeger.

Dawkins, R. 1986 *The Blind Watchmaker*, New York: Norton.

Erteschik-Shir, N. and Lappin, S. 1979 "Dominance and the Functional Explanation of Island Phenomena", *Theoretical Linguistics* 6, 41–85.

Gazdar, G., E. Klein, G. Pullum and Sag, I. 1985 *Generalized Phrase Structure Grammar*, Cambridge MA: Harvard University Press.

Huang, C.T.J. 1981 "Move WH in a Language without WH Movement", *The Linguistic Review* 1, 369–416.

Kasher, A. 1976 "Conversational Maxims and Rationality", in A. Kasher, ed., *Language in Focus: Foundations, Methods and Systems. Essays in Honor of Yehushua Bar-Hillel*, Dordrecht: Reidel, pp. 197–216.

Perlmutter, D. M. 1980 "Relational Grammar", in E. Moravcsik and J. Wirth (eds.), *Current Approaches to Syntax* (*Syntax and Semantics* 13), New York: Academic Press, pp. 195–229.

Ross, J.R. 1967 *Constraints on Variables in Syntax*, MIT dissertation.

Van Valin, R.D., Jr. 1986 "Pragmatics, Island Phenomena, and Linguistic Competence" *CLS* 22/2: *Papers from the Parasession on Pragmatics and Grammatical Theory*, pp. 223–233.

Van Valin, R.D., Jr. 1991 "Functionalist Linguistic Theory and Language Acquisition", *First Language* 11, 7–40.

Van Valin, R.D., Jr. 1993a. "A Synopsis of Role and Reference Grammar", in R. Van Valin (ed.), *Advances in Role and Reference Grammar*, Amsterdam: John Benjamins, 1-164.

Van Valin, R.D., Jr. 1993b. "Towards a Functionalist Account of so-called 'Extraction Constraints", in J. van der Auwera, Devriendt, B. and Goossens, L (eds.), *Complex Structures: A Functionalist Perspective*, Berlin: Mouton de Gruyter (in press).

Wilson, B. & Peters, A. 1988. "What are you cookin' on a hot?: A Three-Year-Old Blind Child's Violation of Universal Constraints on Constituent Movement", *Language* 64, 249–273.

Phenomenology of Perception, Qualitative Physics and Sheaf Mereology

Jean Petitot

1 Introduction

One of the main consequences of cognitive sciences for philosophy concerns *the naturalization of eidetic descriptions*, e.g. eidetic phenomenological descriptions or formal logical ones. "Naturalization" means here experimental devices, physical explanations, mathematical modeling (which is in general completely different from logical formalization) and computational simulation.

This raises immediately a question: what can be the link between natural explanations and phenomenological or logical descriptions of the same phenomena? To investigate this point we need some mediating device which can operate at the same time at two completely different levels:

(i) that of physical explanations, and

(ii) that of eidetic and formal descriptions.

Mathematical modeling provides such a mediation.

My purpose is to give an example of the link between physics and logic which can be worked out using mathematical modeling in cognitive sciences. The example will be very simple but also very characteristic. It concerns the concept of *form*. As I want to clarify some difficult philosophical points I will take the *simplest* case – too simple of course – that of two-dimensional static forms.[1]

In the first part of my paper I will show briefly that there exists a remarkable convergence between:

(i) the phenomenological description of forms worked out by Husserl in the third *Logical Investigation*;

(ii) the topologico-geometrical description proposed by Thom in the late sixties;

(iii) the physical (in the sense of objective, external, in the outside world) explanation of forms in morphodynamics and qualitative physics;

[1] For a mathematical investigation of three-dimensional dynamical forms in computational vision see my paper "Le Physique, le Morphologique, le Symbolique. Remarques sur la vision" (Petitot 1990).

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 387-408

(iv) the physical (in the sense of cognitive, internal, in the brain) explanation of forms in computational vision.

This convergence will lead us to what can be called a morphological geometry which shares the twofold status of a descriptive eidetics (in Husserl's sense) and of a mathematical modeling of physical (external and internal) explanations.

In the second part of my paper I will show that this morphological eidetics yields a mereology and a logic which can help to solve some traditional philosophical problems. We will see that morphological eidetics is naturally related to the geometrical concept of a *sheaf* and we will use the essential and deep link established by Lawvere and Tierney between sheaf theory and (intuitionistic) logic by means of the categorical concept of a *topos*.[2]

2 The Eidetic Nucleus of the Concept of Form

Let us start with the phenomenological pure eidetic description given by Husserl in the third *Logical Investigation*. Husserl begins with the difference between "*abstrakten*" and "*konkreten Inhalten*". He identifies it with the Stumpfian opposition between dependent (*unselbständigen*) and independent (*selbständigen*) contents.

It is only in the second chapter "*Gedanken zu einer Theorie der reinen Formen von Ganzen und Teilen*" that Husserl develops an axiomatics of whole/part relations. In the first chapter "*Der Unterschied der selbständigen und unselbständigen Gegenstände*", he in fact develops a "material" analysis of empirical morphologies.

The central problem analyzed by Husserl is that of the *unilateral* dependence between qualitative moments (e.g. colour) and spatial extension (*Ausdehnung*). According to him, qualities are abstract essences (species) and categorized manifolds. He thought of the "quality → extension" dependence as an eidetic law binding generic abstracta or types.

> "The dependence [*Abhängigkeit*] of the immediate moments [*der unmittelbaren Momente*] concerns a certain relation conforming to a law existing between them, a relation which is determined only by the immediately super-ordered abstracta of these moments" (Husserl 1900–1, p. 233).

There is a functional dependence (*funktionelle Abhängigkeit*) between the immediate moments of quality and extension: it associates to every point x of the extension W the value $q(x)$ of the quality q. But it is objectively legalized by a pure law (*objektiv-ideale Notwendigkeit*, *reine und objektive Gesetzlichkeit*) which acts only at the level of pure essences (*reinen Wesen*). This "ideal a priori necessity grounded in the material essences" (*in den sachlichen Wesen gründenden idealen oder apriorischen Notwendigkeit*) is, according to Husserl, a typical example of the synthetic *a priori*.

[2] For the use of topos theory in cognitive sciences, see also the contributions of A. Peruzzi and G. White in these Proceedings. See also Peruzzi 1991.

In the §§8–9, Husserl analyzes the difference between the contents which intuitively stand out against a background (*anschaulich sich abhebenden Inhalten*) and the contents which are intuitively merged and fused together (*verschmolzenen*). Perception presupposes a global unity of the intuitive moments and a "*phänomenal Abhebung*", that is, a saliency in Thom's sense. It is such a saliency which is expressed by the difference between, on the one hand, contents intuitively separated (*gesonderten, sich abhebenden, sich abscheidenden*) from the neighboring ones and, on the other hand, contents merged with the neighboring ones (*verschmolzenen, überfließenden, ohne Scheidung*).

The concept of fusion or merging – of *Verschmelzung* – is a key one. It expresses the spreading of qualities, that is the topological transition from the local level to the global one. Its complementary concept is that of separation, of disjunction – of *Sonderung*. *Sonderung* is an obstacle to *Verschmelzung*. It generates boundaries which delimit parts. At the intuitive, synthetic *a priori*, level, the "whole/part" difference is grounded on the "*Verschmelzung/Sonderung*" one.

Remark 1 It must be pointed out that Husserl's pure description fits very well with contemporary research. For instance Stephen Grossberg, one of the leading specialists of vision, concludes from his numerous works that there are essentially two fundamental systems in visual perception.

(i) The *Boundary Contour System* (BCS) which controls the segmentation of the visual scene. It detects, sharpens, enhances and completes edges, especially boundaries, by means of a "spatially long-range cooperative process". The boundaries organize the image geometrically (i.e. morphologically).

(ii) The *Featural Contour System* (FCS) which performs featural filling-in, that is spreading of qualities. It stabilizes qualities such as color or brightness. These diffusion processes are triggered and limited by the boundaries provided by the BCS.

Therefore, according to Grossberg,

> Boundary Contours activate a boundary completion process that synthesizes the boundaries that define perceptual domains. Feature Contours activate a diffusion filling-in process that spreads featural qualities, such as brightness or color, across these perceptual domains" (Grossberg 1988, p. 35).□

Remark 2 In fact, the concept of *Verschmelzung* does not come from Stumpf but from the German psychologist Johann Friedrich Herbart (1776-1841) who developed a *continuous* theory of mental representations. Es-

sentially in the same vein as Peirce after him, Herbart was convinced that mental contents are *vague* and can vary continuously. For him, a "serial form" (*Reihenform*) was a class of mental representations which undergo a graded fusion (*abgestufte Verschmelzung*) glueing them together via continuous transitions. He coined the neologism of *synechology* for his metaphysics (Peirce's neologism of *synechism* is clearly equivalent). It is not sufficiently known that Herbart's point of view was one of the main influences on Bernhard Riemann when he developed the key concept of a Riemannian manifold. Even if Riemann did not agree with Herbart's metaphysics, he strongly claimed that he was "a Herbartian in psychology and epistemology". Erhard Scholtz (1992) has shown that in Riemann's celebrated "Über die Hypothesen, welche der Geometrie zu Grunde liegen" (1867) the role of the differentiable manifold underlying a Riemannian manifold "is taken in a vague sense by a Herbartian-type of 'serial form', backed by mathematical intuition".□

Still in the §8 of the third *Logische Untersuchung*, Husserl claims "Sonderung beruht ... auf *Diskontinuität*". *Verschmelzung* corresponds to a continuous (*stetig*) spreading of qualities in an undifferentiated unity (*unterschiedslose Einheit*) (Husserl 1900–1, p. 244), while *Sonderung* corresponds to qualitative discontinuities of the way that qualities cover extension (*Deckungszusammenhang*).

These qualitative discontinuities are *salient* only if:

(i) they are contiguously unfolded (*sie angrenzend ausgebreitet sind*) against the background of a moment which varies continuously (*ein kontinuierlich variierendes Moment*), namely the spatial and temporal moment;

(ii) they present a sufficient gap (threshold of discrimation).

Husserl's morphological description is precise and remarkable:

> It is from a spatial or temporal limit [*einer Raum- oder Zeitgreuze*] that one jumps from a visual quality to another. In the continuous transition [*kontinuierlichen Übergang*] from a spatial part to another, one does not progress also continuously in the covering quality [*in der überdeckenden Qualität*]: in some place of the space, the adjacent neighboring qualities [*die angrenzenden Qualitäten*] present a finite (and not too small) gap [*Abstand*]" (Husserl 1900–1, p. 246).

This Husserlian pure eidetic description of the unilateral dependence "quality → extension" yields therefore the following homologies set out in Table 1.

But even if it is precise and remarkable, Husserl's morphological eidetic description raises nevertheless a fundamental problem. As we have seen, qualitative discontinuities

Totality (Whole)	**Parts**
Verschmelzung	Sonderung
Spreading activation (featural filling-in)	Boundaries
Continuity	Discontinuity

Table 1

> concern the minimal specific differences [*die niedersten spezifischen Differenzen*] in a same immediately superordinate [*übergeordnet*] pure genus [*Gattung*] (Husserl 1900–1, p. 246).

They are discontinuities of the concrete functional dependences "quality → extension". But, according to Husserl, it is impossible to formalize them. Formalization can only operate at a higher level of abstraction, the level of the general eidetic law of dependence.

We meet here a formalist thesis, which subordinates the regional material ontologies to formal ontology, and therefore the synthetic *a priori* laws to the analytic *a priori* ones. As regards its material content, the eidetic law of dependence "quality → extension" belongs to the sphere of "*der vagen Anschaulichkeiten*". But, according to Husserl, these vague – inexact – morphological essences cannot be *geometrically* constructed. As he claims at the beginning of §9,

> "Kontinuität und Diskontinuität sind natürlich nicht in mathematischer Exaktheit zu nehmen" (Husserl 1900–1, p. 245).

It is not possible to clarify here this fundamental point. But nevertheless I want to emphasize the fact that one of the main limits of phenomenology is its divorce of any "material descriptive eidetic" of *Erlebnis* from any form of geometry. Husserl always rejected the possibility of a morphological geometry. In some outstanding sections of *Ideen* I (§§ 71-75), he explains the fundamental difference between, on the one hand, the vague inexact descriptive concepts correlated with morphological essences, and, on the other hand, exact ideal mathematical concepts. According to him, *ideation*, which brings exact essences to ideality, is drastically different from *abstraction*, which brings inexact essences to genericity (categorization and typicality). This opposition is a key one for Husserl. Nevertheless, it is no longer acceptable.

3 Some Convergent Scientific Explanations of Phenomenological Description

We want now to stress that the phenomenological pure eidetic description fits very well with several scientific explanations which are remarkably convergent.

3.1 Topologico-Geometrical Schematization (Thom)

In this section 'internal' means "internal to the material system under consideration".

Phenomenologically, a material system S occupying a spatial domain W manifests its form through observable and measurable qualities $q^1(w)$, ..., $q^n(w)$, which are characteristic of its actual internal state A_w at every point $w \in W$, and which, as we will see in a moment, are sections of fibrations having as typical fibers the quality types Q^1,...,Q^n (colour, texture, etc.). When the spatial parameter w varies smoothly in W, A_w varies smoothly. If A_w remains the actual state, then the q^i also vary smoothly. But if the actual state A_w bifurcates towards another actual state B_w when w crosses some critical value, then some of the q^i must present a discontinuity. Thom called *regular* the points $w \in W$ where locally all the qualities q^i vary smoothly, and *singular* the points $w \in W$ where locally some of the q^i present a qualitative discontinuity. The set R_W of regular points is by definition an open subset of W and its complement K_W, the set of singular points, is therefore a closed set. We say that K_W is the *morphology* yielded by the internal dynamical behavior of the system S.

The singular points $w \in K_W$ are the *critical values* of the control parameters and, in physical cases, the system S has, for these values, critical internal behavior. Thom was one of the first scientists to stress the point that qualitative discontinuities are phenomenologically dominant, that every qualitative discontinuity is a sort of critical phenomenon and that a general mathematical theory of morphologies presented by general systems has to be an enlarged theory of critical phenomena.

Now it is clear that Thom's description is an exact topological version of Husserl's one. Regular points correspond exactly to *Verschmelzung*, and singular ones to *Sonderung*.

3.2 Morphodynamical and Cognitive Explanation

Let S be a material substrate. The problem is to explain its observable morphology, K.

1 At the most peripherical level, visual processing is a signal analysis and it is well known that the most basic signal analysis is Fourier analysis.

David Marr (1982) introduced the hypothesis that the main function of the ganglion cells of the retina is to extract locally the qualitative discontinuities (zero-crossings) which are encoded in the signal and that the higher levels of visual processing are grounded in this early morphological organization of the image (the primal sketch). In fact, it has been shown that the convolution of the signal by the receptive profiles of the ganglion cells (which are essentially Laplacians of Gaussians), is *wavelet analysis*, that is, spatially localized and multiscale Fourier analysis. Now wavelet analysis

is actually the best known device for detecting discontinuities.[3]

Stéphane Mallat has proved "Marr's conjecture": an image can be reconstructed from its qualitative discontinuities scanned at different scales. This shows that the morphological nucleus can be also recovered through signal analysis.

2 Multiscale and multichannel edge detections can serve as input for image segmentation algorithms. The problem is to segment in an optimal way an image $I(x, y)$ defined on a domain W, that is to partition it in maximally homogeneous domains limited by boundaries K.To do this we need a functional $E(W, K)$ which optimises the possible partitions. E must contain two terms: a term which measures the variance of $I(x, y)$ on the connected components of $W - K$, and a term which controls the length, the smoothness, the parsimony and the location of the boundaries.[4] Such a variational algorithm optimizes the way in which one can merge neighboring pixels in homogeneous domains separated by qualitative discontinuities. It provides therefore a *variational* approach to the *Verschmelzung/Sonderung* duality.

3 In what concerns the cortical processing of information, the problem of parts and wholes that is the problem of constituency is also called the *binding problem.* It is evident. At the early stages of perception the features of the objects are extracted in a local, distributed and parallel manner. How can these localized features (parts) be re-integrated in spite of their distributed encoding?

The main idea is that the binding of different features of an object (e.g. the segmentation of visual scenes) may be realized using a temporal coding. The coherence of features and constituents would be encoded in the synchronization (phase locking) of oscillatory neuronal responses to stimuli. And therefore different phases can code for different constituents. This hypothesis is also called the labeling hypothesis.

There is a large amount of experimental evidence concerning synchronized oscillations in the cortical (hyper)columns (in the frequency range of 40 70 Hz) which are sensible to the coherence of the stimulus.[5] With such a mechanism one can explain the global phenomena of fusion and separation: fusion $\equiv$ synchronisation, segmentation $\equiv$ desynchronisation.

To model the fundamental fact of synchronization we need the theory of networks of weakly coupled oscillators, the frequency of which depends on the intensity of the stimulus. These networks are typical *complex* physical systems.

[3] For an introduction to the use of wavelet analysis in computational vision, see also Mallat-Zhong 1989. For a discussion of the link with morphological phenomenology, see e.g. Petitot 1989b, 1990, 1993b,c.

[4] See, for example, Mumford & Shah 1989.

[5] See, for example, Engel, König, Gray and Singer 1992).

	Totality (wholes)	**Parts**
Phenomenological description	*Verschmelzung*	*Sonderung*
Topologico-morphological description	Continuity	Discontinuity
Morphodynamico-physical explanation	Stability of attractors	Bifurcation of attractors
Cognitive explanation I: Wavelet analysis	Behaviour of the amplitude of the wavelet transform of the signal	
Cognitive explanation II: oscillator networks	Synchronised oscillations (phase locking)	desynchronised oscillations

Table 2

A lot of work has been devoted to analysis using qualitative dynamics (e.g. George Bard Ermentrout and Nancy Kopell) or statistical physics (e.g. Y. Kuramato). In a nutshell, the theory of weakly coupled oscillators:

(i) shows that such systems enhance and complete existing boundaries;

(ii) can generate new virtual boundaries (which are not in the inputs);

(iii) confirms the labeling hypothesis.

3.3 Returning to the Morphological Nucleus

There is therefore a remarkable convergence of several different models of the morphological nucleus: physical, morphodynamical, geometrico-topological, sensorial (wavelet analysis), cortical (networks of oscillators). All these models confirm and naturalize the eidetic phenomenological pure description of forms.

4 Sheaf Mereology and the Internal Logic of Topoi

4.1 From Morphological Geometry to Formal Ontology and Logic

I come now to the second part of my paper. It will be rather technical and concern the link of the morphological nucleus, as it is eidetically described by geometry, with logic, mereology and formal ontology. In establishing such a link the key concept will be that of a *sheaf*. Why?

We have seen that at many converging levels, a form is essentially a set of discontinuities (a segmentation, a homogeneity breaking, that is a symmetry breaking) of a covering relation "quality → extension". We need therefore a deeper analysis of the concept of "*Überdeckung*". Now the geometrical concept of a sheaf is exactly the mathematical one which does the job.

Before tackling this point, we must bear in mind that the concept of covering relation is philosophically of utmost importance. I refer to the works of Kevin Mulligan, Barry Smith and Peter Simons.

(i) Covering relations yield prototypical examples of *dependent moments*, which are particular, monadic and static.

(ii) This dependence relation is unilateral and non-conceptual. The fact that every qualitative moment depends *generically* (i.e. as a type) on some extension makes the dependence relation "quality → extension" *internal* at the generic level (it is an *a priori* law between essences in Husserl's sense). But the fact that it depends *specifically* (as a token) on its extension makes the relation *external* (it is a contingent functional dependence in Husserl's sense).

(iii) Individual independent things exist in spatio-temporal domains W. Their relations with their qualities are external.

(iv) But the relations between the qualities themselves are internal.

In his paper *Internal Relations* (1992), Kevin Mulligan quotes Wittgenstein (*Remarks on Colour*):

> A language-game: Report whether a certain body is lighter or darker than another. But now there's a related one: State the relationship between the lightnesses of certain shades of colour. The form of the propositions in both language games is the same "X is lighter than Y". But in the first it is an external relation and the proposition is temporal. In the second it is an internal relation and the proposition is timeless.

This deep remark asserts that the dependence on space-time of relations involving qualities is not linguistically representable. I will try to explain how it can be mathematically fully justified.

The main problem is to unify a *logical axiomatics* of dependence relations with the *geometric eidetics* of covering relations. In general, geometry is sacrificed to logic (this is an aspect of the reluctance on the part of many philosophers to accept the transcendental concept of synthetic *a priori*). Here we will do justice to geometry and give a faithful sheaf model of Husserl's synthetic a priori law of covering. Then we will use one of the main discoveries of the logic of the last twenty years, namely that every category of sheaves (in the mathematical sense of "category") yields a logic (which is called its "internal" logic). Such a logic explains why Wittgenstein is right, that is why it is the *same* linguistic expressions which are used for external relations of token dependence and internal relations of type dependence.

4.2 Fiber Bundles and Sections

There is a fundamental geometric structure which fits perfectly well with Husserl's eidetic pure description of the spreading [*Ausbreitung*] of a quality in an extension or, equivalently, of the covering [*Überdeckung*, *Deckungszusammenhang*] of an extension by a quality. It is the key geometrical concept of fiber bundle or *fibration*.

Let the spatial substrate (*Ausdehnung*) of the form be modeled by a differentiable manifold W. Let Q be the qualitative genus under consideration (e.g. the space of colors). Q can be modeled by a manifold endowed with a categorization, that is with a decomposition in domains (categories) centered around central values (prototypes).

We have seen that a spreading-covering relation between W and Q can be naively defined as a map $q : W \to Q$ which, given any point $x \in W$, associates the value $q(x) \in Q$ of the quality at this point. This models Husserl's functional dependence. *Verschmelzung* is then expressed by the differentiability of q and *Sonderung* by discontinuities of q. These discontinuities constitute a closed subset K of W which expresses geometrically the salient morphology profiled in W.

But this naive model is too naive. Indeed, we need to have *all* the space Q at hand at *every* point $x \in W$. This requirement is imposed by Husserl's pure description. But it is also a perceptual fact. It has been shown by many neurological experiments that the covering of extensions by qualities such as colors or by local geometrical elements such as directions are neurally implemented by (hyper)columns, that is by retinotopic structures where, "over" each retinian position, there exists a "column" implementing the same set of possibilities.

This leads to the fundamental and pervasive concept of *fibration*, which was introduced by Whitney, Hopf and Stiefel and which concerns, in modern geometry and mathematical physics, all the situations where fields of non spatio-temporal entities functionally depend on space-time positions.

Mathematically, a fibration is a differentiable manifold E endowed with a *canonical projection*, a differentiable map $\pi : E \to M$ over another manifold M. M is called the *base* of the fibration, E its *total space*. The inverse images $E_x = \pi^{-1}(x)$ of points $x \in W$ by π are called the *fibers* of the fibration. They are the subspaces of E which are projected to points.

In general a fibration is required to be *locally trivial*, that is to satisfy the two following axioms:

$\mathbf{F}_1$ All the fibers E_x are diffeomorphic with a typical fiber F.

$\mathbf{F}_2$ $\forall x \in M$, $\exists U$, a neighborhood of x, such that the inverse image $E_U = \pi^{-1}(U)$ of U is diffeomorphic to the direct product $U \times F$ endowed with the canonical projection $U \times F \to U$, $(x, q) \mapsto x$.

In our case, we have $M = W$ and $F = Q$. How can we interpret the concept of functional dependence in this new context? It corresponds to the

key concept of a *section* of a fibration. Let $\pi : E \to M$ be a fibration and let $U \subset M$ be an open subset of M. A section s of π over U is a lift of U to E which is compatible with π. More precisely, it is a map $s : U \to E$, $x \in U \mapsto s(x) \in E$, and such that $\pi \circ s = \mathrm{Id}_U$. In general s is supposed to be continuous, differentiable, analytic. It can presents discontinuities along a singular locus.

It is conventional to write $\Gamma(U)$ for the set of sections of π over U. A local trivialization of π over U (i.e. $E_U \times U \approx U \times F \to U$) transforms every section $s : U \to E$ into a map $x \mapsto s(x) = (x, f(x))$, that is into a map $f : U \to F$. Therefore, the concept of section generalizes the classical concept of map, that is of functional dependence.

We can now establish the link with Husserl's description.

(1) The functional dependences determined at the level of minimal specific differences [*die niedersten spezifischen Differenzen*] correspond exactly to *particular* sections $\sigma : W \to E$ of *particular* fibrations $\pi : E \to W$ with fiber Q. These sections model external relations of token dependence.

(2) The qualitative salient discontinuities are discontinuities of sections $\sigma \in \Gamma(U)$.

(3) The eidetic law "concretely determined by its material contents" corresponds to a particular fibration $\pi : E \to W$ with fiber Q, but without any particular given section. Such a fibration models an abstract relation between the genus W and Q (the first level of abstraction). It implicitly contains an infinite universe of potential functional dependences, namely, all the sets of sections $\Gamma(U)$ for $U \subset W$. It models internal relations of type dependence.

(4) The synthetic *a priori* law of dependence "quality → extension" corresponds to the general mathematical structure of fibration. It concerns the most abstract genus – the essences – of space and quality (the second level of abstraction).

(5) Last but not least, the "analytic axiomatization" of this synthetic law in the framework of formal ontology corresponds to the *axiomatics* of fibrations.

This interpretation of Husserl's description in terms of fibrations uses only globally trivial fibrations and is therefore equivalent to a more classical one, using only functional dependences. But, as we will see, the concept of section will allow us to link the problem of covering relations (*Überdeckung*) with a very deep synthesis between geometry and logic.

4.3 Sections and Sheaves

At an abstract level, a fibration is characterized by the sets of its sections $\Gamma(U)$ over the open sets $U \subset M$. If $s \in \Gamma(U)$ is a section over U and if

$V \subset U$, we can consider the *restriction* $s|_V$ of s to V. Restriction is a map $\Gamma(U) \to \Gamma(V)$. It is clear that if $V = U$, then $s|_V = s$, and that if $W \subset V \subset U$ and $s \in \Gamma(U)$, then $(s|_V)|_W = s|_W$ (transitivity of restriction). We get, therefore, what is called a *contravariant functor* $\Gamma : \mathcal{O}(M) \to \mathbf{Sets}$ from the category $\mathcal{O}(M)$ of open sets of M to the category **Sets** of sets. (The objects of $\mathcal{O}(M)$ are the open sets of M, and its morphisms are the inclusions of open sets.)[6]

Conversely, let Γ be such a functor – what is called a *presheaf* on M. To have a chance of being the functor of the sections of a fibration, Γ must clearly satisfy the two following axioms.

S_1 Two sections which are locally equal must be globally equal. Let $\mathcal{U} = (U_i)_{i \in I}$ be an open covering of M. Let $s, s' \in \Gamma(M)$. If $s|_{U_i} = s'|_{U_i}$ $\forall i \in I$, then $s = s'$.

S_2 Compatible local sections can be collated into a global one. Let $s_i \in \Gamma(U_i)$ be a family over $\mathcal{U} = (U_i)_{i \in I}$. If the s_i are compatible, that is if $s_i|_{U_i \cap U_j} = s_j|_{U_i \cap U_j}$ when $U_i \cap U_j \neq \emptyset$, then they can be glued together: $\exists s \in \Gamma(M)$ such that $s|_{U_i} = s_i$ $\forall i \in I$.

These axioms actually characterize a more general structure – one even more pervasive in contemporary mathematics – than the structure of fibration, namely the structure of a *sheaf*. It can be shown that if the axioms (S_1) and (S_2) are satisfied, then one can represent the functor Γ by a general fibered structure $\pi : E \to M$ (called an "étale" space and which is not necessarily a locally trivial fibration) in such a way that $\Gamma(U)$ becomes the set of sections of π over U.

The concept of sheaf expresses essentially *glueing conditions*, that is the way by which local data can be collated in global ones. It is the right mathematical tool for formalizing the covering relations between spatial locations and non spatial determinations. We will see now that this geometric eidetics of local/global covering relations is, in an essential manner, linked to logic.

4.4 The Topos $\mathbf{Sh}(M)$

In the following sections if A is a sheaf on M, $A(U)$ will denote the set $\Gamma_A(U)$ of sections of A over U.

It is easy to show that the sheaves on a base space M constitute a category $\mathbf{Sh}(M)$. Now the main point is that:

(i) this category shares fundamental properties which are characteristic of what is called a *topos* structure, and

(ii) a topos structure is exactly what is needed for doing logic.

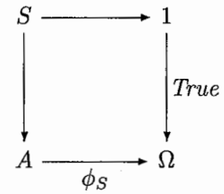
[6] We suppose that the reader is acquainted with category theory. For an introduction see e.g. Peruzzi 1991 and Petitot 1979, 1982.

4.4.1 Exponentials

$\mathbf{Sh}(M)$ is a *Cartesian closed* category. This means that it has products and fibered products or pullbacks, a terminal object – classically denoted by 1: the constant sheaf defined by $1(U) = U$ for every U – and *exponentials* B^A. An exponential object is an object which "internalizes" in the objects of $\mathbf{Sh}(M)$ the morphisms $f : A \to B$. Such "internalizations" of functorial structures are called *representable* functors. Technically, the functor $(\cdot)^A$ is the *right adjoint* of the functor $A \times (\cdot)$. This means that we have for every object C of $\mathbf{Sh}(M)$ a functorial isomorphism $\mathrm{Hom}(C, B^A) \cong \mathrm{Hom}(A \times C, B)$. For example, for $C = 1$, we get $\mathrm{Hom}(1, B^A) \cong \mathrm{Hom}(A, B)$. But an arrow $f : 1 \to B^A$ is like an "element" of B^A. In fact, if A is a sheaf, an arrow $s : 1 \to A$ is a *global section* of A, that is an element $s \in A(M)$.

4.4.2 Subobject Classifier

$\mathbf{Sh}(M)$ also possesses what is called a *subobject classifier* Ω, that is an object which "internalizes" sets of subobjects, making the subobject functor representable. A subobject $m : S \hookrightarrow A$ is a monomorphism (an injective map in the case of the category of sets). This means that if f and g are two morphisms from an object R to S, then $m \circ f = m \circ g$ implies $f = g$. It is equivalent to say that the fibered product $S \times_A S$ defined by m is isomorphic to S. A subobject classifier is a monomorphism $True : 1 \to \Omega$ such that every subobject can be retrieved from *True* by a pull-back:



We get therefore a functorial isomorphism $\mathrm{Sub}(A) \cong \mathrm{Hom}(A, \Omega)$. ϕ_S is called the *characteristic map* of the subobject S.

In the category **Sets** of sets, $\Omega = \{0, 1\}$ is the classical set of boolean truth-values. Here – and this is perhaps the main difference between a topos like $\mathbf{Sh}(M)$ and the classical topos **Sets** – $\Omega(U)$ depends essentially on the topological structure. It expresses the way that truth behaves locally in a sheaf topos: in this case, by definition, $\Omega(U) := \{W \subset U\}$. It is trivial to verify that Ω is a sheaf. The map $True : 1 \to \Omega$ is defined by $True(U) : 1 \mapsto U \in \Omega(U)$, that is, by the *maximal* element of $\Omega(U)$: to be true over U is to be true "everywhere" over U.

4.4.3 Elements, Properties and Parts

In a topos, the morphisms $a : B \to A$ are called *generalized elements* of A, or elements *defined on* B (this denomination comes from algebraic geometry

and, more precisely, from Grothendieck's theory of schemes). Among the elements, the most important are those defined on open sets U, that is, the sections $s \in A(U)$. The elements defined on the terminal object 1 are "global". We will see that an arrow $\theta : A \to \Omega$ is a "predicate" on A, that is a "property" of its generalized elements. Among all predicates, there is the predicate $True_A : A \to 1 \to \Omega$. It is easy to verify that an element $a : B \to A$ factorizes through a subobject $S \hookrightarrow A$ iff the composite of A and the characteristic map of S, $\phi_S \circ a = True_B$. ϕ_S is therefore the predicate on A which is true exactly for those elements of A which are in S. The uniqueness of ϕ_S expresses the extensionality principle.

Using the exponentials and the subobject classifier we can define the *parts* of an object A as another object $\mathcal{P}(A) = \Omega^A$. We get the functorial isomorphisms:

$$\mathrm{Sub}(A) \cong \mathrm{Hom}(A, \Omega) \cong \mathrm{Hom}(A \times 1, \Omega) \cong \mathrm{Hom}(1, \Omega^A) = \mathrm{Hom}(1, \mathcal{P}(A)).$$

We have therefore $\Omega = \mathcal{P}(1)$.

This shows that there are three equivalent descriptions of a subobject $m : S \hookrightarrow A$.

(i) its "extension" S: we will see that it can be symbolized as in **Sets** by $\{a | \phi_S(a)\}$;

(ii) its characteristic map $\phi_S : A \to \Omega$ which is a "predicate" on A;

(iii) the global section $s : 1 \to \mathcal{P}(A)$ which is its "name".

4.4.4 Towards Logic

The existence of an intuitionistic "internal logic" in a topos of sheaves depends essentially on the fact that, for each open set, $\Omega(U)$ is itself the set of subobjects of U, i.e. of open subsets W of U regarded as a topological space. Consequently, for each U, $\Omega(U)$ is a *Heyting algebra*, and Ω itself is therefore a sheaf of Heyting algebras (or, in more category-theoretic terms, a Heyting algebra object in $\mathbf{Sh}(M)$). The consequence is that the "external" set of subobjects $\mathrm{Sub}(A)$) and the "internal" one $\mathcal{P}(A)$) are also Heyting algebras, the canonical isomorphism $\mathrm{Sub}(A) \cong \mathrm{Hom}(1, \mathcal{P}(A)$ being an isomorphism of Heyting algebras.

4.5 Topoi and Logic

Now, the central fact is that a topos is exactly the categorical structure which is needed for doing logic. Furthermore, this logic is *spatially localized*. (For details see Mac Lane, Moerdijk 1992.)

4.5.1 Types and Localization

We can associate with each topos $\mathbf{Sh}(M)$ a formal language $\mathcal{L}_M$, called its *Mitchell-Bénabou language*, and a semantics based on forcing, called its *Kripke-Joyal semantics*. The crucial point is that a sheaf X can be considered as a *type* for variables x, which are themselves interpreted as *sections* $s \in X(U)$ of X. We get therefore, at the same time, a *logical* typing and a *spatial* localization of the variables. This achievement fits perfectly well with Husserl's description and explain Wittgenstein's remark. It provides them with a correct mathematical status.

(i) Sections are tokens denoted by variables belonging to types (species, essences). They are "concretely" particularized by the specification of their localization U and by their specific values. But as an element of type X, a section s particularizes an abstract unilateral relation of dependence, the relation "quality $\to$ extension" which is constitutive of X.

(ii) The relations between particular sections $s \in X(U)$, $t \in Y(V)$ are *external*. The relations between X and Y are *internal*. Nevertheless the "linguistic" expressions which express them are formulas in the formal language $\mathcal{L}_M$ associated to $\mathbf{Sh}(M)$, and are therefore the same.

4.5.2 Syntax

How are the terms and the formulas of $\mathcal{L}_M$ syntactically constructed? Here is a summary of their inductive construction. The intuitive idea is that classical entities are *relativized* to open sets U, in as much the same way as, in the Kripkean conception of modal logic, they are relativized to possible worlds.

A term σ of type X constructed using variables y, z of respective types Y, Z has a *source* $Y \times Z$ and is interpreted by a morphism $\sigma : Y \times Z \to X$ which expresses its structure.

(i) To each $X \in \mathbf{Sh}(M)$, considered as a type, are associated variables x, x', ... They are each interpreted by the identity map $1_X : X \to X$.

(ii) Terms $\sigma : U \to X$, $\tau : V \to Y$, of respective types X and Y, yield a term $\langle \sigma, \tau \rangle$ of type $X \times Y$ (the ordered pair), interpreted by $\langle \sigma, \tau \rangle : U \times V \to X \times Y$.

(iii) Terms $\sigma : U \to X$, $\tau : V \to X$ of the same type X yield the term $(\sigma = \tau)$ of type Ω interpreted by:

$$(\sigma = \tau) : W = U \times V \longrightarrow X \times X \xrightarrow{\delta_X} \Omega,$$

where δ_X is the characteristic function of the diagonal subobject $\Delta : X \to X \times X$, $x \mapsto \langle x, x \rangle$.

(iv) A term $\sigma : U \to X$ of type X and a morphism $f : X \to Y$ yield, by composition, a term $f \circ \sigma$ of type Y.

(v) Terms $\theta : V \to Y^X$ and $\sigma : U \to X$ of respective types Y^X and X yield a term $\theta(\sigma)$ of type Y interpreted by:

$$\theta(\sigma) : W = V \times U \longrightarrow Y^X \times X \xrightarrow{e} Y,$$

where e is the *evaluation* map which associates to every "functional" element $f \in Y^X$, and every element $x \in X$, the value $f(x)$ of f at x.

(vi) In particular, terms $\sigma : U \to X$ and $\tau : V \to \Omega^X$ yield a term $\sigma \in \tau$ (the *membership relation*) of type Ω, which is interpreted by:

$$\sigma \in \tau : W = V \times U \longrightarrow X \times \Omega^X \xrightarrow{e} \Omega.$$

(vii) A variable x of type X and a term $\sigma : X \times U \to Z$, of type Z and with source $X \times U$, yield a λ-term of type Z^X interpreted by $\lambda x\sigma : U \to X^Z$.

(viii) Ω is the type of the *formulas* of $\mathcal{L}_M$. As Ω is a Heyting algebra, we get the logical operations of propositional calculus: $\phi \wedge \psi$, $\phi \vee \psi$, $\phi \Rightarrow \psi$, $\neg\phi$. It is easy to verify that if $\phi(x, y) : X \times Y \to \Omega$ is a formula, we can write the subobject of $X \times Y$ classified by its interpretation in a "set-theoretic" manner: $\{\langle x, y\rangle \in X \times Y | \phi(x, y)\}$.

(ix) One of the most remarkable facts of topos theory is that it is possible to define *quantification* in a purely categorical manner. Let $f : A \to B$ be a morphism of $\mathbf{Sh}(M)$, and consider the "inverse image" functor $f^* : \mathrm{Sub}(B) \to \mathrm{Sub}(A)$ defined by composition with f. Its internal version is the morphism $\mathcal{P}(f) : \mathcal{P}(B) \to \mathcal{P}(A)$. The fact is that $\mathcal{P}(f)$ has two adjoint functors: a left adjoint, $\exists_f : \mathcal{P}(A) \to \mathcal{P}(B)$, and a right adjoint $\forall_f : \mathcal{P}(A) \to \mathcal{P}(B)$. They generalize the two adjunctions in **Sets**.

These categorical constructs show that the formal language $\mathcal{L}_M$ is, on the linguistic level, exactly of the same nature as the classical formal language of sets. The main difference is that we have introduced a subtle dialectics between the logical type of the variables and their spatial localization.

4.5.3 Semantics

The Kripke-Joyal semantics of topoi is a forcing semantics which generalizes Cohen's. A variable x of type X denotes a section $s \in X(U)$, which can be redescribed as a morphism $U \to X$, where U is now the sheaf defined by U. The semantic rules inductively define a forcing relation $U \Vdash \phi(s)$ (U forces $\phi(s)$). Let $s : U \to X$, and let $\mathrm{Im}(s) \in \mathrm{Sub}(X)$ be the image of s. One defines

$$U \Vdash \phi(s) := \mathrm{Im}(s) \subseteq \{x|\phi(x)\}$$

that is, $U \Vdash \phi(s)$ iff $U \xrightarrow{s} X \xrightarrow{\phi} \Omega$ factorizes through $\{x|\phi(x)\}$:

$$U \longrightarrow \{x|\phi(x)\} \longrightarrow 1 \xrightarrow{True} \Omega.$$

The semantic rules are:

(i) $U \Vdash \phi(s) \wedge \psi(s)$ iff $U \Vdash \phi(s)$ and $U \Vdash \psi(s)$.

(ii) $U \Vdash \phi(s) \wedge \psi(s)$ iff there exists an open covering $(U_i)_{i \in I}$ of U such that, for every $i \in I$, $U \Vdash \phi(s|_{U_i})$ or $U \Vdash \psi(s|_{U_i})$. (This is the intuitionistic rule for disjunction).

(iii) $U \Vdash \phi(s) \Rightarrow \psi(s)$ iff, for all $V \subseteq U$, $V \Vdash \phi(s|_V)$ implies $V \Vdash \psi(s|_V)$.

(iv) $U \Vdash \neg\phi(s)$ iff there does not exist $V \subseteq U$, $V \neq \emptyset$, such that $V \Vdash \phi(s|_V)$. (This defines an intuitionistic negation, because Ω is a Heyting algebra and not a Boolean one.)

(v) $U \Vdash \exists y \phi(s, y)$ (with y of type Y) iff there is an open covering $(U_i)_{i \in I}$ of U and sections $\beta_i \in Y(U_i)$ such that, for every $i \in I$, $U_i \Vdash \phi(s|_{U_i}, \beta_i)$.

(vi) $U \Vdash \forall y \phi(s, y)$ iff, for every $V \subseteq U$ and every $\beta \in Y(V)$, we have $V \Vdash \phi(s|_V, \beta)$.

4.6 Sheaf Mereology

The concept of a section is a key mathematical one which allows us to build up models for a large class of dependent parts. It therefore provides a good model (in the sense of model theory) for axiomatic mereology,[7] which we shal call a *sheaf* model. This is why, to conclude this paper, I will investigate the relation between the axiomatics of fibrations and the piece of formal ontology proposed by Barry Smith in his "Ontology and the logistic analysis of reality" (1993).

Smith proposed an axiomatisation for mereology based on two primitives:

(i) a purely mereological one $x\mathsf{C}y$: "x is a *constituent* of y";

(ii) a topological one $x\mathsf{P}y$: "x is an *interior part* of y".

A fundamental axiom is that, for every predicate ϕ, one can define the *sum* – the fusion, the merging – of all the x which satisfy ϕ:

DC4 $[x : \phi(x)] = \iota y(\forall w(w\mathsf{O}y \Leftrightarrow \exists v(\phi(v) \wedge w\mathsf{O}v)))$,

where O is the *overlapping* relation defined by the axiom:

DC1 $x\mathsf{O}y := \exists z(z\mathsf{C}x \wedge z\mathsf{C}y)$.

[7] For mereology in the framework of formal ontology, see Poli 1992.

Of course, we need axioms to guarantee that the matrix of **DC4** is a definite description to which Russell's operator ι can be applied.

If we apply this general axiomatisation of mereology — which belongs to formal ontology – to the sheaf model of Husserl's pure eidetic description, we meet some far reaching discrepancies.

The basic elements of our universe are sections of a sheaf defined by a contravariant functor $\Gamma : \mathcal{O}^*(M) \rightarrow \mathbf{Sets}$. Given such an object $s \in \Gamma(U)$, we must carefully distinguish between:

(i) its domain of definition $\mathrm{Dom}(s) = U \subset M$, which is a detachable part of a geometrical extensive whole (the base manifold M), and

(ii) its values $s(x)$, which belong to an intensive space of qualities (the fiber F).

It must be emphasized again that the key concept of section supports a dialectic relating local and global: *restriction* from global to local and *glueing* from local to global. The domains of sections correspond to the purely extensional (or topological) part of the axiomatisation. Their values correspond on the other hand to the truly mereological part.

As the base space M is a manifold, all the concepts of open set, interior part, boundary, closure, etc. are *ipso facto* well defined.[8]But the concepts of fibration and sheaf deepen what it is for a section s to be a *constituent* of another section t. There are in fact (at least) two meanings of constituency, a weak one and a strong one.

The Weak Sense: $s \in \Gamma(U) \mathsf{C} t \in \Gamma(V) := U \subset V$ (that is, the domain of s is included in that of t.)

The Strong Sense: $s \in \Gamma(U) \mathsf{C} t \in \Gamma(V) := (U \subset V) \wedge (t|_U = s)$ (that is, s is a restriction of t.)

Of course, it is the strong sense which is the most interesting. With it, mereological overlaps become glueing conditions. More precisely, the glueing condition $s|_{U \cap V} = t|_{U \cap V}$ is a condition for *maximal* overlap (that is, of overlapping over $\mathrm{Dom}(s) \cap \mathrm{Dom}(t)$).

This strong sense of constituency is imposed by Smith's AC2 axiom:

AC2 $x \mathsf{C} y \wedge y \mathsf{C} x \Rightarrow x = y$.

It is also imposed by the axioms **AP1** and **AP2a,b** linking the primitives C and P.

AP1 $x \mathsf{P} y \Rightarrow x \mathsf{C} y$;

[8]In some cases one can generalize the concept of section and define it for non-open subsets of M. But in general "good" sections must, away from their singular locus, share some properties of continuity, differentiability, analyticity, etc. These are all *local* properties, which are well defined only on open sets.

AP2a $x \mathsf{P} y \wedge y \mathsf{C} z \Rightarrow x \mathsf{P} z$ (left monotonicity);

AP2b $x \mathsf{C} y \wedge y \mathsf{P} z \Rightarrow x \mathsf{P} z$ (right monotonicity);

Indeed, if $s \in \Gamma(U)$, the unique plausible meaning for $s \mathsf{P} t, t \in \Gamma(V)$, is that $U \subset \mathrm{Int}(V)$ and $t|_U = s$.[9] But then **AP1** and **AP2a,b** imply immediately $s \mathsf{C} t$ in the strong sense.

Now, the main point is that the mereological concept of *sum* (union and fusion) splits into two different concepts.

(1) The *union* of any two sections $s \in \Gamma(U)$ and $t \in \Gamma(V)$ can be defined as the "section" $s \cup t \in \Gamma(U \cup V)$ such that $s \cup t(x) = \{s(x), t(x)\}$ (we put $s(x)$, $t(x) = \emptyset$ if $x \notin U, V$). If $s(x) \neq t(x)$, $s \cup t$ is a *multivalued* section.

(2) The *fusion* of two sections $s \in \Gamma(U)$ and $t \in \Gamma(v)$ is more restrictive. It requires the glueing condition $s|_{U \cap V} = t|_{U \cap V}$. The fundamental consequence is that, if $\phi(s)$ is a predicate of sections, the sum $[s : \phi(s)]$ is no longer a *single* element. It is the set of *maximal* sections satisfying ϕ.

228z The concept of sum as fusion depends on the the concept of *prolongation* of a section. Let ϕ be a predicate of sections. Let us call ϕ-section a section satisfying ϕ. We look at the possibility of extending a ϕ-section $s \in \Gamma(U)$ to a larger open set $V \supset U$. We look therefore at sections $t \in \Gamma(V)$, where $V \supset U$, $s \mathsf{C} t$, and $\phi(t)$. t is a maximal ϕ-section if:

$$\forall r(\phi(r) \wedge t \mathsf{C} r \Rightarrow r = t).$$

A maximal ϕ-section t satisfies the matrix of **DC4**:

$$\forall w(w \mathsf{O} t \Leftrightarrow \exists v(\phi(v) \wedge w \mathsf{O} v)),$$

but this is no longer the matrix of a definite description.

One consequence is that the universe is no longer a single element: it is, rather, the set $\Gamma_m(M)$ of maximal sections. We have, of course, $\Gamma(M) \subset \Gamma_m(M)$: global sections are maximal. But in general there are maximal sections which are not global and nevertheless cannot be extended.

We can therefore conclude that the mereology of sections – in the sheaf model which axiomatizes Husserl's pure eidetic description – shows that some mereological axioms are "evident" only for purely *extensional* mereology, and are by no means "evident" for more sophisticated sort of (non extensional) mereological models.

Another non standard feature of sheaf mereology concerns the concept of *points*. In the classical topos **Sets** the points are the arrows $1 \rightarrow A$, and **Sets** is a *well-pointed* topos in the sense that it has "enough" points: if

[9]Of course Int means here the topological interior. In general, V will be open and therefore $\mathrm{Int}(V) = V$.

$f, g : A \to B$ are two different maps then there exists a point $x \in A$ such that $f(x) \neq g(x)$. It can be proved that a well-pointed topos is boolean (i.e. that Ω is a Boolean algebra) and two-valued (1 and $0 = \emptyset$ are the only subobjects of 1, that is, there are only two global truth-values). Therefore, a topos $\mathbf{Sh}(M)$ will *not* be well-pointed in general: there will not be enough points (that is, global sections) to differentiate different arrows.

5 Conclusion: The Analytic/Synthetic Distinction

The formalization of Husserl's pure phenomenological description in terms of topos theory allows a remarkable clarification of the celebrated transcendental distinction between analytic and synthetic *a priori*. The formal language $\mathcal{L}_M$ corresponds to the logical, analytic component of the formalization. But this is not the end of the story. The other component concerns the *localization of truth*. The fact that truth-values are indexed on open sets and the forcing status of Kripke-Joyal semantics show that "space" is irreducible to logical analyticity and constitutes a *sui generis* geometrical dimension of the *a priori* stance. According to Wittgenstein's remark, it is unrepresentable by the "linguistic" (syntactic) form of the formula of $\mathcal{L}_M$, and this explains why the "linguistic turn" of analytic philosophy has completely occulted it. This geometrical dimension of truth corresponds exactly to what Husserl calls, after Kant, the synthetic *a priori*. In that sense the concept of "synthetic a priori" is, contrary to a widely spread belief, a perfectly sane one. That it has been dramatically misunderstood by logical positivism must not hide the fact that it is basic for ontology as soon as we want to unify spatial intuition and logical typing.

References

Dreyfus, H. (ed.) 1982 *Husserl, Intentionality and Cognitive Science*, Cambridge MA: MIT Press.

Dreyfus, H. 1982 "Husserl's Perceptual Noema", in Dreyfus (ed.) 1982.

Engel, A., König, P., Gray, C., Singer, W. 1992 "Temporal Coding by Coherent Oscillations as a Potential Solution to the Binding Problem: Physiological Evidence", in H. Schuster (ed.), *Non linear dynamics and Neural Networks*, Berlin: Springer.

Grossberg S. (ed.) 1988 *Neural Networks and Natural Intelligence*, Cambridge MA: MIT Press.

Holenstein, E. 1992 "Phenomenological Structuralism and Cognitive Semiotics" in R. Benatti (ed.), *Scripta Semiotica* 1, Peter Lang, pp. 133-158.

Husserl, E. 1900-1901 *Logische Utersuchungen*, Halle: Max Niemeyer. Second edition Halle: Max Niemeyer 1913.

Husserl, E. 1913 *Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie*, Husserliana III-IV.

Koenderink, J.J., van Doorn, A.J. 1976 "The Singularities of the Visual Mapping", *Biological Cybernetics* 25, 51–59.

Mallat, S.G. 1989 *Review of Multifrequency Channel Decompositions of Images and Wavelet Models*, Technical Report, CEREMADE, Paris.

Mallat, S.G., Zhong, S. 1989 "Complete Signal Representation with Multiscale Edges", *Technical Report n° 483*, Department of Computer Sciences, New-York University.

Marr, D. 1982 *Vision*, San Francisco: Freeman.

McIntyre , R., Woodruff Smith, D. 1982 "Husserl's Identification of Meaning and Noema", in Dreyfus (ed.) 1982.

McIntyre, R. 1986 "Husserl and the Representational Theory of Mind", *Topoi* 5, 101–113.

MacLane, S., Moerdijk, I. 1992 *Sheaves in Geometry and Logic*, New York: Springer.

Mulligan, K. 1992 "Internal Relations", *Australian National University Metaphysics Conference.*

Mumford, D., Shah, 1989 "Optimal Approximations by Piecewise Smooth Functions and Associated Variational Problems", *Comm. Pur. Appl. Math* XLII, 4.

Peruzzi, A., 1991 "Categories and Logic", *Problemi fondazionali nella teoria del significato*, (G. Usberti ed.), Firenze: Olschki.

Petitot, J. 1979 "Locale/Globale", *Enciclopedia Einaudi* VIII, 429–490, Torino: Einaudi.

Petitot, J. 1982 "Unità delle matematiche", *Enciclopedia Einaudi* XV, 1034–1085, Torino: Einaudi.

Petitot, J. 1985 *Morphogenèse du Sens*, Paris: Presses Universitaires de France.

Petitot, J. 1986 "Structure", *Encyclopedic Dictionary of Semiotics* (T. Sebeok, ed.), Vol 2, 991–1022, New-York: de Gruyter.

Petitot, J., 1989a. "Structuralisme et Phénoménologie", in Petitot (ed.) 1989, 345-376.

Petitot, J., 1989b " Forme", *Encyclopædia Universalis* XI, 712–728, Paris.

J. Petitot, (ed.) 1989 *Logos et Théorie des Catastrophes: René Thom Cerisy Symposium*, Genève: Patiño.

Petitot, J. 1990 "Le Physique, le Morphologique, le Symbolique. Remarques sur la Vision", *Revue de Synthèse* 1-2, 139–183.

Petitot, J. (ed.) 1990. "Sciences cognitives: quelques aspects problématiques", *Revue de Synthèse* IV, 1-2.

Petitot, J. 1992a *Physique du Sens*, Paris: Editions du CNRS.

Petitot, J. 1992b "Matière-Forme-Sens: un problème transcendantal", *Les Figures de la Forme* (J. Gayon, J.J. Wunenburger eds.), Paris: L'Harmattan.

Petitot, J. 1993a "Topologie phénoménale. Sur l'actualité scientifique de la *phusis* phénoménologique de Maurice Merleau-Ponty", *Merleau-Ponty. Le philosophe et son langage* (F. Heidsieck ed.), *Cahiers Recherches sur la philosophie et le langage* 15, 291–322, Paris: Vrin.

Petitot, J. 1993b "Attractor Syntax", *Mind as Motion*, (T. van Gelder, R. Port, eds), Cambridge MA:MIT Press.

Petitot, J. 1993c "Phénoménologie naturalisée et Morphodynamique", *Intellectica* 17, 79–126.

Petitot, J., Smith, B. 1991 "New Foundations for Qualitative Physics", *Evolving Knowledge in Natural Science and Artificial Intelligence*, (J.E. Tiles, G.J. McKee, G.C. Dean eds.), 231–249, Pitman: London.

Poli, R., 1992. *Ontologia formale*, Genova: Marietti.

Scholtz, E. 1992 "Riemann's Vision of a New Approach to Geometry", *1830-1930: a Century of Geometry*, (L. Boi, D. Flament, J.M. Salanskis eds.), Berlin: Springer.

Smith, B. (ed.) 1982 *Parts and Moments. Studies in Logic and Formal Ontology*, Philosophia: Vienne.

Smith , B. (ed.) 1988 *Foundations of Gestalt Theory*, Philosophia: Munich.

Smith, B. 1993 "Ontology and the Logistic Analysis of Reality", Preprint.

Smith, B., Mulligan, K. 1982 "Parts and Moments : Pieces of a Theory", in Smith (ed.) 1982, 15–109.

Thom, R. 1972 *Stabilité structurelle et morphogenèse*, New York: Benjamin.
Thom, R. 1980 *Modèles mathématiques de la Morphogenèse*, Paris: Christian Bourgois.
White, G., 1993. "Mereology, Combinatorics, and Categories", to appear in *The Monist.*

A Comparison of Structures in Spatial and Temporal Logics

A. G. Cohn, J. M. Gooday and B. Bennett

1 Introduction

Representing and manipulating knowledge about time and space is recognised as a crucially important part of commonsense reasoning (Davis 1990). If we are ever to construct a truly autonomous AI system then practical ways of reasoning about about these two domains must be found. Furthermore, as temporal and spatial reasoning are often bound together (consider, for example, the concept of motion) it would seem sensible to use similar approaches for both. In the last 15 years, a number of temporal reasoning formalisms have been proposed for AI applications amongst which Allen's (Allen 1981) interval calculus has proved to be the most practically useful. Recent work in spatial reasoning has resulted in formalisms that have many similarities to Allen's temporal reasoning system. It therefore seems appropriate to draw on the results of the interval-based temporal reasoning community in order to improve spatial reasoning techniques.

In this paper we examine some of the similarities between Allen's temporal reasoning formalism and the spatial systems of Randell, Cui and Cohn (1992b). An efficient method for generating the *composition tables* used by these systems is presented and we show how techniques developed for Allen's interval calculus can be used to compact the composition tables of spatial calculi. Finally, we show how it is possible to produce certain kinds of *transition graphs* for temporal and spatial calculi directly from composition tables.

2 Temporal and Spatial Calculi

Over the last decade Allen's interval calculus (Allen 1981) has been widely used with systems that require some form of temporal reasoning capability, such as planners (Allen, Kautz, Pelavin and Tenenberg 1991) and natural language understanding systems (Vieu 1991). Interval calculus makes use of the thirteen basic dyadic relations that can hold between two time intervals

The support of the SERC under grant no. GR/G36852 and GR/H 78955 is gratefully acknowledged. This work has also been partially supported by CEC ESPRIT basic research action MEDLAR II, 6471. The authors are particularly grateful to Nick Gotts for his comments on earlier versions of this paper.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 409-422

(Allen does not allow time points[1]) shown in figure 1. Transitivity properties can be used to determine which relations may hold between pairs of intervals. For example, given that interval i_1 is *before* i_2 and also that i_2 *meets* i_3 it can be inferred that i_1 is *before* i_3. Allen encoded all such transitivity information in a 13×13 table thus making the task of reasoning in interval calculus a simple matter of table look-up.

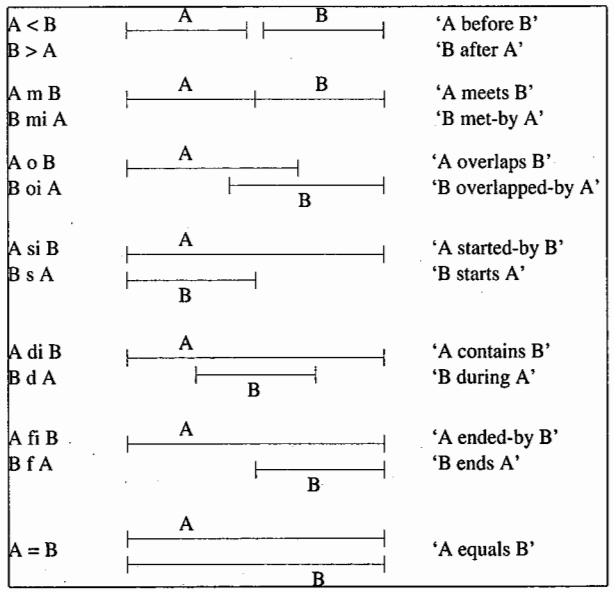


Figure 1: Allen's 13 interval-interval relations

Randell et al. (1992b) have developed interval calculus-like formalisms for the spatial domain based on Clarke's logic of connection (Clarke 1981, Clarke 1985). The basic RCC formal theory assumes a primitive dyadic relation: $C(x, y)$ read as 'x connects with y' which is defined on non-null regions. C is reflexive and symmetric. In terms of points incident in regions, $C(x, y)$ holds when regions x and y share a common boundary point. Using the relation C, a set of 8 mutually exhaustive and pairwise disjoint *base relations* are defined (in a sorted first order logic (Cohn 1987)). These relations, shown in figure 2, are *DC* (is disconnected from), *EC* (is externally connected with), *PO* (partially overlaps), *TPP* (is a tangential proper part of), *NTPP* (is a nontangential proper part of), *TPPI* (inverse of *TPP*), *NTPPI* (inverse of *NTPP*) and *EQUAL*. This set of base relations will be called *RCC-8*. In fact, more general relations such as DR (distinct region) are also defined in the general *RCC* theory, but these are expressible in terms of disjunctions of the basic relations (e.g. DR is equivalent to EC $\vee$ DC).

[1]See Allen and Hayes 1989 for a discussion on how to incorporate time points into the calculus.

(Randell, Cohn and Cui 1992a)

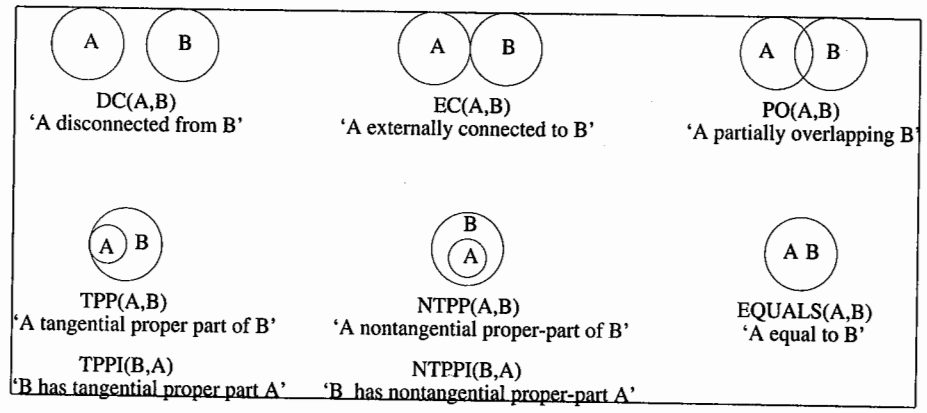


Figure 2: RCC-8 region-region relations

A more expressive calculus may be produced by introducing an additional primitive function *conv(x)* 'the convex hull of x', which is axiomatised and used to define further dyadic relations not expressible in terms of C alone. These additional relations are used to describe regions that are either inside, partially inside or outside other regions. In particular, a set of base relations *RCC-15* is defined, in which *EC* and *DC* are replaced by nine more specialised relations: *DR(O,O), DR(P,O), DR(O,P), DR(I,O), DR(O,I), DR(P,P), DR(I,P), DR(P,I), DR(I,I)*. Here, *DR(P,O)(x,y)* denotes that x and y are Distinct Regions, x is Partially overlaps *conv(y)*, and y is Outside *conv(x)*.

Just as in Allen's interval calculus, composition tables for the relation sets *RCC-8* and *RCC-15* can be determined and used to reason about spatial relationships. Moreover, many more relations making finer grained distinctions can be defined e.g. (Cohn, Randell and Cui 1994).

3 Constructing Composition Tables

The expressive power of 1st-order logic allows for straightforward axiomatisation of theories such as *RCC*. However, automated reasoning in 1st-order logic is very difficult as there is no general decision procedure. In this section we outline an alternative representation of spatial relationships based on Tarski's work with intuitionistic propositional calculus ($\mathcal{I}_0$) (Tarski 1956). This new approach provides us with a representation language which is powerful enough to capture a significant subset of what can be expressed in the RCC formalism. Furthermore, entailment in this representation is decidable. Encoding RCC relations into $\mathcal{I}_0$ provides us with an extremely efficient method for generating composition tables.

Tarski showed that the propositional calculus presented by Heyting as a formalisation of intuitionistically valid reasoning could be given a model-

theoretic interpretation in terms of *topological spaces*. Under this interpretation each formula of $\mathcal{I}_0$ denotes an open subset of a topological space, i.e. an open set of points. A model for $\mathcal{I}_0$ is a structure $\langle \mathcal{U}, i, \mathcal{P}, d \rangle$, where $\mathcal{U}$ is a non-empty set, i is a function satisfying appropriate axioms (see e.g. (Kuratowski 1972, p.129)), which maps subsets of $\mathcal{U}$ to their interiors, $\mathcal{P}$ is a denumerably infinite set of propositional constants, and d is a denotation function which assigns to each constant in $\mathcal{P}$ an *open* subset of $\mathcal{U}$. The domain of d is extended to all $\mathcal{I}_0$ formulae formed from these variables by stipulating that:

1. $d(\sim P) = i(\overline{d(P)})$
2. $d(P \wedge Q) = d(P) \cap d(Q)$
3. $d(P \vee Q) = d(P) \cup d(Q)$
4. $d(P \Rightarrow Q) = i(\overline{d(P)} \cup d(Q))$

where for any set S, $\overline{S}$ is the set of all elements of $\mathcal{U}$ which are not elements of S.

The central theorem of Tarski's paper is that under this interpretation, all valid formulae of $\mathcal{I}_0$ have the value $\mathcal{U}$ whatever the values assigned to atomic propositions. On the other hand if we take $\mathcal{I}_0$ formulae as *asserting* that their denotation is $\mathcal{U}$, they can be regarded as constraints on possible models: each formula picks out those assignments (of open sets to propositional letters) for which the formula has the value $\mathcal{U}$. It can be shown *via* Tarski's theorem that entailments between $\mathcal{I}_0$ formulae correspond to entailments between the constraints which they express.

It turns out that all the relations in *RCC-8* can be expressed in terms of the presence and absence of constraints expressible in $\mathcal{I}_0$. The basis of the interpretation is as follows:

- A *region* is identified with an open set of points. (So regions are denoted by propositional letters in the $\mathcal{I}_0$ representation.)
- Regions *overlap* if they share at least one point.
- Regions are *connected* if their *closures* share at least one point. (The closure of a region X is equal to $\overline{i(\overline{X})}$.)

This interpretation is in accord with that suggested for the *RCC* theory in (Randell et al. 1992b).

In general, to represent the RCC relations, we need to specify not only that certain constraints hold but also that other constraints do not hold. Thus we use an extension of $\mathcal{I}_0$, which we call $\mathcal{I}_0^+$. An $\mathcal{I}_0^+$ representation of a topological situation is simply a pair of sets of $\mathcal{I}_0$ formulae — one set representing positive constraints (called model constraints) and the other negative constraints (called entailment constraints — this is because of their role in the reasoning algorithm described below). Amongst the negative (entailment) constraints we always have for each region X involved in a situation, the requirement that $\sim X$ does not hold, since such a formula can only denote $\mathcal{U}$ if the region X is the null-region. We enforce that all regions

must be non-null, otherwise situations which are intuitively impossible become possible (in fact without this stipulation, according to the $\mathcal{I}_0^+$ reasoning algorithm all situations would be possible as long as all the regions involved were null). Table 1 gives the $\mathcal{I}_0^+$ representation for each of the RCC-8 basic relations:

Relation	*Model Constraint*	*Entailment Constraints*
DC(X,Y)	$\sim X \vee \sim Y$	$\sim X,\ \sim Y$
EC(X,Y)	$\sim(X \wedge Y)$	$\sim X \vee \sim Y, \sim X,\ \sim Y$
PO (X,Y)	—	$\sim(X \wedge Y),\ X \Rightarrow Y,\ Y \Rightarrow X, \sim X, \sim Y$
TPP(X,Y)	$X \Rightarrow Y$	$\sim X \vee Y,\ Y \Rightarrow X, \sim X, \sim Y$
TPPI(X,Y)	$Y \Rightarrow X$	$\sim Y \vee X,\ X \Rightarrow Y, \sim X, \sim Y$
NTPP(X,Y)	$\sim X \vee Y$	$Y \Rightarrow X, \sim X, \sim Y$
NTPPI(X,Y)	$\sim Y \vee X$	$X \Rightarrow Y, \sim X, \sim Y$
EQUAL(X,Y)	$X \Leftrightarrow Y$	$\sim X, \sim Y$

Table 1: $\mathcal{I}_0^+$ representation for *RCC-8* relations

That the model constraints given in this table must hold if the corresponding RCC relation holds is easily verified by considering the interpretation of the formulae given above. The set of entailment constraints serves to exclude certain models. For example the model constraint $\sim(X \wedge Y)$ holds if either $EC(X,Y)$ or $DC(X,Y)$ holds; but adding the formula $\sim X \vee \sim Y$ as an entailment constraint for $EC(X,Y)$ serves to prohibit $DC(X,Y)$.

The $\mathcal{I}_0^+$ representation provides a decision procedure for testing the consistency of topological situation descriptions. Given a situation description consisting of a set of relations of the form $R(A,B)$, where R is one of the relations characterisable in $\mathcal{I}_0^+$, and A and B are constants denoting regions, the following simple algorithm will decide whether the description describes a possible situation:

- For each relation $R_i(\alpha_i, \beta_i)$ in the situation description find the corresponding $\mathcal{I}_0^+$ representation $\langle \mathcal{M}_i, \mathcal{E}_i \rangle$.
- Construct the overall $\mathcal{I}_0^+$ representation $\langle \bigcup_i \mathcal{M}_i,\ \bigcup_i \mathcal{E}_i \rangle$.
- For each formula $F \in \bigcup_i \mathcal{E}_i$ use an intuitionistic theorem prover to determine whether the entailment $\bigcup_i \mathcal{M}_i \vdash_{Io} F$ holds.
- If any of the entailments determined in the last step does hold then the situation is impossible.

For example we may want to know whether the following situation is possible: A, is a non-tangential proper part of B; B is externally connected to C; and, A is a tangential proper part of C. The $\mathcal{I}_0^+$ representations of the three spatial relations are:

NTPP(A,B): $\quad\langle\{\sim A \vee B\},\{B \Rightarrow A, \sim A, \sim B\}\rangle$,
EC(B,C): $\quad\langle\{\sim(B \wedge C)\},\{\sim B \vee \sim C, \sim B, \sim C\}\rangle$,
TPP(A,C): $\quad\langle\{A \Rightarrow C\},\{\sim A \vee C, C \Rightarrow A, \sim A, \sim C\}\rangle$

so the overall $\mathcal{I}_0^+$ representation is

$$\langle\{\sim A \vee B, \sim(B \wedge C), A \Rightarrow C\},\ \{B \Rightarrow A, \sim B \vee \sim C, \sim A \vee C, C \Rightarrow A, \sim A, \sim B, \sim C\}\rangle.$$

We determine that this situation is impossible since

$$\sim A \vee B, \sim(B \wedge C), A \Rightarrow C \ \vdash_{Io}\ \sim A.$$

A full explanation of this method and proof of its correctness can be found in (Bennett 1994).

Consistency checking in $\mathcal{I}_0^+$ enables compositions of spatial relations to be computed very efficiently. Given R_1 and R_2, which are members of some basis set $\mathcal{B}$, one simply checks for all values of R_3 taken from $\mathcal{B}$, whether the situation described by $R_1(A,B)$, $R_2(B,C)$, $R_3(A,C)$ is possible. A composition table for *RCC-8* was computed in under 142 seconds on a Sparc1 workstation. This method has also been successfully applied to much larger sets of basic relations obtained by subdividing the *RCC-8* set to take account of finer distinctions among spatial relationships e.g. *RCC-15*.

4 Conceptual Neighbourhoods

In the real, dynamic world spatial and temporal relationships between objects may change from one situation to the next. For example, considered 2-dimensionally, the spatial region occupied by a pebble on a beach (we denote the pebble by P) will be disconnected from the region occupied by the sea (S) at low tide. As the tide rises S expands and its boundary approaches P until the two regions touch i.e. S and P become externally connected. As the tide continues to rise the regions partially overlap and then P becomes first a tangential proper part and then a nontangential proper part of S. We can directly represent this process as a sequence of *RCC-8* relations:

$$DC(S,P) \Rightarrow EC(S,P) \Rightarrow PO(S,P) \Rightarrow TPP(S,P) \Rightarrow NTPP(S,P)$$

One can easily imagine alternative *RCC-8* sequences that could be used to represent other real-world processes. However, it is not the case that a direct transition exists between every pair of relations. For example, in the physical world, two initially disconnected regions cannot subsequently overlap unless they have first come into external contact i.e. there can be no direct transitions from *DC* to *PO*. We call the graph of all possible direct transitions from one relation to another the *transition graph*. Connected pairs in this graph are called *conceptual neighbours*, a term first used by Freksa (1992) with respect to Allen's interval calculus. A *conceptual neighbourhood* is defined as any connected subgraph of the transition graph (including single nodes and the full graph itself).

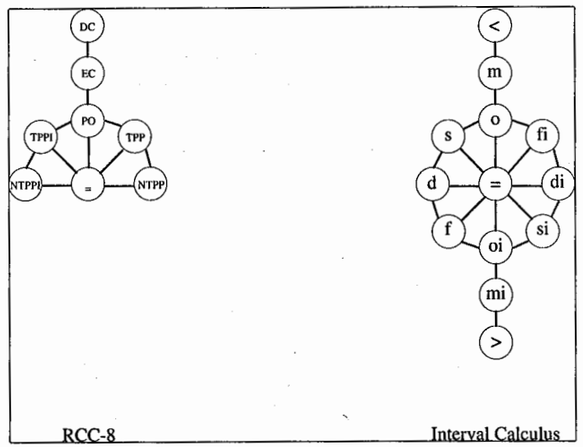


Figure 3: Conceptual neighbourhoods for two calculi

It comes as no surprise to find that the transition graph for Allen's interval calculus is very similar to that of *RCC-8* (see figure 3). This is because *RCC-8* can be thought of as a simplified version of interval calculus: *RCC-8* does not contain any spatial equivalent of temporal ordering hence it is asymmetric. Extending the calculus by introducing the notion of left/right would produce relations opposite of *DC,EC,PO,TPP* and *TPPI* resulting in a spatial version of interval calculus with a (symmetric) transition graph similar to that of Allen's formalism.

Transition graphs describe all the physically possible transitions between relations that can occur when intervals are deformed by continuous expansion, contraction and translation. Restricting the ways in which an interval can be deformed produces alternative graphs. Freksa identified three basic kinds of deformation that can occur in interval calculus. B-deformation occurs when the duration of an interval remains constant but its position on the time line alters. C-deformation corresponds to uniform expansion or contraction of an interval about its centre-point. A-deformation is characterized by expansion or contraction of an interval with one end point fixed. Individually, these give rise to restricted transition graphs shown in figure 4 which Freksa referred to as A, B and C-neighbourhoods. Collectively, they form the full transition graph. Spatial equivalents of the three kinds of deformation can be found for the *RCC* calculi. A-deformation of a spatial region can be viewed as an increase or decrease in the region's area with shape and the position of at least one point on the region's boundary remaining unchanged. The spatial version of B-deformation is the movement

of a region in space with area and shape unchanged. C-deformation refers to a change in area with centre point and shape kept constant. The graphs that follow from these definitions are exact analogies of Freksa's A, B and C neighbourhoods for interval calculus.

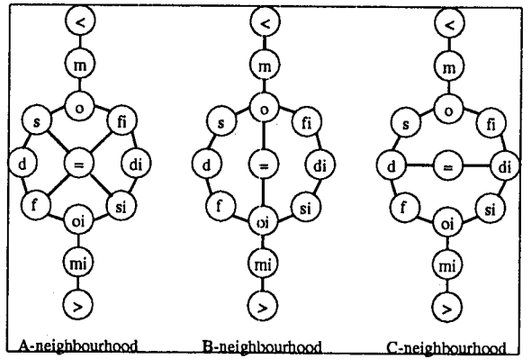


Figure 4: A,B and C neighbourhoods

It has become apparent that the relationship between conceptual neighbourhoods and composition tables is very strong. Freksa (1992) has pointed out that every entry in the composition table for Allen's interval calculus defines a conceptual neighbourhood. Furthermore, although there are 169 table entries only 29 of these are different. This is a surprising result considering that the transition graph for Allen's calculus gives rise to 1255 different conceptual neighbourhoods. Similar properties hold for the *RCC-8* composition table here we find only 21 different entries, all of which also form conceptual neighbourhoods. This has led to two separate lines of research (1) investigation of new composition tables based on conceptual neighbourhoods rather than individual relations and (2) methods of directly generating transition graphs from the data in composition tables.

5 Constructing Composition Tables From Neighbourhoods

Freksa has investigated the possibility of replacing the interval calculus composition table by one that describes transitivity properties between conceptual neighbourhoods (where at least some of these neighbourhoods are non singletons). He argued that as conceptual neighbourhoods tend to be disjunctions of basic relations they capture the uncertainty prevalent in the everyday world. The main difficulty in constructing such neighbourhood-based tables is how to select an appropriate set of neighbourhoods upon which to base the table. The interval calculus contains 1255 different conceptual neighbourhoods so it is impractical to investigate all possible combinations in search of the 'best' set. Freksa recognised this problem and suggested plausible criteria that might be used to reduce the search space

(Freksa 1993).

1. Each of the basic relations in the original calculus should be readily obtainable, either by their direct inclusion in the new set or from the intersection of two of its elements. This ensures that it is possible to use the new table to reason about basic relations with a minimum of computational overheads.
2. The set should consist only of neighbourhoods that are also entries from the original transitivity table. The justification for this is that these neighbourhoods are in some way more important than others as they freely occur even in the original table.

The second criterion limits the search to the 31 neighbourhoods that occur in Allen's table, drastically reducing the search space. Using a specially developed hypercard 'thinktool' Freksa was able to identify a set of 10 neighbourhoods from these that also satisfied the first criterion (shown as the first entry in table 2). This resulting composition table based on these contained only 10×10 entries a 40% reduction in size compared with Allen's original. We have formally verified that Freksa's solution is minimal and have determined a further eight minimal solutions (last eight entries of table 2). However, it is hard to say which, if any, of the sets is the best. Freksa's

{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{o, fi, di}	{di, si, oi}	{o, s, d}	{d, f, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{<, m, o, fi, di}	{d, f, oi, mi, >}	{di, si, oi}	{o, s, d}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{<, m, o, fi, di}	{di, si, oi, mi, >}	{o, s, d}	{d, f, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{<, m, o, fi, di}	{di, si, oi}	{o, s, d}	{d, f, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{d, f, oi, mi, >}	{<, m, o, s, d}	{o, fi, di}	{di, si, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{d, f, oi, mi, >}	{o, fi, di}	{di, si, oi}	{o, s, d}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{<, m, o, s, d}	{di, si, oi, mi, >}	{o, fi, di}	{d, f, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{<, m, o, s, d}	{o, fi, di}	{di, si, oi}	{d, f, oi}
{<}	{m}	{mi}	{>}	{si, =, s}	{fi, =, f}	{di, si, oi, mi, >}	{o, fi, di}	{o, s, d}	{d, f, oi}

Table 2: Minimal sets of conceptual neighbourhoods for Allen's interval calculus

solution has the property that it uses the smallest conceptual neighbourhoods (each contains three or less disjuncts) but this is not necessarily an advantage as larger neighbourhoods may capture more general information about uncertainty. Furthermore, none of the solutions appear to be any less 'cognitively plausible' than any of the others.

We have recently applied Freksa's technique to *RCC-8* and *RCC-15* in an attempt to produce reduced tables. Using a low-level C language program in which neighbourhoods were expressed (and manipulated) as bit vectors we were able to perform an exhaustive analysis of *RCC-8*. Our results showed that Freksa's approach does not produce a compact neighbourhood-based table for this formalism, which seems surprising considering the similarity between this and Allen's interval calculus. The main reason for failure is that Freksa's second criterion rules out a great number of conceptual neighbourhoods that would otherwise be considered as viable candidates. Those

that remain are simply too few in number and not sufficiently different to one another to yield a smaller table in accordance with the first criterion.

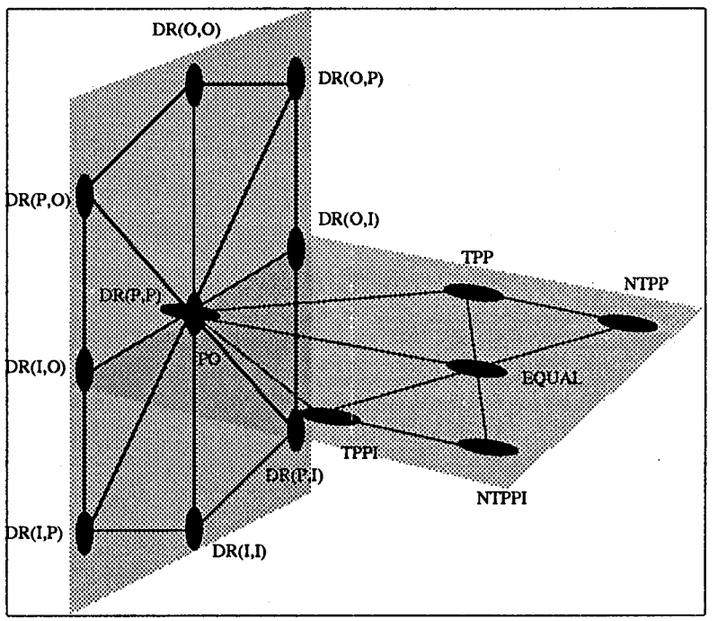


Figure 5: Transition graph for RCC-15

RCC-15 has a very much larger composition table than both *RCC-8* and Allen's interval calculus 225 entries, of which 66 are unique. Even after applying Freksa's second criterion it is not possible to exhaustively check each potentially suitable set of neighbourhoods against the first criterion as there are $\approx 10^{14}$ sets with 15 or less elements. Fortunately, the conceptual neighbourhood structure of *RCC-15* lends itself readily to simplification. Figure 5 shows how *RCC-15* can be viewed as consisting of two separate conceptual neighbourhoods one for the *DR* relations (vertical plane) and one for *PO,TPP, TPPI, NTPP, NTPPI* and *EQUAL* (horizontal plane) that connect through *PO* (PO is connected to every DR relation). We applied Freksa's technique to each of these separately and combined the resulting minimal solutions. This yielded a set of 13 neighbourhoods which could be used as the basis of a 169 entry neighbourhood-based composition table. Although this represents a 25% reduction in table size compared with the original, this is still small compared with the 40% reduction Freksa obtained with Allen's interval calculus.

The results of applying Freksa's approach to *RCC-8* and *RCC-15* call into question Freksa's second criterion: neighbourhoods used as relations

in the new tables can only be chosen from those that appear as entries in the original composition table. In fact, if this requirement is relaxed then 24 alternative eight-neighbourhood solutions are obtained for *RCC-15*, each representing a 75% reduction in table size! One such solution is illustrated in figure 6. Similarly, a number of 6×6 solutions can be obtained for *RCC-8*,

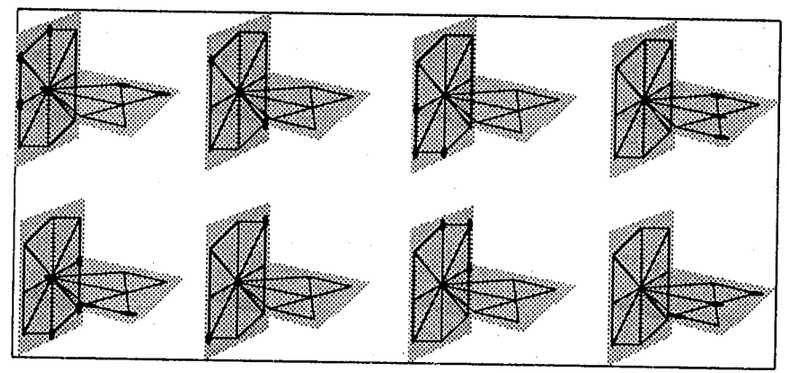


Figure 6: Conceptual neighbourhood set for RCC-15 compact table

representing a 44% reduction in table size.

Compacted composition tables are ideal for use in a multi-processor environment. Although the tables are much smaller than the non-compacted originals, it is often necessary to perform more look-up operations with the new tables. For example, the table entry for two basic relations can be obtained via a single look-up operation using a non-compacted table. However, for the compact table basic relations are generally obtained by intersecting two of the neighbourhood relations upon which the table is based, so four look-ups are required. However, if a number of simple processing elements are available, the look-ups can be performed simultaneously leading to a considerable improvement in efficiency overall. This is particularly noticeable with large tables and such an approach will make implementation of an enhanced spatial calculus (with at least a 100×100 composition table) described in (Cohn et al. 1994) practical. Yet larger relation calculi could easily be generated by considering, e.g., orientation information (Zimmermann and Freksa 1993).

6 Generating Transitivity Graphs From Composition Tables

Although most current spatial/temporal calculi are relatively simple and contain few relations, it is anticipated that more complex formalisms with far greater expressive capabilities will be used in future (we have already begun work on a 23-relation spatial calculus). It is often a tedious and surprisingly difficult task to produce transition graphs for such calculi. A solution to this problem is to generate the graphs directly from information

stored in composition tables.

One approach to constructing transition graphs is to generate all possible graphs and test these for suitability against constraints derived from the composition table. It is believed that every composition table entry forms a conceptual neighbourhood. Conceptual neighbourhoods can be viewed as representing path linked nodes in the transition graph and this enables us to test potential solutions to ensure that they only contain paths corresponding to these and that all such paths are represented. Unfortunately, for all but the simplest calculi it is infeasible to generate and test every possible graph. A practical alternative is to use constraint-based reasoning techniques to prune the search space. We start by labelling a node with each relation (singleton entry in the composition table). Next, we take each two-relation composition table entry in turn and connect the corresponding nodes. For each three-relation entry, if the relations are not path connected then we connect the appropriate nodes. This process is continued for four-relation entries etc. If we apply this technique to RCC-8 we obtain the three solutions shown in figure 7. Combining the three neighbourhoods gives the transition graph corresponding to B and C deformation. Although we are unable to obtain the two missing links (corresponding to A deformation) it is nevertheless encouraging that an almost complete transition graph can be generated automatically.

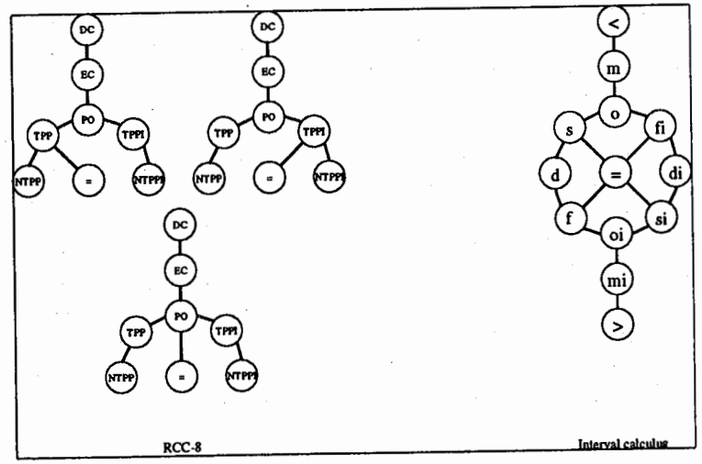


Figure 7: Transition graphs derived from composition tables

Applying the technique to interval calculus gives a total of 36 solutions, many of which contain links that are incorrect. Clearly, additional constraints are required. One possibility is to consider the degrees of freedom associated with a relation. In interval calculus, if $x < y$ holds then it can also hold (in at least some cases) when x or y are either A, B or C deformed. We say that $<$ has 3 *degrees of deformation*. *x meets y* has only one degree of deformation as it will not hold if either x or y are B or C deformed. Only under A deformation can it remain true. Adding the constraint that two graph nodes can only be linked if they have different degrees of deformation produces a single, interval calculus transition graph corresponding to combined B and C deformation. In fact, degree of deformation corresponds to Galton's notion of 'dominance' (Galton 1993), an alternative approach to describing the constraints on transitions between relations in a calculus.

7 Further Work and Conclusions

We intend to apply these compaction and efficiency enhancement techniques to a wider variety of spatial and temporal calculi with a view to confirming the utility of these techniques in logics with a large number of base relations. Decidable calculi to model these languages also need to be developed.

To conclude: we have presented an efficient method for generating composition tables and shown how Freksa's table reduction technique, originally intended for Allen's interval calculus, can be applied to RCC-8 and RCC-15. In doing this we verified Freksa's result and showed that Freksa's second criterion was not appropriate and, in some cases, far greater reductions in table size could be achieved by ommitting it. Finally, we showed that it is possible to construct the B/C deformation transition graphs for RCC-8 and Interval Calculus directly from information in the composition tables.

References

Allen, J. F.: 1981, An interval-based representation of temporal knowledge, *Proceedings 7th IJCAI*, pp. 221–226.

Allen, J. F. and Hayes, P. J.: 1989, Moments and points in an interval-based temporal logic, *Computational Intelligence* **5**, 225–238.

Allen, J. F., Kautz, H. A., Pelavin, R. N. and Tenenberg, J. D.: 1991, *Reasoning About Plans*, Morgan Kaufmann, Sab Mateo.

Bennett, B.: 1994, Spatial reasoning with propositional logics, *in* J. Doyle, E. Sandewall and P. Torasso (eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference (KR94)*, Morgan Kaufmann, San Franciscon, CA. To appear.

Clarke, B. L.: 1981, A calculus of individuals based on connection, *Notre Dame Journal of Formal Logic* **23**(3), 204–218.

Clarke, B. L.: 1985, Individuals and points, *Notre Dame Journal of Formal Logic* **26**(1), 61–75.

Cohn, A. G.: 1987, A more expressive formulation of many sorted logic, *Journal of Automated Reasoning* **3**, 113–200.

Cohn, A. G., Randell, D. A. and Cui, Z.: 1994, Taxonomies of logically defined qualitative spatial relations, *in* N. Guarino and R. Poli (eds), *Formal Ontology in Conceptual Analysis and Knowledge Representation*, Kluwer. To appear.

Davis, E.: 1990, *Representations of commonsense knowledge*, Morgan Kaufmann, San Mateo.

Freksa, C.: 1992, Temporal reasoning based on semi-intervals, *Artificial Intelligence* **54**, 199–227.

Freksa, C.: 1993, Personal communication.

Galton, A. P.: 1993, Perturbation and dominance in the qualitative representation of continuous state-spaces, Submitted for publication.

Kuratowski, K.: 1972, *Introduction to Set Theory and Toplolgy*, 2nd edn, Pergamon Press.

Randell, D. A., Cohn, A. G. and Cui, Z.: 1992a, Computing transitivity tables: A challenge for automated theorem provers, *Proceedings CADE 11*, Springer Verlag, Berlin.

Randell, D. A., Cui, Z. and Cohn, A. G.: 1992b, A spatial logic based on regions and connection, *Proceedings 3rd International Conference on Knowledge Representation and Reasoning*, Morgan Kaufmann, San Mateo, pp. 165–176.

Tarski, A.: 1956, Sentential calculus and topology, *Logic, Semantics, Metamathematics*, Oxford Clarendon Press, chapter 17. trans. J.H. Woodger.

Vieu, L.: 1991, *Sémantique des relations spatiales et inférences spatio-temporelles*, PhD thesis, Université Paul Sabatier, Toulouse.

Zimmermann, K. and Freksa, C.: 1993, Enhancing spatial reasoning by the concept of motion, *in* A. Sloman (ed.), *Prospects for Artificial Intelligence*, IOS Press, pp. 140–147.

On the Boundary Between Mereology and Topology

Achille C. Varzi

1 Introduction

Much recent work aimed at providing a formal ontology for the commonsense world has emphasized the need for a mereological account to be supplemented with topological concepts and principles. There are at least two reasons underlying this view. The first is truly metaphysical and relates to the task of characterizing individual integrity or organic unity: since the notion of connectedness runs afoul of plain mereology, a theory of parts and *wholes* really needs to incorporate a topological machinery of some sort. The second reason has been stressed mainly in connection with applications to certain areas of artificial intelligence, most notably naive physics and qualitative reasoning about space and time: here mereology proves useful to account for certain basic relationships among things or events; but one needs topology to account for the fact that, say, two events can be continuous with each other, or that something can be inside, outside, abutting, or surrounding something else.

These motivations (at times combined with others, e.g., semantic transparency or computational efficiency) have led to the development of theories in which both mereological and topological notions play a pivotal role. How exactly these notions are related, however, and how the underlying principles should interact with one another, is still a rather unexplored issue. One can see mereology and topology as two independent chapters; or one may grant priority to topology and characterize mereology derivatively, defining parthood in terms of connection; or, again, one may privilege mereology and explain connection in terms of parthood *and* other predicates or relations. It is also possible, on some assumptions, to develop a unified framework based on a single mereo-topological primitive of connected parthood. The purpose of this paper is to offer a first assessment of these alternative routes, discussing their relative merits and examining to what extent their adequacy, and more generally the boundary between mereology and topology, depends on the ontological fauna that one is willing to countenance.

I am grateful to George Bealer, Tony Cohn, Nicola Guarino and Barry Smith for helpful discussion and valuable comments on earlier drafts.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 423-442

2 The Bounds of Mereology

Mereology is by definition concerned with parts that is, with the relation holding between two things when one is part of the other. On a weak understanding this means that a mereological theory is first and foremost an attempt to explicate the meaning of the word 'part' and to set out the principles underlying our correct use of it, and of kindred notions. For instance, virtually every mereological theory agrees on treating parthood as a partial ordering, which in a way reflects some very basic meaning postulates for 'part'.[1] Here, however, I am interested in the stronger interpretation, according to which mereology may provide a fundamental framework for the task of ontological investigations. It is a view that influenced much Greek and scholastic philosophy, and that made its way into modern philosophy *via* Husserl's third *Logical Investigation*.[2] Mereology tells us how reality is *constituted*. In this sense not just any partial ordering will qualify as a part-relation, and the question of what additional principles are involved becomes a truly philosophical (as opposed to merely terminological) question. Modern formal systems of mereology also owe their birth to this view, regardless of whether the relation 'x is (a) part of y' is taken as a primitive (as in Leśniewski's mereology) or defined in terms of cognate relations such as, for instance, 'x extends over y' (Whitehead's *Enquiry*), 'x is disjoint from y' (Leonard and Goodman's calculus of individuals), or 'x overlaps y' (Goodman's *Structure of Appearance*).[3] Sometimes this has been associated with a nominalistic stand and mereology has been presented as a parsimonious alternative to set theory, dispensing with abstract entities or, better, treating all entities as individuals. However there is no necessary internal link between mereology and nominalism. Mereology can be credited a fundamental ontological role whether or not we take the entire universe to be describable in terms of parthood relationships. I think the recent revival of mereology and its ascent in artificial intelligence and other cognitive sciences can be seen in this light. The question is, rather, how far we can go with it how much of the universe can be grasped and described by means of purely mereological notions.

It is in this perspective that the limitations mentioned at the beginning become relevant, particularly if our concern is the ontology of the macroscopic, common-sense world. Our common-sense picture of reality requires some means of distinguishing between things that are all of a piece and things that are scattered in space or time, or between things that are continuous and things that are not. Yet it is not clear how this can be done mereologically,

[1]There are exceptions. In particular, the transitivity of the part-relation has been disputed at least since Rescher 1955. See Cruse 1979, Winston *et al.* 1987, and the recent plea for naive mereology in Sanford 1993.

[2]Husserl 1901. On the role of mereology in ancient and scholastic philosophy, see e.g. Burkhardt and Dufour 1991, Henry 1991.

[3]See Leśniewski 1916, Whitehead 1919, Leonard and Goodman 1940, and Goodman 1951, respectively. For a thorough overview see Simons 1987.

starting from the relational concept of part (or overlapping, disjointness, and the like). In spite of the natural tendency to present mereology as a theory of parts *and* wholes, wholeness cannot be explained in terms of parthood, hence of mereology, except in the trivial sense that everything qualifies as the complete whole of its parts. Moreover, according to classical mereology every class of parts determines a complete whole (its mereological sum, or fusion), which makes the latter an utterly ineffective notion.

This latter point is not in itself undisputed. Avoiding explicit reference to classes, the underlying principle is that every satisfied property or condition picks out a unique entity consisting of all things satisfying that property or condition. It is usually expressed as follows:[4]

$$\exists x \phi x \rightarrow \exists x \forall y (\mathsf{O}yx \leftrightarrow \exists z (\phi z \wedge \mathsf{O}yz)) \qquad (1)$$

where 'O' stands for the relation of overlapping (i.e., sharing a common part). This principle is probably the most commonly criticised feature of classical mereology, the usual objection being that it has counter-intuitive instances, i.e., "unnatural" sums of widely scattered, disparate, unrelated, or otherwise ill assorted entities, such as the totality of red things, or my eyebrows and your favorite Chinese restaurant.[5] The classical mereologist's reply is simply that the criticism is off target, and I go along with that. If you already have some things, allowing for their sum is no further commitment: the sum *is* those things.[6] One may feel uncomfortable with treating unheard-of Goodmanian mixtures as individual wholes. But it is not a task of mereology to specify which wholes are more natural than others. In effect, telling which entities constitute natural wholes is presumably not even a metaphysical task, but the concern of empirical sciences[7] (just as ascertaining which sentences are true is not a semantic task but an empirical issue). The real source of difficulty, as I see it, is different. It is that the question of what constitutes a natural whole cannot even be *formulated* in mereological terms. As soon as we allow for the possibility of scattered entities we lose the possibility of discriminating them from integral, connected

[4]Some classical systems. such as Tarski 1929 or Leonard and Goodman 1940, give a formulation of the principle involving reference to classes of individuals rather than just using predicates or open formulas. Here I stick to a class-neutral formulation simply for expository convenience. The difference is nonetheless to be noted, since an ordinary first-order language has a denumerable supply of predicates or formulas, so that at most denumerably many classes (in any given universe of discourse) can be specified. For the nominalist this limitation is of course negligible insofar as classes do not exist except as nomina. Compare Eberle 1970, especially pp. 67f.

[5]See, for example, the early criticisms of Lowe 1953, Rescher 1955, or Chisholm 1976. Of course a similar complaint arises in set theory, as discussed in Smith 1991.

[6]The *locus classicus* is Goodman 1956, 1958. For a recent statement see Lewis 1991, who stresses that "it is in virtue of this thesis that mereology is ontologically innocent: it commits us only to things that are identical, so to speak, to what we were committed to before" (p. 82); the "so to speak" is explained as in Baxter 1988a, 1988b.

[7]This point is made in Simons 1982: 149.

wholes; but of course we cannot just keep the latter without some means of discriminating them from the former.

Whitehead's early attempts to characterize his ontology of events (the primary natural entities of the *Enquiry*) provide a good exemplification of this mereological dilemma.[8] Whitehead's system does not satisfy (1), for the intended domain is one which excludes scattered or disconnected entities (that is, events). Of course it is not maintained that there are *no* mereological sums. Rather, the suggestion is that a necessary condition for two events to have a sum is that they be "joined" to each other. This relation of joining does not coincide with overlapping (for otherwise no event could be dissected into separated proper parts), and it is explicitly considered that two events may be *ad*joined, i.e., joined without sharing any common parts. Joining is thus a more general notion than overlapping: it is intuitively meant to hold whenever two events are continuous with each other, be they discrete or not. Whitehead eventually defines it along the following lines:

$$\mathsf{J}xy =_{\mathrm{df}} \exists z(\mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{P}wz \rightarrow \mathsf{O}wx \vee \mathsf{O}wy)) \qquad (2)$$

where 'P' indicates parthood. Now, this definition does indeed say that two events are joined just in case their mereological sum exists. But precisely for this reason, it is immediately verified that (2) falls well short of capturing the intended notion of topological connectedness or continuity. For there is nothing to guarantee that the piece overlaying two joined events be itself connected. The pattern reproduced in Figure 1, with two disconnected discs x and y partially overlapping a disconnected piece z, illustrates a simple counter-example.[9]

These considerations apply *mutatis mutandis* to other attempts to subsume topological connectedness within a bare mereological framework.[10] Of course one can succeed if the assumption is made that only self-connected entities can inhabit the domain of discourse. This would indeed support a restricted conception of wholeness in which one would deny (1) in its generality, and in which a plurality of entities can be said to make up an integral whole just in case they have a sum (or at least an upper bound relative to the part-relation[11]). But this is no satisfactory way out, for it just is not possible to make the assumption explicit. If the lack of any specific notion of whole is indicative of the neutrality of mereology, and hence of its strength and generality,[12] it is a fact that this lack is in turn a lack of expressiveness, hence a sign of weakness too.

[8] Whitehead 1919. The definition of joining given below is actually from Whitehead 1920, but the difference from the earlier account is inessential.

[9] The figure is adapted from Simons 1987: 337, where a similar point is made (compare the discussion at pp. 81-86). See also Simons 1991.

[10] See Needham 1981.

[11] See Bostock 1979.

[12] As recently emphasized by Eschenbach and Heydrich 1993: 207.

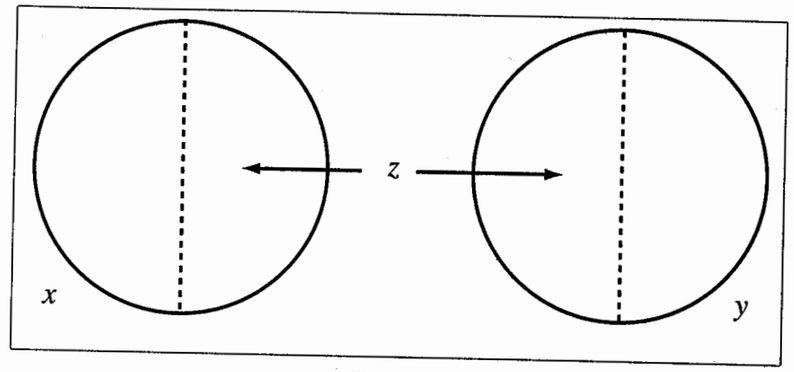


Figure 1

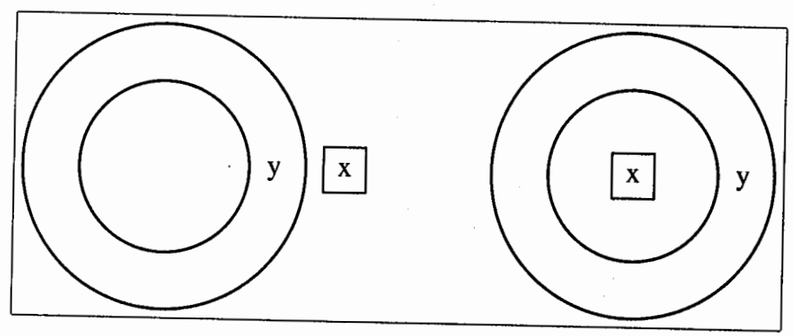


Figure 2

Nor is this exclusively a metaphysical concern. This deficiency of mereology has shown up in various ways in recent work in linguistics, knowledge representation, and qualitative reasoning in artificial intelligence.[13] In all of these contexts mereology is now credited a central role in accounting for certain fundamental relationships among the entities in the domain of discourse – be they events, spatial regions, physical objects, or what have you. But as I already anticipated, these contexts also tend to confirm the limits of mereology when it comes to accounting for relations that entail a step into the territory of topology. Mereologically the two situations depicted in Figure 2 involve no difference, though of course we may want to keep track of the basic opposition in terms of spatial inclusion of the square, x, in the annulus, y.

The same difficulty arises when we consider relations among things

[13] I shall give detailed references in the next section.

that are just touching each other, or straddling one another, or neighbouring other things. All of these – and, indeed, many others – are relations that any theory concerned with spatio-temporal entities should supply, but which cannot be defined directly in terms of plain mereological primitives.

3 Three Ways to Topology

This need to overcome the bounds of mereology has been handled in the literature in various ways, but I think three main strategies can be distinguished. The first is, in a sense, the obvious one: if topology cannot be made to fit mereology, and if its importance is to be fully recognized, then we may just *add* it to a mereological basis as an independent chapter. From this point of view, mereology can be seen as the ground theory on which theories of greater and greater complexity (including topology as well as, say, morphology or kinematics) can be built by supplying the necessary notions and principles. The second strategy is more radical: if topology eludes the bounds of mereology, we may try and turn things around: start from topology right away and define mereological notions in terms of topological primitives. From this point of view, just as mereology can be seen as a generalisation of the even more fundamental theory of identity (parthood, overlapping, and even fusion subsuming singular identity as a definable special case), likewise topology can be seen as a generalisation of mereology, where the relation of joining or connection takes over overlapping and parthood as special cases. Finally, the third strategy is a sort of vindication of mereology, building precisely on its formal generality: on this view topology is simply a domain-specific chapter of mereology, connection and kindred notions being accounted for in terms of part-relations among entities of a specified sort. I shall further scrutinize these three strategies in turn.

The First Way. The first strategy is, as I said, the most obvious and has in effect been followed by most authors (sometimes on quite independent grounds). It was pioneered by Tarski's work on the foundations of the geometry of solids (where a mereological basis is supplemented with a single primitive predicate 'x is a sphere' to allow a definition of solid geometric correlates of all ordinary point-geometric notions) and was reproposed by Lejewski in his outline of a Leśniewskian theory of time, or *Chronology* (where a mereological basis is supplemented with a primitive relation 'x is wholly earlier than y' to account for the main topological feature of time structures, viz. precedence).[14] The same approach underlies Tiles' analysis of events (where topology is introduced by means of the primitive 'x lies in the interior of y') as well as much linguistics-oriented work on time, tense, and aspect, such as Kamp's analysis of temporal reference and Bach's or

[14] See Tarski 1929 and Lejewski 1982, respectively. Lejewski also hypothesises that combining Chronology with a cognate theory of space, or Stereology, would yield a favorable framework for developing a general Kinematics. See Lejewski 1986.

Link's algebraic semantics for event structures (where the relation of overlapping defined on temporal entities is typically paired with a strict ordering of total precedence).[15] Chisholm's recent work on spatial continuity can also be viewed in this light.[16] I suppose the basic idea has been exploited in other areas, though presumably the range of actual implementations and choice of primitives is not much wider.[17] It is however the mereo-topological framework recently proposed by Smith that can be regarded as the outstanding representative of this approach, also because it is effectively much more general – and with much more far-reaching foundational ambitions – than its special-purpose precursors.[18]

Smith uses the same primitives as Tiles, namely the relations of being a part and of being an interior part. The former has a standard interpretation, while the latter is intuitively meant to hold when an entity is part of another and does not overlap its boundary. I shall not go into the details of the axiomatization: it is rather straightforward and justifies the claim that topology can conveniently be grounded on mereology rather than set theory.[19] The aim, however, is "to go further and capture mathematically certain ontological intuitions pertaining to ordinary material objects... to capture, if one will, the mathematical structures characteristic of the common-sense world".[20] Using 'P' and 'IP' to indicate parthood and interior parthood, respectively, here is for example how such basic notions as boundary ('B'), connection ('C'), or self-connectedness ('SC') are captured within the proposed theory:[21]

$$\mathsf{B}xy =_{\mathrm{df}} \forall z(\mathsf{P}zx \to \forall w(\mathsf{IP}zw \to \mathsf{O}wy \wedge \neg \mathsf{P}wy)) \quad (3)$$

$$\mathsf{C}xy =_{\mathrm{df}} \mathsf{O}xy \vee \exists z(\mathsf{O}zx \wedge \mathsf{B}zy \vee \mathsf{O}zy \wedge \mathsf{B}zx) \quad (4)$$

$$\mathsf{SC}x =_{\mathrm{df}} \forall y \forall z(\forall w(\mathsf{O}wx \leftrightarrow \mathsf{O}wy \vee \mathsf{O}wz) \to \mathsf{C}yz). \quad (5)$$

We need not for the moment discuss the intuitive adequacy of these notions. They do the job, as far as topology goes. Moreover, the resulting framework does allow one to sketch a first formulation of some basic ontological intuitions that go well beyond the repertoire of standard topology. For instance, Smith suggests a first rendering of the Brentanian thesis that boundaries are dependent things, i.e., can only exist as boundaries *of* something[22] (contrary to the set-theoretic conception of boundaries as sets

[15] See Tiles 1981, Kamp 1979, Bach 1986, Link 1987, and van Benthem 1983.

[16] Chisholm 1992/93.

[17] I myself have recently followed this approach in joint work with R. Casati on the metaphysics of holes and holed things: see Varzi 1993, Casati and Varzi 1994.

[18] Smith 1992, 1993.

[19] Following Menger 1940.

[20] Smith 1993: 61.

[21] I deviate slightly – but inessentially – from the original notation and formulation to avoid unnecessary intermediate definitional detours.

[22] See Brentano 1976. Of course a full statement of the thesis requires further work, as is shown in White 1993. See also Chisholm 1984, 1989 (ch. 8).

of ordinary, ontologically independent points):

$$\exists y \mathsf{B}xy \rightarrow \exists z \exists w(\mathsf{B}xz \wedge \mathsf{P}xz \wedge \mathsf{IP}wz) \tag{6}$$

or, more strictly, that self-connected boundaries are boundaries of self-connected wholes:

$$\mathsf{SC}x \wedge \exists y \mathsf{B}xy \rightarrow \exists z \exists w(\mathsf{SC}z \wedge \mathsf{B}xz \wedge \mathsf{P}xz \wedge \mathsf{IP}wz). \tag{7}$$

On the other hand, one thing to be noticed is that much of this involves a conceptual detour that could effectively be avoided. We could just assume as primitive the very notion of a boundary (which we are actually to presuppose in the intuitive interpretation of interior parthood),[23] or the relation of connection, or even the property of being self-connected, and then define interior parts accordingly – as by the following general equivalences:

$$\mathsf{IP}xy \leftrightarrow \mathsf{P}xy \wedge \forall z(\mathsf{P}zx \rightarrow \neg \mathsf{B}zy) \tag{8}$$

$$\mathsf{IP}xy \leftrightarrow \mathsf{P}xy \wedge \forall z(\mathsf{C}zx \rightarrow \mathsf{O}zy) \tag{9}$$

$$\mathsf{IP}xy \leftrightarrow \mathsf{P}xy \wedge \forall z(\forall w(\mathsf{SC}w \wedge \mathsf{O}wy \wedge \neg \mathsf{P}wy \rightarrow \mathsf{O}wz) \rightarrow \neg \mathsf{O}xz. \tag{10}$$

I shall indeed come back to this point, for I think this is where a lesson is to be drawn. First, however, I shall move on to considering in greater detail the other two strategies mentioned above.

The Second Way. The second way of bridging the gap between mereology and topology exploits the intuition that the latter is truly a more basic and more general framework subsuming the former in its entirety, at least relative to certain domains. This view can be traced back to De Laguna's work on solid geometry (based on the primitive relation 'x connects y to z') and was taken over by Whitehead himself in the final version of his theory in *Process and Reality* (where all notions are explained in terms of the single topological primitive 'x is extensionally connected with y').[24] The approach was fully worked out by Clarke in his resourceful reformulation of the calculus of individuals and has recently been employed by Randell, Cui and Cohn for work in spatio-temporal reasoning and naive physics, and by Aurnague and Vieu for the analysis of spatial prepositions in natural language.[25] To my knowledge, not many other developments or applications have been put forward, if we exclude the interval logics for the representation of time – based on a primitive relation of temporal precedence – which have

[23] Smith himself considers the possibility of using a primitive closure operation.

[24] See De Laguna 1922 and Whitehead 1929.

[25] See Clarke 1981, 1985; Randell and Cohn 1989, 1992, Randell 1991, Cohn *et al.* 1993, Randell *et al.* 1992a, 1992b, 1992c; Vieu 1991, Aurnague and Vieu 1993a, 1993b.

been a very active and yet independent research area in artificial intelligence, particularly under the impact of Allen's work.[26]

In all of these systems, parthood (and consequently the other principal mereological relations) is characterized derivatively in terms of topological connection ('C') in accordance with the following definition:

$$\mathsf{P}xy =_{\mathrm{df}} \forall z(\mathsf{C}zx \rightarrow \mathsf{C}zy). \tag{11}$$

As is clear, much of the intuitive appeal of this reduction depends on the intended interpretation of the basic topological relation (which is axiomatized as a reflexive and symmetric relation – I shall again skip the details). If we take 'C' to mean the same as 'O', then (11) becomes a standard mereological equivalence; but things may change radically on different interpretations. Clarke follows Whitehead and explicitly suggests that one might "interpret the individual variables as ranging over spatio-temporal regions and the two-place primitive predicate, 'x is connected with y', as a rendering of 'x and y share a common point'".[27] This account has been subscribed to by other authors as well. Since points are not regions, sharing a point does not imply overlapping, which therefore does not coincide with (even though it is included in) connection. This means that things may be *externally* connected:

$$\mathsf{EC}xy =_{\mathrm{df}} \mathsf{C}xy \wedge \neg \mathsf{O}xy. \tag{12}$$

But, whereas on Smith's more standard rendering of 'C' (4) this would be explained in terms of overlapping of a common boundary (though not of a common part, recalling that boundaries need not be parts of the things they bound), here the explanation is left open, for boundaries are just not included in the domain. Thus, on this account things can be topologically "open" or "closed" without there being any corresponding mereological difference – a feature that some have found philosophically unpalatable.[28]

Recently, Randell, Cui and Cohn have proposed a modified version of their theory in which 'x is connected with y' is taken as a rendering of 'the topological closures of regions x and y share a common point'.[29] The reason

[26] See Allen 1981, 1984, Allen and Hayes 1985 (though much can already be found in Hamblin 1969, 1971). See Galton 1993 for a unified theory of space, time and motion.

[27] Clarke 1981: 205. See also Gerla and Tortora 1992.

[28] See Simons: "What we are being asked to believe is that there are two kinds of individuals, 'soft' (open) ones, which touch nothing, and partly or wholly 'hard' ones, which touch something. Yet we are not allowed to believe that there are any individuals which make up the difference. We can discriminate individuals which differ by as little as a point, but are unable to discriminate the point. It is hard to find satisfaction in this picture." (Simons 1987: 98) One is reminded here of Brentano's protest against the "monstrous doctrine that there would exist bodies with and without surfaces, the one class containing just so many as the other, because contact would be possible only between a body with a surface and another without". (Brentano 1976: 146-47; the reference is to Bolzano 1851) A way of recovering the notion of a boundary within Clarke's framework (relative to finite domains) is indicated in Vieu 1991 and Aurnague and Vieu 1993.

[29] Randell *et al.* 1992a, 1992b, Cohn *et al.* 1993.

is precisely to do justice to the intuition that "from the naive point of view, the distinction between open, semi-open and closed regions is not drawn", as well as to avoid the consequence that "if we map bodies to closed regions (as the spaces they occupy), then their complements become open, which is a less agreeable result".[30] This shift of interpretation is reflected, formally, in the abandonment of the quasi-Boolean operation of complementation originally used by Clarke

$$x' =_{\mathrm{df}} \iota y \forall z(\mathsf{C}zy \leftrightarrow \neg \mathsf{P}zx) \tag{13}$$

in favour of the following weaker variant:

$$x' =_{\mathrm{df}} \iota y \forall z((\mathsf{C}zy \leftrightarrow (\mathsf{P}xz \vee \neg \mathsf{IP}zx)) \wedge (\mathsf{O}zy \leftrightarrow \neg \mathsf{P}zx)) \tag{14}$$

where 'IP' is defined as follows (in contrast to (9) above):

$$\mathsf{IP}xy =_{\mathrm{df}} \mathsf{P}xy \wedge \neg \exists z(\mathsf{C}zx \wedge \mathsf{C}zy \wedge \neg \mathsf{O}zx \wedge \neg \mathsf{O}zy). \tag{15}$$

This guarantees that every (non-universal) region be connected with its own complement and, more generally, it avoids the above-mentioned feature of Clarke's original formulation: the "remainder principle" of classical mereology:

$$\mathsf{P}xy \wedge \neg \mathsf{P}yx \rightarrow \exists z(\mathsf{P}zy \wedge \mathsf{O}zx) \tag{16}$$

is in fact a theorem of the modified theory. However, there are some drawbacks as well. For instance, the resulting theory does not support models with atoms (regions with no proper parts). For an atom would have the property that every region connected with it would be connected with its complement, and by (11) that would imply the absurdity that every atom is part of its complement.[31]

So much for the intuitive modeling. It is apparent that the effective meaning of (11) – and consequently the mereological system that one effectively obtains – can drastically change depending on the particular interpretation that one considers. But from our present perspective the interesting question is even more fundamental; it concerns the very basic idea of relying on something like (11) when 'C' is not interpreted mereologically. And it is just here that I have reservations. *If* spatial regions are the only entities of our domain, then the proposed definition is really all there is to mereology, and the different interpretations reflect neither more nor less than a natural variety of possible implementations of the same idea (to which there corresponds a variety of more or less standard mereologized topologies). In fact both Clarke's original system and the developments that followed prove fit

[30] Randell *et al.* 1992a: 394–95.

[31] This is noted in Randell *et al.* 1992b: 173, correcting a mistake of Randell *et al.* 1992a. Three alternative ways of dealing with atoms are made available, but they all determine some departure from the basic framework.

to account for a fair deal of mereo-topological reasoning. Simulation programs have also been built using which one can go as far as to model some rather complicated biological or mechanical processes, such as the cycle of operations in a force pump.[32] If, however, we are to take an open-faced attitude towards other entities than just regions, with or without boundaries, then we do not have much choice. Either we insist on the idea that things can be mapped to the regions that they occupy, or we maintain that the topology of regions is really all we need insofar as the same principles apply to the entities of a common-sense ontology as well. Both views seem to me rather difficult to defend, except perhaps for special purpose representations. A shadow does not overlap the wall onto which it is cast. And an object can be wholly located inside a hole, hence totally connected with it, without actually being part of it. The region that it occupies is part of the region occupied by the hole, but that's all. Or think of Lewis's angels dancing forever on the head of one pin: "At every moment, each occupies the same place as the other. Still they are two distinct proper parts of the total angelic content of their shared region".[33] For the purposes of naive mereo-topological reasoning, these are all cases of things that are connected but not overlapping. They are, following (12), externally connected. But they are not adjacent, which is what the notion of external connection is supposed to account for. The wall in no sense abuts the shadow. And the hole does not squeeze to the side to leave room for its guest. Holes are immaterial and can be *interpenetrated*: if the object is inside the hole, then each part of the object is connected with some part of the hole and it makes no sense to characterize this as *external* connection. From here intuitions diverge rapidly: the notion of connection that we get by reasoning exclusively in terms of regions, no matter which specific interpretation we choose, is just too strong for the general case.

Indeed, from this point of view Smith's definition of connection in terms of boundaries and overlapping as given in (4) is likewise unacceptable. What is required is, rather, a weaker interpretation of connection as *co-localization at* (rather than *sharing of*) some point in space-time. We could try the following: A thing x is connected with a thing y iff either x and the closure of y or y and the closure of x are co-localized – but not necessarily overlapping – at some point (where the closure is the thing together with its boundaries). This would allow one to keep track of the distinction between being part-of and the more general relation of being spatially enclosed-in ('E'), hence between overlapping and the relation of spatially intersecting ('I'), hence, again, among things that are connected and things that are superimposed ('S') or merely adjacent ('A'):

$$\mathsf{E}xy \quad =_{\mathrm{df}} \quad \forall z(\mathsf{C}zx \rightarrow \mathsf{C}zy) \tag{17}$$

[32] See Randell *et al.* 1992c.

[33] The example is from Lewis 1991: 75.

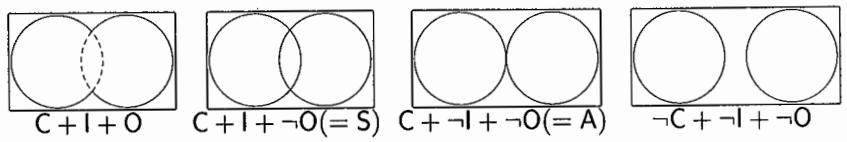


Figure 3

$$\mathsf{I}xy \;=_{\text{df}}\; \exists z(\mathsf{E}zx \wedge \mathsf{E}zy) \tag{18}$$
$$\mathsf{S}xy \;=_{\text{df}}\; \mathsf{I}xy \wedge \neg \mathsf{O}xy \tag{19}$$
$$\mathsf{A}xy \;=_{\text{df}}\; \mathsf{C}xy \wedge \neg \mathsf{I}xy. \tag{20}$$

These distinctions and the corresponding relations in terms of logical inclusion are schematically illustrated in Figure 3.

But of course this weaker interpretation – or others along these lines – would not support (11). If x is a part of y then everything connected with x would be connected with y, but the converse implication would fail. In this sense 'P' does elude 'C'. Hence the basic assumption of the "second way" falls short, and we are back to the first way (though with a new notion of connection).[34]

The Third Way. We thus come to the third possibility mentioned above, which to my knowledge has only been put forward in very recent work by Eschenbach and Heydrich.[35] Here the idea is that a topological framework like Clarke's can effectively be regained in a purely mereological setting (rather than *vice versa*), provided that we embed it in a less restrictive domain. The embedding is straightforward and exploits the non-mereological domain-specific concept of a region. This was the only ontological category admitted in Clarke's domains. In Eschenbach and Heydrich, however, points and other boundaries are also admitted. Accordingly, connection is neither more nor less than overlapping of regions, and yet the topological idea of external connection is made safe by the fact that the common part of two overlapping regions need not itself be a region. Using 'R' to indicate the relevant (primitive) domain-specific region-predicate, the following definitions are all we need in order to reconstruct a mereologized topology of the sort discussed above:

$$\mathsf{C}xy \;=_{\text{df}}\; \mathsf{O}xy \wedge \mathsf{R}x \wedge \mathsf{R}y \tag{21}$$
$$\mathsf{EC}xy \;=_{\text{df}}\; \mathsf{C}xy \wedge \forall z(\mathsf{P}zx \wedge \mathsf{P}zy \rightarrow \neg \mathsf{R}z). \tag{22}$$

As is clear, this approach allows one to retain standard mereology *holus bolus*. Some principles – like the "remainder principle" (16) – may not hold

[34] This is the approach that I followed in Varzi 1993.

[35] Eschenbach and Heydrich 1993.

unrestrictedly in the restricted domain of regions; but this simply mirrors the fact that such a domain (the extension of the non-universal predicate 'R') is deprived of some topologically relevant elements, points and boundaries in the first place. In the comprehensive domain this principle is just as unproblematic as any other. In fact, the main point can be put in the form of a general translation theorem to the effect that the mereology resulting upon restricting the ontology to only include spatial regions is exactly the subtheory that can be obtained from the unrestricted mereology by uniformly restricting the range of quantifiers. Thus, for instance, principle (16) is valid but its restricted variant

$$\mathsf{P}xy \wedge \neg \mathsf{P}yx \rightarrow \exists z(\mathsf{R}z \wedge \mathsf{P}zy \wedge \neg \mathsf{O}zx) \tag{23}$$

(with 'R' interpreted as an ordinary predicate constant) is not.[36]

I find this illuminating, for it shows that mereology needs very little help in order to cope with certain basic topological notions and principles. Formally it is only a matter of restricted quantification. Moreover, this way of looking at things is very general and one may consider exploiting different interpretations of 'R', or referring to other domain-restricting devices (I shall consider some possibilities in a moment). If this amounts to saying that topology is exclusively about regions of space and few other region-related entities such as points and boundaries (or about whatever selected entities we employ to fill in the extension of 'R'), then of course it contrasts the general desiderata expressed above. Whether we try to explain mereological relations among things in terms of topological relations among the corresponding regions, or topological relations among regions in terms of mereological relations among things of a kind, we miss out on something important for the ontology of the everyday world. However, the present approach draws no necessary reduction of parthood to spatial connection, and this gives new content to the idea that topological reasoning about ordinary things can be inferred from the topology of the regions they occupy: on this approach we may safely confine ourselves to reasoning about regions and yet keep track of the relevant difference between enclosure and parthood, or between intersection and overlapping. Thus, the limitation is not substantial: the third way is wide enough to support analogues of the general interpretation of connection suggested above.

[36] To be more precise, I believe the point can be put as follows. First, let L be a mereological language with 'P' as primitive, let L_t be the language obtained from L by replacing 'P' with 'C', and let L_r be obtained from L by adding a new predicate symbol 'R'. Next, for any sentence ϕ of L, let ϕ_t be the sentence of L_t obtained from ϕ by replacing each atomic component of the form '$\mathsf{P}xy$' with '$\forall z(\mathsf{C}zx \rightarrow \mathsf{C}zy)$', and let ϕ_r be the sentence of L_r obtained from ϕ by recursively replacing each quantified component of the form '$\forall x\psi$' or '$\exists x\psi$' (with 'x' free in ψ) with '$\forall x(\mathsf{R}x \rightarrow \psi)$' and '$\exists x(\mathsf{R}x \wedge \psi)$' respectively. Lastly, let M be a mereological system in L and let M_t and M_r be corresponding systems in L_t and L_r obtained by replacing each axiom ϕ of M with ϕ_t or ϕ_r, respectively. Then, for every thesis ϕ of M, the sentence ϕ_t is a thesis of M_t iff ϕ_r is a thesis of M_r.

The Fourth Way. There is also a fourth way. I only mention it now because I am not aware of any serious proposal in this direction, but the basic idea seems to me simple and attractive. If connection is too strong for the purpose of doing mereology, and parthood too weak for doing topology, why not just put the two notions together to get the right blend? Why not build a unified framework based on a single mereo-topological primitive covering both territories? An obvious possibility is to use as a primitive the ternary relation 'x and y are connected parts of z'. Indicating this with '$\mathsf{CP}xyz$', we can define parthood and connection as follows:

$$\mathsf{P}xy =_{\mathrm{df}} \mathsf{CP}xxy \tag{24}$$

$$\mathsf{C}xy =_{\mathrm{df}} \exists z\mathsf{CP}xyz \tag{25}$$

From here we can then go on as we wish. For instance, we can define interior parthood using (9) and then proceed as in Smith's account. Or we can follow alternative routes, including an account to the effect that the equivalence corresponding to (11) holds.

This strategy has in fact one obvious advantage, namely that it remains neutral with respect to the actual interpretation of the notions defined: 'part' and 'connection' can be characterized axiomatically (and interpreted intuitively) as if they were two independent primitives. The only mutual constraints are that for (24) to make good sense, connection must be a reflexive relation, whereas (25) presupposes that every pair of connected entities have a mereological sum. But these are perfectly uncontroversial presuppositions. In particular, the latter is not only an obvious consequence of the sum principle of classical extensional mereology (1), but also a principle held by the opponents of (1) (sometimes with the precise intent of stressing a distinction between admissible and inadmissible sums). From this point of view, this fourth "way" embodies the same formal generality as the first way, but since it only requires one primitive it also enjoys a certain conceptual economy that can be seen as an advantage of the second way.

4 The Ease of Mereo-Topology

We have, then, a rather comprehensive taxonomy of possible strategies and theories. Each of them reflects some way of overcoming the bounds of a plain mereology in dealing with topological notions and properties. And each does so without requiring a significant departure from the general outlook that led to the development of modern mereological systems (witness the fact that all theories considered above are compatible with a nominalistic stand).

Now *which* way is actually to be preferred is not a question that I here intend to address any further. The main strategies have been developed mostly on independent grounds and with disparate purposes, and putting them in the same sandbox and under the same light is only a first step towards a critical appraisal of their relative limits and potentialities. What I

wish to emphasize, rather, is that the difference between the various alternatives is not only a matter of applicative purposes, or formal thoroughness, or computational efficiency. Although each of these concerns may have played an important role in the development of each single theory, I think the difference lies first and foremost in the ontological status that certain entities – from boundaries to ordinary things – are accorded. As we have seen, where and how the domains of mereology and topology are bridged depends heavily on the ontological fauna that one is willing to countenance, on the variety of entities that one is ready to allow in the universe of discourse. And it is just here that I would like to add some remarks.

We have a picture that looks like a network connecting two extreme positions. At one extreme we find Whitehead's early stand as reflected in the definition of 'join' discussed at the beginning of this paper: If we only allow for self-connected entities, then mereology may even subsume topology, though we cannot expect it to be quite classical (the sum principle (1) must fail). At the other extreme we find Clarke's exploitation of Whitehead's late approach: If we only have regions, then topology alone suffices and mereology can be subsumed easily, though again we cannot expect the outcome to be very classical (the remainder principle (16) must be sacrificed). In between we have a variety of intermediate and perhaps not always directly comparable positions, each according greater weight to some entities over others. Now, one thing that is remarkable in this picture is that in spite of the apparent conflict the two extreme positions can be seen as implementing the very same idea. It is, indeed, a matter of restricted quantification in the spirit exemplified by Eschenbach and Heydrich. Just as Clarke's account can be viewed as the result of restricting classical mereology to a domain of regions ('R'), in the precise sense that every universal or existential statement amounts to a restricted quantification of the forms '$\forall x(\mathsf{R}x \rightarrow \psi)$' and '$\exists x(\mathsf{R}x \wedge \psi)$', so Whitehead's theory can be obtained by restricting quantified statements using a predicate for self-connectedness ('SC') i.e., by transforming them into restricted statements of the form '$\forall x(\mathsf{SC}x \rightarrow \psi)$' and '$\exists x(\mathsf{SC}x \wedge \psi)$'. Mereology has no "predicate" for self-connectedness, but nothing prevents us from borrowing it from somewhere else, just as we can borrow a region predicate 'R'. As a matter of fact, if we do so then the defective definition of 'join' (2) becomes perfectly adequate:

$$\mathsf{J}xy =_{\mathrm{df}} \exists z(\mathsf{SC}z \wedge \mathsf{O}zx \wedge \mathsf{O}zy \wedge \forall w(\mathsf{SC}w \wedge \mathsf{P}wz \rightarrow \mathsf{O}wx \wedge \mathsf{O}wy)). \tag{26}$$

Perhaps this comes as no surprise, since both systems are Whitehead's after all. It is, however, instructive that two apparently opposite positions support essentially the same interpretation (modulo specific differences in the axiomatization).

We can also see how this relates to the intermediate positions in the network. If, as seems to be the case, topology is to a great extent a matter of domain-specific predicates, then in the end the bounds of mereology – at

least the ones considered above – do not seem to determine any dramatic conceptual limitation. As long as we are capable of specifying what we are talking about, the bounds are easily overcome. Now this brings us back to a remark that I left open when discussing Smith's choice of primitives. What I find interesting in the viability of many possible primitives for a system like Smith's (but the same applies to other systems of the sort) is that there is no *prima facie* ontological or methodological reason for preferring one choice to another – e.g., for preferring 'IP' to 'B' or 'C'. In fact we saw that we can even take as a topological primitive the predicate 'SC'. Thus, also in this case topology can be viewed as a business of simple predication and restricted quantification. The difference with the extreme positions mentioned above is that in this case the domain's composition is left open. No assumption is made concerning the extension of the distinguishing topological notions – and this supports a non-trivial (hence open-faced) use of 'SC'. From this perspective the bounds of mereology are the bounds of any fundamental theory. Mereology tells us how the world is constituted *in general*, but if we want to talk about certain things as opposed to others, if we want to pick out certain classes of entities instead of others, we need some means of referring to them. Whether we then make this into a theory in many chapters (first way), a compact monograph (fourth way), or something in between (second and third way) is of little importance as long as the choice is recognized to be a matter of ontological transparency.

There remains a question of "what's next", to use Lejewski's phrase.[37] I have been talking of topology as a necessary next step after mereology, and I have done this by concentrating on connection and related notions. However, there is of course much more than this to topology. I do not mean to say that we still need to do a lot of work to get close to the topology actually used by mathematicians – I am not even sure that that is necessary. Rather, I want to underline that there are many important topological notions that cannot be captured by any of the systems outlined here and which nevertheless play a very basic role in our everyday reasoning about the world. For instance, how can we distinguish between things with holes and things without – between a sphere and a torus? We need, it seems, an additional predicate of "simple connectedness", or some means to distinguish surfaces of different genus. How can we account for such basic spatial relations as being inside or outside a given object (or region) when this is not a matter of pure topological closure? For instance, how can we say whether the fly is inside or outside the glass? We need, some authors have suggested, an additional topological operation capturing some notion of "convex hull", not definable in terms of "connection" and the like.[38] Whether or not such additions yield an adequate treatment of the examples mentioned is of course a complex matter. It seems, however, that from this point of view one can hardly feel

[37] From Lejewski 1982.

[38] This is the strategy followed e.g. by Randell, Cui and Cohn in most of the works cited above. The fly example is from Vieu 1991 (adapted from Herskovits 1986).

satisfied with simply expanding a mereological framework with a notion of connectedness. One needs much more just to accomplish some very basic pieces of topological reasoning. And, more importantly, even when a satisfactory amount of topological reasoning could be regained, we would need to move into other provinces to account for equally basic commonsensical intuitions concerning, for instance, movement of parts or interactions among wholes. After all the bounds of topology are pretty narrow too. The world of topology is initially a world of spheres and toruses and little else, and we need to step into morphology – the theory of qualitative discontinuities – just to account for certain basic differences in shape; we need to step into kinematics just to account for certain basic differences of behavior.

My provisional conclusion is thus two-sided. On the one hand, the move from mereology to a mereo-topology is an important and yet rather easy matter of specialization (comparable to the move from, say, set theory to set-theoretical topology). On the other, if we go the way of saying that topology is required for the purposes of investigating the common-sense world, then we can hardly stop there. Many other boundaries have to be crossed – which in effect is a way of saying that many more things have to be taken at face value.

References

Allen J. F. 1981 "An Interval-Based Representation of Temporal Knowledge", *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, Vancouver: IJCAI [Morgan Kaufmann], Vol. 1, pp. 221–26.

Allen J. F. 1984 "Towards a General Theory of Action and Time", *Artificial Intelligence* 23, 123–54.

Allen J. F., Hayes P. 1985 "A Common-Sense Theory of Time", *Proceedings of the 9th International Joint Conference on Artificial Intelligence*, Los Angeles: IJCAI [Morgan Kaufmann], Vol. 1, pp. 528–31.

Aurnague M., Vieu L. 1993a "A Three-Level Approach to the Semantics of Space", in C. Z. Wibbelt (ed.), *The Semantics of Prepositions: From Mental Processing to Natural Language Processing*, Berlin: Mouton de Gruyter, pp. 393–439.

Aurnague M., Vieu L. 1993b "Toward a Formal Representation of Space in Language: A Commonsense Reasoning Approach", in F. Anger, H. Guesgen and J. van Benthem (eds.), *Proceedings of the Workshop on Spatial and Temporal Reasoning. 13th International Joint Conference on Artificial Intelligence*, Chambéry: IJCAI, pp. 123–58.

Bach E. 1986 "The Algebra of Events", *Linguistics and Philosophy*9, 5-16.

Baxter D. 1988a "Identity in the Loose and Popular Sense", *Mind* 97, 575–82.

Baxter D. 1988b 'Many-One Identity', *Philosophical Papers* 17, 193–216.

Benthem J. van 1983 *The Logic of Time*, Dordrecht: Kluwer (2nd ed. 1991).

Bolzano B. 1851 *Paradoxien des Unendlichen*, hrsg. aus dem schriftlichen Nachlasse des Verfassers von F. Přihonský, Leipzig (English translation by D. A. Steele, *Paradoxes of the Infinite*, London: Routledge and Kegan Paul 1950).

Brentano F. 1976 *Philosophische Untersuchungen zu Raum, Zeit und Kontinuum*, hrsg. von S. Körner und R. Chisholm, Hamburg: Meiner (Eng. trans. by B. Smith, *Philosophical Investigations on Space, Time and the Continuum*, London: Croom Helm, 1988).

Burkhardt H., Dufour C. A. 1991 "Part/Whole I: History", in H. Burkhardt and B. Smith (eds.), *Handbook of Metaphysics and Ontology*, Munich: Philosophia, pp. 663–73.

Casati R., Varzi A. C. 1994 *Holes and Other Superficialities*, Cambridge, MA: MIT Press.

Chisholm R. M. 1976 *Person and Object. A Metaphysical Study*, London: Allen and Unwin.

Chisholm R. M. 1984 "Boundaries as Dependent Particulars", *Grazer Philosophische Studien* 10, 87–95.

Chisholm R. M. 1989 *On Metaphysics*, Minneapolis: University of Minnesota Press.

Chisholm R. M. 1992/93 "Spatial Continuity and the Theory of Part and Whole. A Brentano Study", *Brentano Studien* 4, 11–23.

Clarke B. L. 1981 "A Calculus of Individuals Based on 'Connection'", *Notre Dame Journal of Formal Logic* 22, 204–18.

Clarke B. L. 1985 "Individuals and Points", *Notre Dame Journal of Formal Logic* 26, 61–75.

Cohn A. G., Randell D. A., Cui Z. 1993 "A Taxonomy of Logically Defined Qualitative Spatial Regions", in N. Guarino and R. Poli (eds.), *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova: Ladseb-CNR, pp. 149–58.

Cruse D. A. 1979 "On the Transitivity of the Part-Whole Relation", *Journal of Linguistics* 15, 29–38.

De Laguna T. 1922 "Point, Line, and Surface as Sets of Solids", *Journal of Philosophy* 19, 449–61.

Eberle R. A. 1970 *Nominalistic Systems*, Dordrecht: Reidel.

Eschenbach C., Heydrich W. 1993 "Classical Mereology and Restricted Domains", in N. Guarino and R. Poli (eds.), *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova: Ladseb-CNR, pp. 205–17.

Hamblin C. 1969 "Starting and Stopping", *The Monist* 53, 410–25.

Hamblin C. 1971 "Instants and Intervals", *Studium Generale* 24, 127–34.

Galton A. P. 1993 "Towards an Integrated Logic of Space, Time, and Motion", in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, Chambéry: IJCAI [Morgan Kaufmann], Vol. 2, pp. 1550–55.

Gerla G., Tortora R. 1992 "La relazione di connessione in A. N. Whitehead: Aspetti matematici", *Epistemologia* 15, 351–64.

Goodman N. 1951 *The Structure of Appearance*, Cambridge, MA: Harvard University Press (3rd ed. Dordrecht: Reidel, 1977).

Goodman N. 1956 "A World of Individuals", in J. M. Bochenski, A. Church, N. Goodman, *The Problem of Universals. A Symposium*, Notre Dame: University of Notre Dame Press, pp. 13–31.

Goodman N. 1958 "On Relations that Generate", *Philosophical Studies* 9, 65–66.

Henry D. P. 1991 *Medieval Mereology*, Amsterdam: Grüner.

Herskovits A. 1986 *Language and Spatial Cognition. An Interdisciplinary Study of the Prepositions in English*, Cambridge: Cambridge University Press.

Husserl E. 1901 *Logische Untersuchungen. Zweiter Band. Untersuchungen zur Phänomenologie und Theorie der Erkenntnis*, Halle: Niemeyer (2nd ed. 1913; Eng. trans. by J. N. Findlay, *Logical Investigations*, London: Routledge & Kegan Paul 1970).

Kamp H. 1979 "Events, Instants, and Temporal Reference", in R. Bäuerle, U. Egli and A. von Stechow (eds.), *Semantics from Different Points of View*, Berlin: Springer, pp. 376–417.

Lejewski C. 1982 "Ontology: What's Next?", in W. Leinfellner, E. Kraemer and J. Schank (eds.), *Language and Ontology. Proceedings of the 6th International Wittgenstein Symposium*, Vienna: Hölder-Pichler-Tempsky, pp. 173–85.

Lejewski C. 1986 "Logic, Ontology and Metaphysics", in S. G. Shanker (ed.), *Philosophy in Britain Today*, London: Croom Helm, pp. 171–97.

Leonard H. S., Goodman N. 1940 "The Calculus of Individuals and Its Uses", *Journal of Symbolic Logic* 5, 45–55.

Leśniewski S. 1916 *Podstawy ogólnej teoryi mnogości* I, Moskow: Prace Polskiego Koła Naukowego w Moskwie, Sekcya matematyczno-przyrodnicza (Eng. trans. by D. I. Barnett, "Foundations of the General Theory of Sets. I", in S. Leśniewski, *Collected Works*, S. J. Surma, J. Srzednicki, D. I. Barnett and F. V. Rickey (eds.), Dordrecht: Nijhoff 1992, Vol. 1, pp. 129–73).

Lewis D. K. 1991 *Parts of Classes*, Oxford: Blackwell.

Link G. 1987 "Algebraic Semantics for Event Structures", in J. Groenendijk, M. Stockhof and F. Veltman (eds.), *Proceedings of the 6th Amsterdam Colloquium*, Amsterdam: Institute for Language, Logic and Computation, pp. 243–62.

Lowe V. 1953 "Professor Goodman's Concept of an Individual", *Philosophical Review* 62, 117–26.

Menger K. 1940 "Topology Without Points", *Rice Institute Pamphlets* 27, 80–107.

Needham P. 1981 "Temporal Intervals and Temporal Order", *Logique et Analyse* 24, 49–64.

Randell D. A. 1991 *Analysing the Familiar: Reasoning about Space and Time in the Everyday World*, University of Warwick PhD Thesis.

Randell D. A., Cohn A. G. 1989 "Modeling Topological and Metrical Properties in Physical Processes", in R. J. Brachman, H. J. Levesque and R. Reiter (eds.), *Principles of Knowledge Representation and Reasoning. Proceedings of the First International Conference*, Los Altos: Morgan Kaufmann, pp. 357–68.

Randell D. A., Cohn A. G. 1992 "Exploiting Lattices in a Theory of Space and Time", *Computers and Mathematics with Applications* 23, 459–76.

Randell D. A., Cui Z., Cohn A. G. 1992a "An Interval Logic of Space Based on 'Connection'", in B. Neumann (ed.), *Proceedings of the 10th European Conference on Artificial Intelligence*, Chichester: John Wiley, pp. 394–98.

Randell D. A., Cui Z., Cohn A. G. 1992b "A Spatial Logic Based on Regions and Connections", in B. Nebel, C. Rich and W. Swartout (eds.), *Principles of Knowledge Representation and Reasoning. Proceedings of the Third International Conference*, Los Altos: Morgan Kaufmann, pp. 165–76.

Randell D. A., Cui Z., Cohn A. G. 1992c "Naive Topology: Modeling the Force Pump", in B. Faltings and P. Struss (eds.), *Recent Advances in Qualitative Physics*, Cambridge, MA: MIT Press, pp. 177–92.

Rescher N. 1955 "Axioms for the Part Relation", *Philosophical Studies* 6, 8-11.

Sanford D. 1993 "The Problem of the Many, Many Composition Questions, and Naive Mereology", *Noûs* 27, 219–228.

Simons P. M. 1982 "The Formalisation of Husserl's Theory of Wholes and Parts", in B. Smith (ed.), *Parts and Moments: Studies in Logic and Formal Ontology*, Munich: Philosophia, pp. 113–59.

Simons P. M. 1987 *Parts: A Study in Ontology*, Oxford: Clarendon.

Simons P. M. 1991 "Whitehead und die Mereologie", in M. Hampe and H. Maassen (eds.), *Die Gifford Lectures und ihre Deutung. Materialien zu Whiteheads "Prozess und Realität"*, Vol. 2, Frankfurt: Suhrkamp, pp. 369–88.

Smith B. 1991 "Relevance, Relatedness and Restricted Set Theory", in G. Schurz and G. J. W. Dorn (eds.), *Advances in Scientific Philosophy. Essays in Honour of Paul Weingartner*, Amsterdam: Rodopi, pp. 45–56.

Smith B. 1992 "Topology for Philosophers", talk presented at the *15th International Wittgenstein Symposium: Philosophy of Mathematics*, Kirchberg a/W, August 1992.

Smith B. 1993 "Ontology and the Logistic Analysis of Reality", in N. Guarino and R. Poli (eds.), *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova: Ladseb-CNR, pp. 51–68 (revised version in G. Häfliger and P. M. Simons (eds.), *Analytic Phenomenology*, Dordrecht: Kluwer, 1994, pp. 223–245).

Tarski A. 1929 "Les fondements de la géométrie des corps", *Księga Pamiątkowa Pierwszkego Polskiego Zjazdu Matematycznego*, supplement to *Annales de la Société Polonaise de Mathématique* 7, 29-33 (Eng. trans. by J. H. Woodger, "Foundations of the Geometry of Solids", in A. Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, Oxford: Clarendon 1956, pp. 24–29).

Tiles J. E. 1981 *Things That Happen*, Aberdeen: Aberdeen University Press.

Varzi A. C. 1993 "Spatial Reasoning in a Holey World: A Sketch", in F. Anger, H. Guesgen and J. van Benthem (eds.), *Proceedings of the Workshop on Spatial and Temporal Reasoning. 13th International Joint Conference on Artificial Intelligence*, Chambéry: IJCAI, pp. 47–59.

Vieu L. 1991 *Sémantique des relations spatiales et inférences spatio-temporelles: Une contribution à l'étude des structures formelles de l'espace en Langage Naturel*, Université Paul Sabatier de Toulouse PhD Thesis.

Winston M., Chaffin R., Herrmann D., 1987, "A Taxonomy of Part-Whole Relations", *Cognitive Science* 11, 417-44.

White G. 1993 "Mereology, Combinatorics, and Categories", Preliminary Report for the SNF Project 11.31211.91; to appear in *The Monist*.

Whitehead A. N. 1919 *An Enquiry Concerning the Principles of Human Knowledge*, Cambridge: Cambridge University Press.

Whitehead A. N. 1920 *The Concept of Nature*, Cambridge: Cambridge University Press.

Whitehead A. N. 1929 *Process and Reality. An Essay in Cosmologyy*, New York: Macmillan.

The Ontological Level

Nicola Guarino

1 Introduction

In 1979, Ron Brachman discussed a classification of the various primitives used by KR systems at that time.[1] They could be grouped in four levels, ranging from the *implementational* to the *linguistic* level (Table 1). Each level corresponds to an explicit set of primitives offered to the knowledge engineer. At the implementational level, primitives are merely pointers and memory cells, which allow us to construct data structures with no *a priori* semantics. At the logical level, primitives are propositions, predicates, logical functions and operators, which are given a formal semantics in terms of relations among objects in the real world. No particular assumption is made however on the nature of such relations: classical predicate logic is a general, uniform, neutral formalism, and the user is free to adapt it to its own representation purposes. At the conceptual level, instead, primitives have a definite cognitive interpretation, corresponding to language-independent concepts like elementary actions or thematic roles. Finally, primitives at the linguistic level directly refer to verbs and nouns.

Brachman noticed an evident gap in this classification: while primitives at the logical level are extremely general and content-independent, at the conceptual level they get a specific intended meaning that must be taken as a whole, without any account of its internal structure. He proposed the introduction of an intermediate *epistemological* level, where the primitives allow us to specify "the formal structure of conceptual units and their interrelationships *as conceptual units* (independent of any knowledge expressed therein)".[2] In other words, while the logical level deals with abstract predicates and the conceptual level with specific concepts, at the epistemological level the generic notion of a concept is introduced as a knowledge structuring primitive.

Brachman's KL-ONE[3] is an example of a formalism built around these notions. Its main contribution was to give an epistemological foundation to cognitive structures like frames and semantic networks, whose formal contradictions had been revealed in the famous "What's in a link?" paper by Bill Woods.[4] Brachman's answer to Woods' question was that conceptual

[1] Brachmann 1979.
[2] Brachman 1979: 30.
[3] Brachman 1979, Brachman and Schmolze 1985.
[4] Woods 1975.

R. Casati, B. Smith, G. White (eds.) *Philosophy and the Cognitive Sciences* 443-456

Level	Primitives
Implementational	Memory Cells, Pointers
Logical	Propositions, predicates, functions, logical operators
Epistemological	Concept types, structuring relations
Conceptual	Conceptual relations, primitive objects and actions
Linguistic	Linguistic terms

Table 1: Classification of Primitives used in KR formalisms (after Brachman 1979). The epistemological level was the "missing level".

links should be accounted for by epistemological links, which describe the formal knowledge structure needed to justify conceptual inferences. KL-ONE focused in particular on the inferences related to the so-called IS-A relationship, offering primitives to describe the (minimal) formal structure of a concept needed to guarantee "formal inferences about the relationship (subsumption) between a concept and another". This formal structure consists of the sum of the constituents of a concept (primitive concepts and role expressions) and the constraints among them, independently of any commitment about: (i) the meaning of primitive concepts; (ii) the meaning of roles themselves; (iii) the nature of each role's contribution to the meaning of the concept. The intended meaning of concepts remains therefore totally arbitrary: indeed, the semantics of current descendants of KL-ONE is such that – at the logical level – concepts correspond to arbitrary monadic predicates, while roles are arbitrary binary relations. In other words, at the epistemological level, emphasis is more on formal reasoning than on (formal) representation: the very task of representation, i.e. the structuring of a domain, is left to the user.

Current frame-based or object-oriented formalisms suffer from the same problem. For example, the advantage of a frame-based language over pure first-order logic is that some logical relations as those corresponding to classes and slots have a peculiar, structuring meaning, being those upon which a particular model of the domain is founded. This meaning is the result of a number of ontological commitments, which accumulate in layers starting from the very beginning of a knowledge base development process.[5] For a particular knowledge base, its ontological commitments are, however, implicit and strongly dependent on the particular task being considered, since the formalism itself is in general neutral about ontological choices.[6]

In this paper we argue against this neutrality, claiming that a rigorous ontological foundation for knowledge representation can improve the quality of the whole knowledge engineering process, making it easier to build knowledge bases which are at least understandable (if not reusable). We

[5] See for instance Davis, Schrobe *et al.* 1993.

[6] See Genesereth and Nilsson 1987.

contrast the notion of formal ontology, intended as a theory of the *a priori* nature of objects, to that of (formal) epistemology, intended as a theory of meaning connections.[7] We show in the following how theories defined at the epistemological level, based on structured representation languages like KL-ONE, cannot be distinguished from their "flat" first-order logic equivalents unless we make clear their implicit ontological assumptions by stating formally what it means to interpret a unary predicate as a concept (class) and a binary predicate as a "role" (slot).[8] We therefore introduce the notion of ontological level, as an intermediate level between the epistemological and the conceptual one (Table 2). While the epistemological level is the level of structure, the ontological level is the level of meaning. At the ontological level, knowledge primitives satisfy formal meaning postulates, which restrict the interpretation of a logical theory on the basis of formal ontology, intended as a theory of *a priori* distinctions: [9]

- among things, i.e. entities of the world (physical objects, events...);
- among meta-level categories used to model the world (concepts, properties, states, roles, attributes, parts...).

We focus here on the latter kind of distinctions, showing how the basic dichotomy existing in KR systems between concepts like 'Apple' and assertional properties like 'Red' can be understood in terms of the philosophical distinction among sortal and characterising universals.[10] In the next section we present a couple of examples showing the necessity to make such a distinction explicit. In Section 3 we introduce the notion of ontological commitment as a constrained interpretation of a logical theory, and we sketch a basic ontology of meta-level categories of unary predicates. In Section 4 we discuss the role of the ontological level in the current practice of knowledge engineering.

2 Reds and Apples

Suppose we want to state that a red apple exists. In standard first order logic, it is straightforward to write down something like '$\exists x(Ax \wedge Rx)$'. If, however, we want to impose some structure on our domain, the simplest formalism we may resort to is many-sorted logic. Yet, we have to decide which predicate corresponds to a sort: we may write '$\exists x{:}A.Rx$' as well as '$\exists x{:}R.Ax$' (or maybe '$\exists(x{:}A, y{:}R)x = y$'). All these structured formalisations are equivalent to the previous one-sorted axiom, but each one contains an implicit structuring choice. At the *epistemological* level, this choice is up to the user, since the

[7] The use of the term 'epistemology' sounds somehow reductive here, but I believe this reflects its common understanding in the KR literature.

[8] A preliminary ontological analysis of the primitives used in KL-ONE-like formalisms appeared in Guarino 1992.

[9] Formal ontology has recently been defined as "the systematic, formal, axiomatic development of the logic of all forms and modes of being"; Cocchiarella 1991.

[10] See Strawson 1959, Wiggins 1980

semantics of the primitive "sort" is the same as its corresponding first-order predicate. At the *ontological* level, what we want is a formal, restricted semantic account that reflects the ontological commitment intrinsic in the use of a given predicate as a sort. This means that the choice of a particular axiomatisation is still up to the user, but its consequences are formalised in such a way that another user can understand the meaning of the choice itself, and possibly agree on it on the basis of its semantics.

In our case, a statement like '$\exists x{:}R.Ax$' sounds as intuitively odd: what are we quantifying on? Do we assume the existence of "instances of redness" that can have the property of being apples? According to Strawson, the difference between the two predicates lies in the fact that 'Apple' "supplies a principle for distinguishing and counting individual particulars which it collects", while 'Red' "supplies such principle only for particulars already distinguished, or distinguishable, in accordance with some antecedent principle or method".[11] This distinction is known in the philosophical literature as the distinction between sortal and non-sortal (characterising) universals, and is (roughly) reflected in natural language by the fact that the former are common nouns, while the latter are adjectives. The issue is also related to the difference between count and mass terms, and has been a matter of lively debate among linguists and philosophers[12]. The distinction is implicitly present in the KR literature, where sortal universals are usually called "concepts", while characterising universals are called "properties". The difference between the two is however the result of heuristic considerations, and nothing, in the semantics of a concept, forbids it to be an arbitrary unary predicate.

Our position is that, within a KR formalism, the meaning of structuring primitives as sorts (or concepts, in KR terminology) should be at least specified with formal, necessary conditions at the meta-level, which force the user to accept their consequences when he/she decides to use a given predicate as a sort. According to our previous discussion, a predicate like 'Red' should not satisfy such conditions, and should be excluded therefore from being used as a sort. Notice however that this may be simply a point of view: at the ontological level it is still the user who decides which conditions reflect the intended use of the 'Red' predicate: a more rigid choice would be distinctive of higher levels, like the conceptual or the linguistic level. For example, compare the statement mentioned above with others where the same unary predicate 'Red' appears in different contexts (Figure 1).

In case (2), 'Red' is still a unary predicate whose argument refers to a particular colour instead of a particular fruit; in (3) the argument refers to a particular colour gradation belonging to the set of "reds", while in (4) the argument refers to a human-being, meaning for instance that he/she is a communist.

[11] Strawson 1959: 168

[12] See Pelletier and Schubert 1989 for an overview.

Red(x)	this apple is red	(1)
	the colour of this apple is red	(2)
	crimson is a red	(3)
	John is a red	(4)

Figure 1: Varieties of Predication

We face here the difference of positions between the Linguist and the Philosopher discussed by Bill Woods in one of the historical papers on knowledge representation:[13] while the Linguist "is interested in characterising the fact that the same sentence can sometimes mean different things", the Philosopher on the other hand "is concerned with specifying the meaning of a formal notation rather than a natural language". Woods goes on to state that

> philosophers have generally stopped short of trying to actually specify the truth conditions of the basic atomic propositions, dealing mainly with the specification of the meaning of complex expressions in terms of the meanings of elementary ones. Researchers in artificial intelligence are faced with the need to specify the semantics of elementary propositions as well as complex ones.

In a knowledge representation formalism, we are constantly using natural language words within our formulas, relying on them to make our statements readable and to convey the meaning we have not explicitly stated: however, since words are ambiguous in natural language, it may be important to "tag" these words with a semantic category, endowed with a suitable axiomatisation, in order to guarantee a consistent interpretation. This is unavoidable, in our opinion, if we want to share theories across different domains.[14]

How can we account for the semantic differences in the use of 'Red' in the formulas above? In our opinion, they are not simply related to the fact that the argument belongs to different domains: they are mainly due to different *ways of predication*, i.e. different subject-predicate relationships. Studying the formal properties of such relationships is a matter of formal ontology.

3 A Basic Ontology of Unary Predicate Types

A basic ontology which – according to Strawson's intuitions – classifies unary predicates on the basis of their ability to supply an identification principle for their arguments is presented in Figure 2.

[13] Woods 1975.

[14] Gruber 1993.

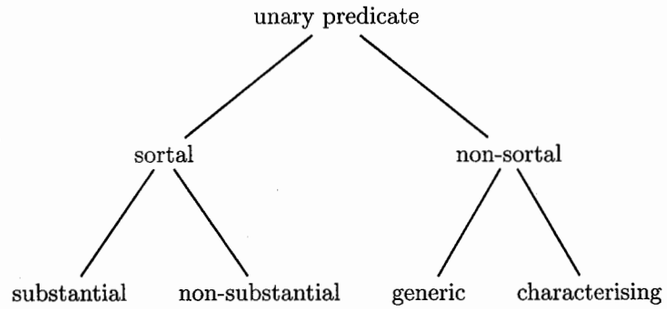


Figure 2: A Basic Ontology of Predicate Types (Wiggins 1980)

Based mainly on Wiggins' work,[15] sortal predicates are divided into substantial (like 'Apple' or 'Human being') and non-substantial (like 'Food' or 'Student') predicates, while non-sortal predicates include generic predicates like 'Thing' and characterising predicates like 'Red'. A preliminary formalisation of such distinctions is presented in the following.

Instead of trying to give a "universal" formal definition of the above categories, we shall pursue here a more modest account: our definitions will be related to a specific knowledge base described by a standard first order theory, which we are interested to "add structure" to. This means that the basic knowledge-building blocks are already fixed, being the predicates of the theory itself; our task will be to offer a formal instrument to clarify their ontological implications for the specific purposes of the understanding and reuse of knowledge bases. We assume therefore that interpretations of our specific theory, rather than describing a real or hypothetical situation in a world that has the same laws of nature of ours,[16] are states of affairs having an "idealised rational acceptability".[17] This choice excludes unwanted metaphysical implications.

Within this framework, let us concentrate on a minimal problem. Suppose to have a first order, non-functional language L with signature $\Sigma = \langle C, P \rangle$, where C is a set of constant symbols and P is a finite set of predicate symbols.

Definition 1 Let L_m be the modal extension of a first order language L obtained by adding to L the usual modal operators, and D be a set. A *rigid model*[18] for L_m based on D is a structure $M = \langle W, \mathcal{R}, D, \mathcal{F}_C, \mathcal{F}_P \rangle$, where W is a set of possible worlds sharing the same interpretation $\mathcal{F}_C$ for constant symbols of L_m, $\mathcal{R}$ is a binary relation on W, D is a domain common to all possible worlds, and $\mathcal{F}_P$is a mapping that assigns to each

[15]Wiggins 1980; Carrara 1992.
[16]Cocchiarella 1993.
[17]Putnam 1981, cited in Aune 1991: 543.
[18]We use here the terminology introduced in Fitting 1993

predicate symbol P of L and world $w \in W$ a unary relation on D.

For a given rigid model M, $\mathcal{R}$ is a relation between worlds (corresponding to interpretations of L) that may differ in the interpretation of predicates while sharing the same interpretation for constants. We want to give $\mathcal{R}$ the meaning of an *ontological compatibility* relation: two worlds are ontologically compatible if they describe plausible alternative states of affairs involving the same elements of the domain. $\mathcal{R}$ will be in this case reflexive, transitive and symmetric (i.e., an equivalence relation), and the corresponding modal theory will be S5.

Definition 2 Let L be a first order language and D a domain. An *ontological commitment* for L based on D is a set C of rigid models for L_m based on D, where the relation $\mathcal{R}$ is an equivalence relation. Such commitment can be specified by an S5 modal theory of L_m, being in this case the set of all its rigid models based on D. A formula Φ of L_m is valid under C ($C \models \Phi$) if it is valid in each model $M \in C$.

Within this modal framework, a preliminary consideration we can make on the distinction between sortal and non-sortal predicates is that the former cannot be necessarily false for each element of the domain: they must be *natural predicates*, in the sense of the following definition:[19]

Definition 3 Let L be a first order language, P a monadic predicate of L, and C an ontological commitment for L. P is called *natural* under C iff $C \models \exists x. \Diamond Px$.

A more substantial observation that comes to mind when trying to formalise the nature of the subject-predicate relationship in the examples above, is that the "force" of this relationship is much higher in 'x is an apple' than in 'x is red'. If x has the property of being an apple, it cannot lose this property without losing its identity, while it doesn't seem the case in the second example. This observation goes back to Aristotelian essentialism, and can be easily formalised as follows:[20]

Definition 4 A predicate P is *ontologically rigid* under C iff it is natural under C and $C \models \forall x. (Px \supset \Box Px)$.

Ontological rigidity seems a useful property for characterising sortals: stating that 'Apple' is rigid and 'Red' is not will clarify the intended meaning of these two predicates in the statement (1) of Figure 1. In this case, if $a \in C$, the two worlds satisfying $(A(a) \wedge R(a))$ and $(A(a) \wedge \neg R(a))$ will turn out to be mutually compatible, while those satisfying respectively $(R(a) \wedge A(a))$ and $(R(a) \wedge \neg A(a))$ will not (due to the constraints imposed on $\mathcal{R}$ by the rigidity of A). Assuming that rigidity is a necessary property for sortals, we can then exclude both $\exists x : R. Ax$ and $\exists (x{:}A, y{:}R). x = y$ from our axiomatisation choices for (1).

[19]Cocchiarella 1993.
[20]Barcan Marcus 1971.

Notice that the naturalness condition in the above definition excludes cases where rigidity would be trivially true due to the impossibility of P. On the other hand, ontological rigidity will be trivially satisfied by predicates being necessarily true for each element of the domain, like 'Thing' or 'Entity'. Yet, according to traditional wisdom they are excluded from being sortals, since no clear distinction criteria are associated to them. Rigidity cannot be therefore be considered as a necessary condition for sortals. We call these "top level" predicates *generic* predicates[21]. In the same category other rigid predicates should be included, that, although being not trivially rigid, are still too general to supply a distinction criterion: 'Object', 'Individual', 'Event'... In our opinion, the distinctive characteristic of generic predicates is that they are rigid but divisive:[22]

Definition 5 Let P be a natural predicate under C, and $<$ be a "proper part" relation assumed as primitive, satisfying the axioms of classical mereology.[23] P is *divisive* under C iff $C \models \exists x.\Diamond(Px \wedge \exists y.y < x) \wedge \forall x.\Box(Px \supset (\forall y.(y < x \supset Py)))$.

Definition 6 Let P be ontologically rigid under C. It is a *generic predicate* in C if it is divisive in C, and a *substantial sortal* in C otherwise.

Within our KR framework, the above definition gives a formal characterisation of the notion of substantial sortals originally introduced by Wiggins, delimiting those rigid predicates that are sortals. We need now a distinction criterion between non rigid predicates: some of them (like 'Student') will presumably be *non-substantial sortals*, while the others (like 'Red') will be *characterising predicates*. The intuition behind the distinction between substantial and non-substantial sortals is that in the first case the identity criterion is given by the predicate itself, while in the second case it is provided by some superordinate sortal. We formalise it as follows:

Definition 7 Let P be a natural predicate which is not ontologically rigid under C. It is a *non-substantial sortal* in C iff there exists a substantial sortal S in C such that $C \models \forall x.\Box(Px \supset Sx)$, and a characterising predicate otherwise.

Since the set of predicate symbols of L is finite and fixed for all possible ontological commitments of L, this definition does not imply any "real" second-order quantification in L_m, nor has any metaphysical implication. For example, suppose that 'Student' is a non-rigid predicate under some commitment C for L. If 'Human being' also belongs to L and is a superordinate substantial sortal under C, then 'Student' will be a non substantial sortal, while otherwise it may simply be a characterising predicate. This

[21] In Pelletier and Schubert 1989 they are called 'super sortals'.

[22] Regarding the criticisms made for instance in Pelletier and Schubert 1989, see the comments in the conclusions.

[23] See for instance Simons 1987. We assume here that L_m and C are suitably extended to include $<$.

means that *a priori* considerations about the real world do not affect our definitions, unless they force the user to revise the original first-order axiomatisation. However, one of the advantages of the ontological level is that an unwanted formal property for a predicate may trigger a knowledge elicitation process: in our case, if 'Student' sounds strange being a characterising predicate, the reason may be that we have forgotten to include 'Human-being' within our axiomatisation.

Adapting some definitions from Cocchiarella,[24] we believe it is important, for knowledge representation purposes, to make some further assumptions on sortals, which characterise what we call a well-founded ontological commitment:

Definition 8 An ontological commitment C based on D is well-founded iff:

- each element of D belongs to a substantial sortal;
- if two substantial sortals are not in the subsumption relationship, then they are mutually disjoint.

From Definitions 7 and 8 it follows that:

Theorem 1 *Two overlapping non-substantial sortals are subordinate to the same substantial sortal.*

Definition 9 Let C be a well-founded ontological commitment. If a substantial sortal S is subordinate to another substantial sortal T under C, then S is called a *kind* of T.

Definition 10 A predicate is called a *sortal* under a commitment C if it is either a substantial or a non-substantial sortal under C.

As a final comment concerning the taxonomy of unary predicates we have discussed in this section, we would like to make the following proposal regarding the relationship between the terminology currently used in KR formalisms and the philosophical terms we have defined here: (Figure 3).

4 The Ontological Level

As well as the logical and epistemological levels are characterised by a (standard) formal semantics, the ontological level is characterised by a formal ontological account, like the one introduced in the previous section. Although we have limited ourselves to a few very basic ontological distinctions, it should be clear that other important distinctions could be done within a similar framework, like for instance those between attributes and arbitrary binary relations. However, the definitions we have given are enough to capture the different uses of the predicate 'Red' shown in Figure 1, which correspond to distinct ontological commitments (Figure 4).

[24] Cocchiarella 1993, Cocchiarella 1977.

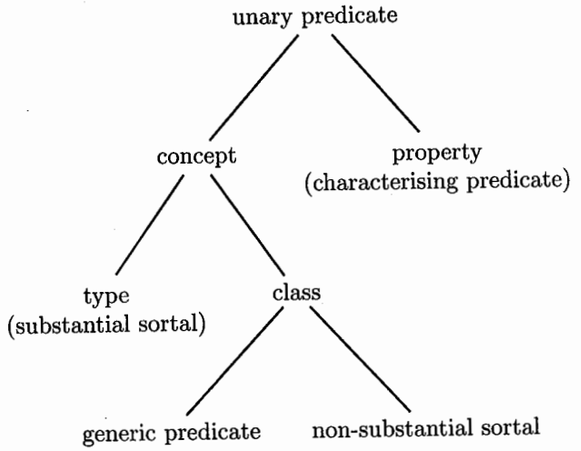


Figure 3: A Terminological Proposal for KR Formalisms

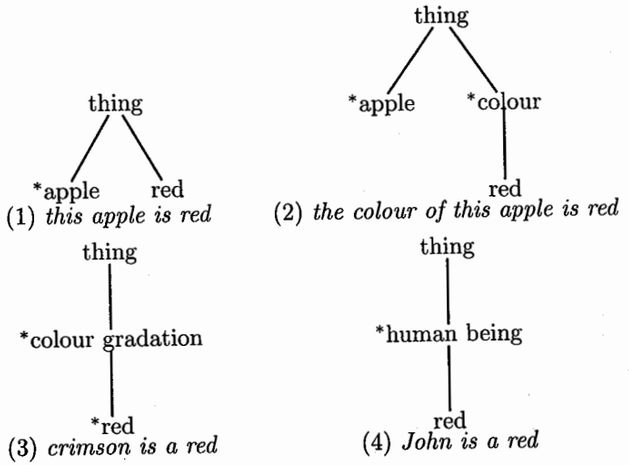


Figure 4: Different Ontological Commitments Capturing Varieties of Predication. Arcs represent subsumption relationships, and asterisks mark substantial sortals.

Level	Primitives	Interpretation	Main Feature
Logical	Predicates, Functions	Arbitrary	Formalisation
Epistemological	Structuring relations	Arbitrary	Structure
Ontological	*Ontological relations*	*Constrained*	*Meaning*
Conceptual	Conceptual Relations	Subjective	Conceptualisation
Linguistic	Linguistic Terms	Subjective	Language Dependency

Table 2: Main Features of the Ontological Level

In case (1), 'Red' is not rigid, and it has no superordinate substantial sortal: it is a characterising predicate, having as argument a physical object. In case (2), 'Red' is still not rigid, but it is subordinate to 'Colour', which is assumed to be rigid and not divisive (a colour has no parts): it is a non substantial sortal, having as argument the colour of a physical object. In case (3), 'Red' is rigid, since its argument is a colour gradation (crimson has to be a red): it is a substantial sortal, and also a kind of colour gradation. Finally, in case (4), 'Red' is used as a contingent property of human-beings and hence is not rigid, but it is not a characterising predicate since it is assumed that being a red implies being a human-being: 'Red' is therefore a non substantial sortal like in (2), under a different ontological commitment.

It is important to stress that, although the notion of ontological commitment we have defined is bound to a quantified modal logic, the computationally bad properties of such a theory have nothing to do with those of the first order language we started with. Even with a language of very limited expressiveness like a description logic,[25] we can embed it in a full quantified modal logic, and use this to define ontological commitment of the original language. This means that we give up to perform any automatic deduction on the modal theory, since we are only interested in its semantic properties. However, given a KR formalism at the epistemological level, we may be interested in expressing somehow its ontological commitment *within the formalism itself*. In other words, this is a matter of *ontological adequacy* of a KR language. We can get this ontological adequacy by suitably restricting the semantics of the epistemological level primitives (assuming for instance that "concepts" used in description logics have the semantics of sortals), or otherwise by having a syntactic way to "tag" a predicate symbol with an ontological category (stating for instance that 'human-being' is a 'substantial sortal', where the latter is a primitive symbol). Some meta-level capability is necessary in the second case.

The main features of the ontological level are compared in Figure 2 to that of other levels. Ontological level is the only level where the intended meaning of a KR language is constrained in a formal way. Lower levels have

[25] See a brief review in Woods and Schmolze 1992.

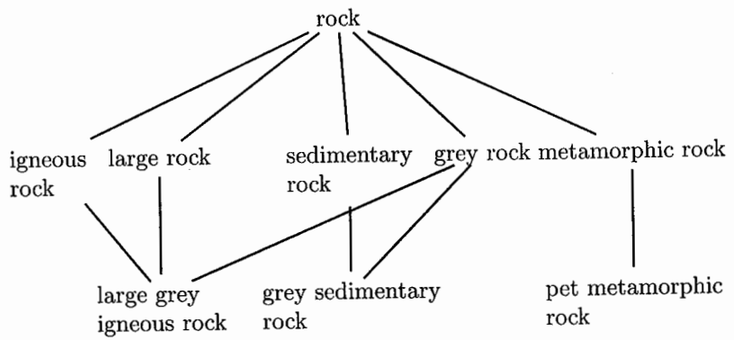


Figure 5: Kinds of Rocks (from Brachman, Fikes *et al.* 1983)

an arbitrary interpretation, since a logical theory admits a number of models much higher than the intended ones; higher levels (which can still be "implemented" at the ontological level) have a subjective interpretation, which can however be the refinement of a formal interpretation already constrained at the ontological level.

5 Conclusions

In Brachman, Fikes *et al.* 1983, the authors discussed the example reported in Figure 5. They argued that a question like "How many kinds of rocks are there?" cannot be answered by simply looking at the nodes subsumed by 'rock' in the network, since the language allows them to easily proliferate. Hence they gave up answering such dangerous questions within a KR formalism, by specifying a functional interface designed to answer "safe" queries about analytical relationships between terms independently of the structure of the knowledge base, as "a large grey igneous rock is a grey rock". On the other hand, the same authors, in an earlier paper,[26] stressed the importance of terminological competence in knowledge representation, stating for instance that an "enhancement mode transistor" (which is "a *kind* of transistor") should be understood as different from a "pass transistor" (which is "a *role* a transistor plays in a larger circuit").

We hope to have shown in this paper that – in the spirit of Woods' statement cited in Section 3 – terminological competence can be gained by formally expressing the ontological commitment of a knowledge base. If, in the example above, predicates corresponding to 'rock', 'igneous rock', 'sedimentary rock' and 'metamorphic rock' are marked as substantial sortals (as they should be according to their ordinary meaning), while all the others are non-substantial sorts (since they are not rigid), then a safe answer to the query "how many kinds of rocks are there?" can be "at least 3".

[26] Brachman and Levesque 1982.

It is important to make clear however that the complete formal characterisation of the taxonomy described in Section 3 – and the taxonomy itself – are still a matter of open discussion. In particular, the notion of divisiveness is still problematic,[27] since we may assume for instance that each part of an igneous rock is still a rock, invalidating the example above (but not its spirit). Our answer to this objection is that in the notion of a rock, and of a physical object in general, there is implicit a notion of external boundary,[28] such that an *undetached* part of a rock is not a rock, bust just a part of it: this is why we can answer a question like "how many rocks are there?", while it is difficult to answer "how many parts of rock are there?". A more thoroughly account of the basic ontology sketched in Section 3 will be however the subject of a forthcoming paper;[29] the the preliminary distinction criteria introduced here have in our opinion the advantage of simplicity, avoiding to make use of subtle notions like ontological foundation, introduced in previous works.[30]

The title chosen for this paper should however suggest the reader that the particular ontology of unary predicates we have proposed here is not the main issue here. Rather, we believe that the main contribution of section 3 is the notion of ontological commitment expressed in terms of a modal framework: the use of a modal logic, used as a tool to constrain the intended semantics of the underlying non modal theory, seems to be unavoidable to express ontological constraints. In the perspective of formal ontology mentioned in the introduction, these constraints should be also related to a-priori distinctions among entities of the world, while we have limited ourselves to meta-level categories. We have tried to show here that (i) *some* formal properties which account for distinctions among predicate types can indeed be worked out, although complete, unproblematic definitions may not be given; (ii) when the semantics of structuring primitives used in KR languages is restricted in order to take into account such formal distinctions at the ontological level, the potential misunderstandings and inconsistencies due to conflicting intended models are highly reduced; (iii) further research in this area is needed, and it should be encouraged within the KR community, in strict co-operation with the philosophical and linguistic communities.[31]

References

Aune, B. 1991 "Metaphysics of Analytic Philosophy", in H. Burkhardt and B. Smith (eds), *Handbook of Metaphysics and Ontology*, Munich: Philosophia 1991.

Barcan Marcus, R. 1971 "Essential Attribution", *The Journal of Philosophy* 7, 187–202.

[27] Pelletier and Schubert 1989.

[28] Smith 1992.

[29] Guarino, Carrara and Giaretta 1994.

[30] Guarino and Boldrin 1993a.

[31] I am especially indebted to Massimiliano Carrara and Pierdaniele Giaretta for having introduced me to the philosophical distinctions among universals. I also thank Luca Boldrin, Matteo Cristani and Claudio Sossai for their valuable comments on earlier versions of this paper.

Brachman, R., R. Fikes, et al. 1983 "Krypton: A Functional Approach to Knowledge Representation", *IEEE Computer* (October), 67–73.

Brachman, R. and H. Levesque 1982 "Competence in Knowledge Representation", in *Proceedings of National Conference on Artificial Intelligence* (AAAI 82), Pittsburgh: American Association for Artificial Intelligence, pp. 189–192.

Brachman, R. J. 1979 "On the Epistemological Status of Semantic Networks", in N. V. Findler (Ed.), *Associative Networks: Representation and Use of Knowledge by Computers*, Academic Press, pp. 3–50.

Brachman, R. J. and J. G. Schmolze 1985 "An Overview of the KL-ONE Knowledge Representation System", *Cognitive Science* 9, 171-216.

Carrara, M. 1992 *Identit e persona nella riflessione filosofica di David Wiggins*, Graduation thesis, Faculty of Philosophy, University of Padova.

Cocchiarella, N. 1977 "Sortals, Natural Kinds and Re-identification", *Logique et Analyse* 80, 439–474.

Cocchiarella, N. B. 1991 "Formal Ontology", in H. Burkhardt and B. Smith (eds), *Handbook of Metaphysics and Ontology*, Munich: Philosophia 1991.

Davis, R., H. Shrobe et al. 1993 "What is in a Knowledge Representation?", *AI Magazine* (Spring 1993), 17–33.

Fitting, M. 1993 "Basic Modal Logic", in D. M. Gabbay, C. J. Hogger and J. A. Robinson (eds.), *Handbook of Logic in Artificial Intelligence and Logic Programming*, Oxford: Clarendon Press, 1993.

Genesereth, M. R. and N. J. Nilsson 1987 *Logical Foundations of Artificial Intelligence*, Los Altos, CA: Morgan Kaufmann.

Gruber, T. 1993 "Toward Principles for the Design of Ontologies Used for Knowledge Sharing", *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova, LADSEB-CNR Int. Rep. 01/93.

Guarino, N. 1992 "Concepts, Attributes and Arbitrary Relations: Some Linguistic and Ontological Criteria for Structuring Knowledge Bases", *Data & Knowledge Engineering* 8, 249–261.

Guarino, N. and L. Boldrin 1993 "Concepts and Relations", *International Workshop on Formal Ontology in Conceptual Analysis and Knowledge Representation*, Padova, LADSEB-CNR Int. Rep. 01/93.

Guarino, N. M. Carrara, and P. Giaretta 1994 "An Ontology of Meta-Level Categories", In D. J. E. Sandewall and P. Torasso (eds), *Principles of Knowledge Representation and Reasoning: Proceedings of the Fourth International Conference* (KR94), San Mateo, CA: Morgan Kaufmann, pp. 270–280.

Mulligan, K., P. M. Simons, B. Smith 1984 "Truth Makers", *Philosophy and Phenomenological Research* 44, 287-321.

Pelletier, F. J. and L. K. Schubert 1989 "Mass Expressions", in D. Gabbay and F. Gnthner (eds), *Handbook of Philosophical Logic*, Reidel 1989.

Putnam, H. 1981, *Reason, Truth, and History* Cambridge: Cambridge University Press.

Simons, P. 1987, *Parts: a Study in Ontology* Oxford: Clarendon Press.

Smith, B. 1992, "Characteristica Universalis" in K. Mulligan (ed.), *Language, Truth and Ontology*, Dordrecht: Kluwer 1992.

Strawson, P. F. 1959 *Individuals: An Essay in Descriptive Metaphysics*, London: Routledge.

Wiggins, D. 1980 *Sameness and Substance*, Oxford: Blackwell 1980.

Woods, W. A. 1975 "What's in a Link: Foundations for Semantic Networks", in D. G. Bobrow and A. M. Collins (eds), *Representation and Understanding: Studies in Cognitive Science*, Academic Press 1975.

Woods, W. A. and J. G. Schmolze 1992 "The KL-ONE Family", in F. W. Lehmann (ed.), *Semantic Networks in Artificial Intelligence*, Oxford: Pergamon 1992.

List of Authors

J. M. Gooday
Division of Artificial Intelligence
School of Computer Studies
University of Leeds, Leeds LS2 9JT, England
gooday@scs.leeds.ac.uk 409

Nicola Guarino
LADSEB–CNR
National Research Council
I-35020 Padova, Italy
guarino@ladseb.pd.cnr.it 443

John Haugeland
Department of Philosophy
University of Pittsburgh
haugelan+@pitt.edu 127

Jaakko Hintikka
Philosophy Department
Boston University
Boston, MA 02215 265

Herbert Hochberg
University of Texas, Austin 193

Frank Jackson
Philosophy Program
Research School of Social Sciences
The Australian National University
fjc@coombs.anu.edu.au 101, 113

Dale Jacquette
Department of Philosophy
The Pennsylvania State University
246 Sparks Building
University Park, PA 16802 89

Eduard Marbach
University of Bern
Institute of Philosophy
Laenggassstrasse 49a
CH-3000 Bern 9
marbach@philo.unibe.ch 247

Johann Christian Marek
Universität Graz
Institut für Philosophie
Heinrichstraße 26
A–8010 Graz
marek@bkfug.kfunigraz.ac.at 139

Joseph Margolis
Temple University
Philadelphia, PA 19122 11

Rita Nolan
Department of Philosophy
State University of New York at Stony Brook
Stony Brook, NY 11777
rdnolan@ccmail.sunysb.edu 221

J. C. Nyíri
Budapest 63

Francesco Orilia
Olivetti, Research & Development
Viale Gramsci 12
56100 Pisa, Italy
Orilia@oliveb.atc.olivetti.com 37

Francis Jeffry Pelletier
Department of Philosophy
University of Alberta 311

Alberto Peruzzi
Dipartimento di Filosofia
Università di Firenze
Via Bolognese 52, 50139 Firenze
Italia 357

Jerzy Perzanowski
N. Copernicus University
Fosa Staromiejska 3
87-100 Toruń 287

Jean Petitot
EHESS/CREA, Paris
petitot@poly.polytechnique.fr 387

J. Proust
CREA, CNRS/Ecole Polytechnique
1 rue Descartes 75005 Paris
proust@poly.polytechnique.fr 233

François Récanati
CREA, CNRS/Ecole Polytechnique
1 rue Descartes 75005 Paris, France
recanati@poly.polytechnique.fr 343

Georges Rey 75
Department of Philosophy
University of Maryland
College Park, MD 20742
rey@umiacs.umd.edu

Gerhard Schurz 297
Institut für Philosophie
Universität Salzburg
Franziskanergasse 1
A-5020 Salzburg

John R. Searle 1
Philosophy Department
University of California at Berkeley
Berkeley, CA 94720
searle@cogsci.berkeley.edu

Ernest Sosa 159
Philosophy Department
Brown University
Providence, RI 02912

Neil Tennant 113, 273
Department of Philosophy and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210

Michael Tye 169
Temple University and King's College, London
tye@templevm.bitnet

Robert D. Van Valin, Jr. 371
Department of Linguistics & Center for Cognitive Science
State University of New York at Buffalo
linvan@ubvms.buffalo.edu

Achille C. Varzi 423
Istituto per la Ricerca Scientifica e Tecnologica
I-38100 Trento, Italy
varzi@irst.it

Andrew Woodfield 319
Department of Philosophy
University of Bristol

Appendix: R. Casati and G. White (eds.), *Philosophy and the Cognitive Sciences*

(Contributions of the Ludwig Wittgenstein Society I)
Kirchberg am Wechsel
The Austrian Ludwig Wittgenstein Society 1993

Distributors

To be ordered from:
The Austrian Ludwig Wittgenstein Society
Markt 2, A-2880 Kirchberg am Wechsel
Austria

Table of Contents

Schriftenreihe der Wittgenstein-Gesellschaft

Hrsg. Rudolf Haller, Elisabeth Leinfellner, Werner Leinfellner, Paul Weingartner

SWG Band 20/2

WITTGENSTEINS PHILOSOPHIE DER MATHEMATIK

Akten des 15. Internationalen Wittgenstein-Symposiums (Teil 2)

WITTGENSTEIN'S PHILOSOPHY OF MATHEMATICS

Proceedings of the 15th International Wittgenstein-Symposium (part 2)
Kirchberg am Wechsel, 16. – 23.August 1992
Hrsg. Klaus Puhl
Wien 1993, 315 Seiten, Leinen; ISBN 3-209-01592-9
DM 98,– / SFr 91,– / öS 680,–

Inhalt/Contents

Verlag Hölder-Pichler-Tempsky
A-1096 Wien, Frankgasse 4
Tel.(0043)-1-438993 / FAX (0043)-1-43899385

Veröffentlichungen des Instituts Wiener Kreis
Herausgegeben von Friedrich Stadler

Band 1
JOUR FIXE DER VERNUNFT
Der Wiener Kreis und die Folgen

Hrsg. Paul Kruntorad
unter Mitwirkung von Rudolf Haller und Willy Hochkeppel
Wien 1991; 294 Seiten, Broschur, ISBN 3-209-01221-0
DM 57,—/SFr 53,—/öS 396,—

Jeden Donnerstag-Abend traf sich ein privater Diskussionszirkel im Mathematischen Institut der Universität Wien und schuf dabei neue Standards, an denen sich jede künftige Philosophie, die als rationale auftreten will, orientieren muß.
Aus dem Logischen Empirismus des „Wiener Kreises" der Zwischenkriegszeit entwickelte sich die weltweit in vielfältiger Weise ausgeprägte analytische Philosophie.
Der Band gibt einen Einblick in Entstehung, Arbeitsweise und Wirkungsgeschichte einer der bedeutendsten philosophischen Gruppierungen des 20. Jahrhunderts.

INHALTSÜBERSICHT:

Verlag Hölder-Pichler-Tempsky
A-1096 Wien, Frankgasse 4
Tel.(0043)-1-438993 / Fax (0043)-1-43899385