# Was dürfen wir glauben?
# Was sollen wir tun?

**Sektionsbeiträge des achten internationalen Kongresses der *Gesellschaft für Analytische Philosophie* e.V.**

**Herausgegeben von**
**Miguel Hoeltje, Thomas Spitzley und Wolfgang Spohn**

gap
gesellschaft für analytische philosophie

Was dürfen wir glauben? Was sollen wir tun?
Sektionsbeiträge des achten internationalen Kongresses der
Gesellschaft für Analytische Philosophie e.V.

Herausgegeben von
Miguel Hoeltje, Thomas Spitzley und Wolfgang Spohn

Online-Veröffentlichung der
Universität Duisburg-Essen (DuEPublico) 2013
ISBN 978-3-00-042332-1

# Vorwort

Vom 17.–20.9. 2012 fand unter dem Titel „Was dürfen wir glauben? Was sollen wir tun?“ und unter der Schirmherrschaft von Frau Ministerin Prof. Schavan (Deutschland), Herrn Minister Prof. Töchterle (Österreich) und Herrn Staatssekretär Dr. Dell'Ambrogio (Schweiz) in Konstanz der achte internationale Kongress der Gesellschaft für Analytische Philosophie statt. Neben rund 35 eingeladenen Sprecherinnen und Sprechern, die in einer Reihe von Hauptvorträgen und Kolloquien zu Wort kamen, gab es in den acht thematischen Sektionen des Kongresses insgesamt mehr als 230 Vorträge und Poster-Präsentationen. Mit rund 450 Einreichungen für Sektionsbeiträge war die Beteiligung außergewöhnlich hoch. Der vorliegende Sammelband umfasst nun 61 auf solchen Sektionsbeiträgen basierende Artikel.

Ein so großer Kongress hätte nicht ohne die Beteiligung und Mithilfe vieler Menschen erfolgreich stattfinden können. Neben den Mitgliedern des GAP-Vorstandes und des GAP.8-Programmkomitees, die im Vorfeld die Planung übernommen hatten, sind hier die Mitglieder der Jurys für den Wolfgang-Stegmüller- sowie den Ontos-Preis zu nennen. Ebenfalls ist den Gutachterinnen und Gutachtern zu danken, die sich der schwierigen Aufgabe angenommen hatten, aus der übergroßen Zahl der Einreichungen die Sektionsbeiträge auszuwählen. Vor Ort in Konstanz haben ganz besonders die Kongressassistentin Gabriele Hahn und ihr Team (Wolfgang Egner, Sandra Vatter u.v.m) für eine hervorragende Planung und einen reibungslosen Ablauf der Konferenz gesorgt. Ohne die Unterstützung der DFG sowie der Universität Konstanz wäre der Kongress so nicht möglich gewesen – auch ihnen gebührt unser Dank. Ferner ist den Sponsoren von GAP.8 zu danken: dem Schweizer Staatssekretariat für Bildung und Forschung, den Fischer-Werken GmbH & Co KG, der Sparkasse Konstanz, den Verlagen de Gruyter, Klostermann, Meiner, Mentis, Ontos, Oxford University Press, Reclam, Springer, Suhrkamp und Synchron Publishers, dem philosophie-Magazin, und schließlich der Stiftung Wissenschaft und Gesellschaft, der Universitätsgesellschaft und dem Verein der Ehemaligen der Universität Konstanz.

An der organisatorischen Vorbereitung und Erstellung dieses Sammelbandes von Sektionsbeiträgen hatte niemand so großen Anteil wie unsere studentische Hilfskraft Katharina Lührmann – für die hervorragende Arbeit möchten wir schließlich ihr ganz herzlich danken!

Die Herausgeber

Miguel Hoeltje, Thomas Spitzley, Wolfgang Spohn

## Inhalt


**1. Sprachphilosophie** **6**

Don't Ask, Look! Linguistic Corpora in Philosophical Analyses 7
Roland Bluhm

Rede über fiktive Kontexte 16
David B. Blumenthal

The Ineliminability of Non-Nominal Quantification 32
David Dolby

Primitive Normativität als Antwort auf den Regelfolgen-Skeptiker 39
Nadja El Kassar

Relativism and Superassertibility 47
Manfred Harth

Has Vagueness Really No Function in Law? 60
David Lanius

A Single-Type Ontology for Natural Language 70
Kristina Liefke

Relevanz anstatt Wahrheit? 85
Theresa Marx

**2. Metaphysik und Ontologie** **95**

The Fundamental Question of Metaphysics and the Question of Fundamentality in Metaphysics 96
Brandon C. Look

Why Dispositions Are Not Higher-order Properties 104
Joshua Mugg

The Point of Action 111
Michael Oliva Córdoba

Bennett on Dismissivism 115
Laura Cecilia Porro

**3. Logik und Wissenschaftstheorie** **125**

Regularity Theories of Mechanistic Constitution in Comparison 126
Jens Harbecke

Vage natürliche Arten 135
Rico Hauswald

Epistemische und nicht-epistemische Werte in der angewandten Forschung 148
Gertrude Hirsch Hadorn

Causation, Dispositions, and Mathematical Physics 162
Johannes Röhl

Between Relativism and Absolutism? – The Failure of Kuhn's Moderate Relativism 172
Markus Seidel

When Is It Rational to Believe a Mathematical Statement? 186
Jendrik Stelling

Statistical and Non-Statistical Normality 199
Corina Strößner

**4. Philosophie des Geistes** **210**

Theory of Mind as Gradual Change Guided by Minimalisms 211
Gerhard Chr. Bukow

Mechanistische Erklärung: Reduktiv oder nicht? 224
Bettina Gutsche

Phenomenal Concepts - Still Battling the Bewilderment of Our Intelligence 236
Max Mergenthaler Canseco

Ein Dilemma für modale Argumente gegen den Materialismus 250
Sebastian J. Müller

How We Know Our Senses 256
Eva Schmidt

The *arche* of Cognition – Grounding Representations in Action 264
Arne M. Weber & Gottfried Vosgerau

Integrating Evaluation and Affectivity into the Intentionality of Emotions 278
Wendy Wilutzky

Nichtwillentliche Aktivität 287
André Wunder

**5. Erkenntnistheorie** **297**

Explanatorisches Verstehen: Ein Definitionsvorschlag 298
Christoph Baumberger

How Gettier Helps to Understand Justification 312
Frank Hofmann

Contextualism and Gradability – A Reply to Stanley 318
Romy Jaster

Intuitions, Heuristics, and Metaphors: Extending Cognitive Epistemology 324
Eugen Fischer

What are Epistemic Duties? 340
Andrea Kruse

The Method of Reflective Equilibrium and Intuitions 352
Julia Langkau

Why Know-how and Propositional Knowledge Are Mutually Irreducible 365
David Löwenstein

Interrogative Formen des Wissens und reduktiver Intellektualismus 372
Pedro Schmechtig

Practical Knowledge 392
Michael Schmitz

**6. Ästhetik und Religionsphilosophie** **404**

Combining Bayesian Theism with Pascal's Wager 405
Stamatios Gerogiorgakis

Zur Rechtfertigung religiöser Überzeugungen durch pragmatische Argumente 412
Christoph Kurt Mocker

Kunst und Moral 418
Lisa Katharin Schmalzried

Praemotio physica und leibnizianischer Molinismus 435
Ruben Schneider

**7. Angewandte Ethik, politische Philosophie, Rechts- und Sozialphilosophie** **450**

Problems of Advance Directives in Psychiatry 451
Simone Aicher

Bildung als Gegenstand der fairen Chancengleichheit bei Rawls 465
Claudia Blöser

Liberalismus, Handlungsfreiheit und Autonomie 477
Christine Bratu

Im Namen der Autonomie? Eine kritische Untersuchung des liberalen Paternalismus am Beispiel von Maßnahmen des kognitiven Enhancements 489
Rebecca Gutwald

Zum Begriff des Kindeswohls: Ein liberaler Ansatz 501
Christoph Schickhardt

Erbschaftssteuern, Obduktionen und die postmortale Konfiszierung von Organen 507
Christoph Schmidt-Petri

Two Problems with the Socio-Relational Critique of Distributive Egalitarianism 525
Christian Seidel

The Role of Economic Analysis in Combating Climate Change 536
Joachim Wündisch

**8. Normative Ethik, Metaethik, Handlungs- und Entscheidungstheorie** **548**

Defending Moral Intuitionism Against Debunking Arguments 549
Anne Burkard

Overdetermination in Inuitive Causal Decision Theory 559
Esteban Céspedes

Double Effect and Terror Bombing 573
Ezio Di Nucci

Counterfactuals and Two Kinds of *Ought* 588
Daniel Dohrn

Thomas Buddenbrook und der Vorrang der Moral 593
Martin Hoffmann

Practical Knowledge 607
David Horst

Sollen, Können und Versuchen 613
Michael Kühler

Drei Arten von Hilfspflichten 623
Jörg Löschke

Willensschwäche – Eine Systematisierung und eine Erklärung 638
Christoph Lumer

The Case against Consequentialism: Methodological Issues 654
Nikil Mukerji

Moralischer Zufall und Kontrolle 666
Julius Schälike

What Makes Moral Values Queer? 676
Julius Schönherr

Konsequentialistische Theorien und der Besondere-Pflichten-Einwand 690
Marcel Warmt

# 1. Sprachphilosophie

# Don't Ask, Look!
# Linguistic Corpora in Philosophical Analyses

Roland Bluhm

Ordinary Language Philosophy has largely fallen out of favour, and with it the belief in the primary importance of analyses of ordinary language for philosophical purposes. Still, in their various endeavours, philosophers not only from analytic but also from other backgrounds refer to the use and meaning of terms of interest in ordinary parlance. In doing so, they most commonly appeal to their own linguistic intuitions. Often, the appeal to individual intuitions is supplemented by reference to dictionaries. In recent times, Internet search engine queries for expressions of interest have become quite popular. Apparently, philosophers attempt to surpass the limits of their own linguistic intuitions by appealing to experts or to factual uses of language. I argue that this attempt is commendable but that its execution is wanting. Instead of appealing to dictionaries or Internet queries, philosophers should employ computer-based linguistic corpora in order to confirm or falsify hypotheses about the factual use of language. This approach also has some advantages over methods employed by experimental philosophers. If the importance of ordinary language is stressed, the use of linguistic corpora is hardly avoidable.

## 1. Introduction

In linguistics, or, more particularly, in lexicography, using text corpora is a well-established practice. The *Oxford English Dictionary*, to take the most famous example, was based on an enormous corpus of paper slips with excerpted quotations. Today, the text corpora used in linguistics are usually computer-based, and at least some are freely accessible on the Internet. Surprisingly, they seem to have been disregarded by philosophers, even those that profess an interest in ordinary language.

Over the last couple of years, I have made extensive use of such corpora, mostly in my research on hope.[1] But although I have used corpora for some time, I have only recently begun to describe explicitly their use in philosophy. The purpose of this paper is to take on this challenge and to recommend the use of linguistic text corpora for philosophical purposes—or, in a slogan, to advertise Computerised Ordinary Language Philosophy.

I will begin, in sections 1 to 3, by spelling out my reasons for advocating the use of corpora in philosophy. I will introduce a very simple model of linguistic analysis (for philosophical purposes). This model allows me to point out the roles of intuition in such analyses and to highlight the benefits of using corpora in philosophy. In section 4, I will then discuss some other options, namely the use of Internet queries and questionnaires. I will round off my account (in section 5) with a qualified plea for using corpora in philosophy.[2]

---

[1] Cf. Bluhm 2012.

[2] I would like to thank my audiences at SOPHA 2012 and GAP.8 for valuable feedback; I thank Peter Hacker especially for a lively discussion of my presentation. As such things go, I very stubbornly insist on my main points, but I have profited greatly from his critique.

## 2. A Simple Model of Linguistic Analysis in Philosophy

Ordinary Language Philosophy, it seems, has largely fallen out of favour, and with it the belief in the primary importance of analyses of ordinary language. Perhaps rightly so. Yet philosophers not only from analytic but also from other backgrounds still consider the ways in which terms of philosophical interest are used in ordinary, non-technical language. In the analytic tradition this practice is, of course, still quite common.

I suspect that some of my readers would be more interested in the justification for doing Ordinary Language Philosophy than in the use of corpora for its purposes. Lest there be disappointment later, let me make clear that I am *not* concerned with this question here. I will give reasons for using corpora, but not for doing Ordinary Language Philosophy. Let me also stress that I do *not* want to advocate Computerised Ordinary Language Philosophy as *the one* method to be employed in philosophy. That would be rather silly. But let us assume for the sake of this exposition that we *are* committed to the analysis of ordinary language for philosophical purposes.

What I would like to do first is to present a model of the practice of such analyses. The model gives a very simplified account of reality, but it will allow me to point out what benefits the use of corpora has.

I believe that analyses of ordinary language in philosophy proceed more or less as follows.

<table>
<tr><td></td><td colspan="2">Steps in the practice of ordinary language analyses</td></tr>
<tr><td></td><td colspan="2">Formation of research interest and hypotheses</td></tr>
<tr><td>1</td><td colspan="2">Decision which expressions are pertinent</td></tr>
<tr><td>2</td><td colspan="2">Formation of hypotheses related to pertinent expressions</td></tr>
<tr><td>3</td><td>Coming up with an example for the use of a pertinent expression</td><td rowspan="3">i.e., testing and refining of hypotheses related to pertinent expressions</td></tr>
<tr><td>4</td><td>Analysis of the example</td></tr>
<tr><td>5</td><td>Iterations and variations</td></tr>
<tr><td>6</td><td colspan="2">Drawing conclusions</td></tr>
</table>

Let me shortly comment on the six steps.

The process of analysis is preceded by the formation of a research interest and the (perhaps tacit) formation of hypotheses related to the issue to be addressed. I am going to disregard this phase in the following considerations (which is why it is not numbered).

1: Usually the task of linguistic analysis in philosophy is described as the attempt to analyse or to explicate a concept. I am not concerned here with the question of what a concept is. However, it is hardly bold to claim that a concept usually can be expressed in various ways in a given language. Therefore, it is not always obvious *which* linguistic phenomena are pertinent for the analytical process. Thus, if we take seriously the idea of approaching some philosophical problem through an analysis of ordinary language, we first need to clarify which expressions are to be considered at all.

2: We form (perhaps tacit) hypotheses about the use of the pertinent expressions.

3–5: We then test and refine these hypotheses. We come up with contexts in which some expression is to be examined, that is, with some utterance (spoken or written) in which the

pertinent expression features prominently, or with a more or less hypothetical setting, an example story, in which the expression of interest plays a significant role. Ideally, the next step is the interpretation or analysis of whatever has been thought of in step 3. Steps 3 and 4 are then repeated for additional expressions, and potentially interesting findings are examined through various iterations and variations.

6: The process ends, ideally, in conclusions being drawn from the results of steps 1 to 5.

## 3. Intuition in Linguistic Analysis

In the course of this six-step process, philosophers have to appeal to their linguistic intuitions. Let me emphasise that 'intuition' here refers to *linguistic competence* in the object language, rather than to a certain kind of belief. It would, of course, be possible to use the label 'intuition' for beliefs that are based on this competence. But I am more concerned here with drawing attention to the *source* than to the result of linguistic judgment.

### *3.1 Some Problems of Intuition*

Intuition, in the sense of linguistic competence, has to be employed in different ways in the different phases of analysis. Here is another very sketchy overview.

<table>
<tr><td></td><td colspan="2">Steps in the practice of ordinary language analyses</td><td>Type of intuition</td></tr>
<tr><td></td><td colspan="2">Formation of research interest and hypotheses</td><td></td></tr>
<tr><td>1</td><td colspan="2">Decision which expressions are pertinent</td><td>Recall/interpretation</td></tr>
<tr><td>2</td><td colspan="2">Formation of hypotheses related to pertinent expression</td><td></td></tr>
<tr><td>3</td><td>Coming up with an example for the use of a pertinent expression</td><td rowspan="3">i.e., testing and refining of hypotheses related to pertinent expressions</td><td>Recall/imagination</td></tr>
<tr><td>4</td><td>Analysis of the example</td><td>Interpretation/evaluation</td></tr>
<tr><td>5</td><td>Iterations /variations</td><td>Recall/imagination</td></tr>
<tr><td>6</td><td colspan="2">Drawing conclusions</td><td></td></tr>
</table>

Again, some comments are in order.

1: In order to think of words pertinent to a specific analytical task (in step 1), one has to employ one's *active* knowledge of the object language: one has to *recall* pertinent linguistic phenomena. And one has to understand, that is, to *interpret* these phenomena, which is an exercise of intuition in a *passive* sense.

3: In a partly similar way, in order to come up with examples, one has to recall or to *imagine* contexts in which the expression can be used.

4: In step 4, passive knowledge of the object language takes priority: the ability required here is that of *interpreting* a given example utterance (with or without context). And one also needs to *evaluate* whether the pertinent expression is acceptable in the example utterance (with or without context). This is on one hand a matter of self-control. Since step 3 is informed by step 2, there is the danger that one may come up with biased examples. On the

other hand, philosophers often would like to know in which way an expression *cannot* be used. In this case, the evaluation of acceptability is of primary importance.

5: Coming up with variations, again, requires *active* linguistic competence.

2 and 6: Ideally, hypotheses and conclusions are not drawn from one's active or passive knowledge of the object language. Probably what is called for in these steps is the faculty of *judgment*.

In the process I have outlined, the reliance on intuition in the sense of *active* knowledge of the object language is problematic in two ways. First, the one who thinks of the examples usually has an investment in specific hypotheses that might bias his examples. Second, everyone's linguistic competence, especially the active part of it, is limited, and can therefore serve only as a limited source of data. This point is also pertinent to the historical Oxford-based Ordinary Language Philosophy, whose proponents sometimes seem to have confused their very sophisticated variety of English with English as such. Also, it is worth mentioning that philosophers who are not native speakers of English have limited linguistic competence and are thus disadvantaged with respect to linguistic analysis of English expressions. This is problematic, since English is, more often than not, the object language of philosophical research.

You might think that my six-step schema misrepresents what is usually done. I readily admit that I have idealised the facts considerably. Most importantly, there are several ways in which philosophers do try to overcome the limits of their own intuition.

### *3.2 Arming the Armchair*

The most time-honoured way to overcome the limits of one's linguistic intuition is simply to consult a dictionary.

J. L. Austin recommended using dictionaries to identify the pertinent expressions for an analytical task at hand. He suggested simply sifting through the dictionary from A to Z to identify expressions of interest. What he might have had in mind, over and above a simple reminder of words one cannot recall instantly, was that dictionaries contain information on word families and word fields. *Word families* consist of words that are etymologically related—those whose lexical roots share a common ancestry. For example, the word family of 'hope' also contains 'hopeful', 'hopelessly', 'unhope', 'wanhope', etc. *Word fields*, on the other hand, contain words with related meanings. The word field of 'hope' contains 'desire', 'wish', 'belief', and 'expectation', but also 'fear', 'despair', etc. Word families and fields are explicitly listed only in special dictionaries, but they form the basic grid of lexicology, and information on them can be gleaned even from alphabetically organised dictionaries.

Dictionaries also, of course, contain accounts of the presumed meanings of words. And although the definitions sought in philosophy differ in function, focus, and degree of precision from the paraphrases of meanings given in lexicography,[3] dictionaries may at least be helpful for formulating preliminary hypotheses about meanings.

However, dictionaries are not to be trusted unquestioningly.

First of all, dictionaries are not without error. Some mistakes may be individual and unsystematic; others, especially omissions, may have systematic causes. For example, not all dictionaries are descriptive. Especially some older dictionaries are normative to some degree. And *all* dictionaries represent the choice of some material over other material. Not everything can be recorded. One way in which this is relevant is that most dictionaries do not record recent developments.

[3] Cf. Wiegand 1989.

Also, new dictionaries partly rely on older dictionaries to supply information about phenomena of language and their interpretation.[4] To put it crudely, dictionary writers crib what other dictionary writers have written. This is hardly avoidable for reasons of economy, and it is also a reasonable thing to do: it is a scientific virtue to preserve knowledge that has already been gained. But there is no way to know, when consulting a dictionary, to what extent the authors of the dictionary have checked the material that they have inherited from their predecessors—to what extent they are preserving not only past knowledge but past mistakes.

Finally, it is important to note that dictionaries rely on intuitions at various points. Older dictionaries, such as the *Oxford English Dictionary*, relied on quotations that were collected by informants and thus relied on the judgment and the passive linguistic competence of those informants. The collected quotations were then processed by the dictionary's writers and editor, who have left their mark on the entries, as well.

## 4. The Benefits of Using Corpora in Linguistic Analysis

If ordinary language is important with respect to some philosophical endeavour, and if we want to avoid the potential errors I have pointed out, we need some basis on which our intuitions (as well as dictionaries' information) can be tested, corrected, and extended. More particularly, we need independent, and thus unbiased, evidence that expressions in which we are interested are used in certain ways. Also, we need an independent basis for testing our hypotheses about the use of these expressions.

Linguistic text corpora can serve these functions and more. Before I go into that, let me briefly indicate what a corpus is.

### *4.1 Linguistic Text Corpora*

Regrettably, a wholly convincing definition of 'corpus' is difficult to obtain. A very wide characterisation is as follows:

> We define a corpus simply as "a collection of texts." If that seems too broad, the one qualification we allow relates to the domains and contexts in which the word is used rather than its denotation: *A corpus is a collection of texts when considered as an object of language or literary study*. (Kilgarriff and Grefenstette 2003: 334)

Other rather broad characterisations point to collecting principles to distinguish corpora from mere collections of text. But it is doubtful whether these are clear criteria:

> If a corpus is defined as a principled or structured [...] collection of texts, it has to be distinguished from a more arbitrary collection of material or "text database". [...] The borderline between a well-defined corpus and a random collection of texts is unlikely to be a clear-cut one, however. (Hundt 2008: 170)

Let us just say that a corpus is a collection of texts (written or spoken) that serves as a primary database for supplying evidence with respect to some linguistic question. That might not be a fully satisfactory definition, but it will suffice for the present purpose.

The more sophisticated corpora are also annotated; they contain information, for example, on parts of speech.

There are many corpora that are freely accessible for scientific purposes. By way of example, let me name two suitable ones.[5] For British English, there is the British National Corpus

[4] Cf. Bergenholtz and Schaeder 1985: 292.

[5] Comprehensive lists can be found in, e.g., Lee 2010 and Xiao 2008.

(BNC); for American English, the Corpus of Contemporary American English (COCA).[6] The BNC is a relatively large, closed corpus of texts of written and spoken language. It contains approximately 100 million words in texts dating from 1960 to 1994.[7] COCA is not closed; every year approximately 20 million words are added. At present, the corpus contains about 450 million words from more than 175,000 texts dated from 1990 to the present (2012).[8] Both BNC and COCA are freely accessible for scientific purposes. The essentially identical search interfaces for both are provided by Brigham Young University in Provo, Utah. The available search algorithms are quite powerful, allowing queries for exact strings as well as lemmata (i.e., words disregarding inflexions, such as 'hope', 'hopes', 'hoped', and 'hoping'). The corpora are annotated, allowing queries for strings of certain grammatical categories (e.g., 'hope' as a noun vs. a verb). It is possible to search for the co-occurrence of expressions within a distance of 10 words (unfortunately, this function ignores sentence boundaries).

### *4.2 Four Benefits of Using Corpora*

So, what are the benefits of using corpora?

First of all, corpora provide data on the basis of which hypotheses can be formulated, they provide data to confirm or falsify hypotheses and conclusions from the analytical process, and they provide data that can be used to exemplify or illustrate specific usages of interest.

And, secondly, all of these data are, by and large, unfiltered.

Not all corpora fulfil all of these functions equally well, of course. All corpora provide some linguistic context for the queried expressions, but wider contexts (more than one or two sentences) are not always provided. There are several corpora that, for copyright reasons, do not give free access to the texts that constitute their basic data. Yet a relatively thorough consideration of context may be required to formulate substantial and interesting hypotheses, especially when an analytical task is first approached. And sometimes the meaning of a word can be understood only when the wider context of its use is known.

Another important property of corpora is size. Hypotheses are best tested with a large corpus that, due to its size, contains rare linguistic phenomena. It is important to keep in mind that hypotheses claiming the non-existence of some phenomenon cannot be proved, and hypotheses claiming the existence of some phenomenon cannot be disproved by a corpus analysis. However, if a corpus is very large and comprises a balanced mixture of texts, we can base at least tentative claims about the existence or non-existence of phenomena on it.

Regardless of the width of the context provided, there are two further benefits of using corpora.

Thirdly, the contexts in which the queried expressions are found give insights into the variety of real-life situations in which the phenomenon referred to by the concept occurs.

And finally, the contexts often provide excellent raw material for thought experiments with regard to the concept *and* the phenomenon in question.

## 5. A Few Remarks on Other Options

There are two alternatives to the use of corpora that I would like to address briefly. The first is the use of Internet search engine queries.

---

[6] The corpora are accessible at http://corpus.byu.edu/bnc/and http://corpus.byu.edu/coca/.

[7] The online interface dates the texts in BNC to "1970s–1993" (cf. http://corpus.byu.edu/bnc/). My deviation from this information is based on Leech, Rayson, and Wilson 2001: 1.

[8] For the sake of comparison, the largest accessible German corpus, DEREKO, contains 10 million texts with 2.3 billion ($10^9$) words.

### *5.1 Web as Corpus*

In recent times, it has become somewhat popular to search the Internet to find out how an expression of interest is actually used. Apparently, the attempt is to surpass the limits of one's own linguistic intuitions by appeal to factual language uses. This attempt is commendable, I believe, but its execution is wanting.

The Internet is indeed an alluring data source because of its sheer size and the relative ease with which data can be compiled from it. A considered estimate from 2008 calculates that 60% of the generally accessible Internet is in English, consisting of an estimated 3 trillion ($10^{12}$) word tokens.[9] Access to all these data is not only relatively easy, but mostly free of charge. However, using the Internet as a corpus by accessing it with one of the common general search engines is problematic in a number of respects.[10]

Most importantly, the common Internet search engines offer only few (and changing) search algorithms. The Internet is also not linguistically annotated, and thus lacks information that could be employed for sophisticated hypothesis testing. Internet search engines do not allow the deduction of reliable statistical information, at least not without refined methods. (They list page hits instead of tokens of the queried expression, and they yield different search results at different times.) The number of queries that are allowed on one day with a given search engine is limited. And last but not least, English (or some language that resembles English) is used on the Internet by a large number of speakers who have limited linguistic competence because it is not their native language. The fact that, for example, certain constructions are used is not necessarily good evidence for their acceptability.

All in all, to simply use the web as a corpus is not advisable. The alternative is to use the web as raw material for building a corpus. This process would involve refining the data gathered from the Internet and is a completely different kettle of fish. However, although using the web for building a corpus is respectable, that is not what is done when philosophers type a query into a general search engine.

### *5.2 Questionnaires*

The second alternative to the use of corpora, which has recently been put to sophisticated use by experimental philosophers, is to ask informants to answer a questionnaire in a controlled setting in order to obtain their views on how a specific concept is used. One can do so either indirectly, by asking whether certain constructions are or are not objectionable, or directly, by asking how the informants would characterise the meaning of a specific concept.

One thing to be said in favour of employing questionnaires is that they supplement or even substitute for the researcher's intuitions regarding the use of expressions under discussion. Another is that questionnaires allow researchers to pose questions regarding very specific and infrequent uses of expressions.

Against the practice, it must be noted that questionnaires usually are given to a very limited number of test subjects and therefore do not necessarily solve the problem of limited active linguistic competence. However, the major flaw of the questionnaire method is to draw the informants' attention to their use of the language and to thereby invite answers that do not provide information on how informants in fact use a specific concept, but on how they *believe* they use or should use the concept.

To give an example: Patricia Bruininks and Bertram Malle found that test subjects associated more important objects with 'hope' than they did with 'optimism', 'desire', 'wanting', and

[9] Cf. Bergh and Zanchetta 2008: 313.

[10] Cf. Kilgarriff 2007 and Bergh and Zanchetta 2008 for the following.

'wishing'.[11] I believe that this result is due to a bias effect. There is a powerful ideology of hope. People seem to think that hope *should* be for something important, where 'important' may be read subjectively as 'something one attaches great importance to' or objectively as 'something that is worth attaching great importance to'. But that is not the way the word is in fact used. It is also used in cases of tepid hope and hope for trivial things. Thus, the use of the word in *unprompted* contexts does not seem to bear out the tendency that is evident in the questionnaire.

## 5. On the Advisability of Using Corpora

Do all of us, then, have to do Computerised Ordinary Language Philosophy? In the beginning I have said no. But the answer is rather yes *and* no.

No, because the method is not suitable for addressing all philosophical questions. I do not believe that all philosophical issues can be reduced to questions that can be answered by conceptual analysis. And even if we *could* narrow it all down to conceptual analysis, this would not imply that it is paramount to consider ordinary language.

We must also observe that analysing the use of a word field or a word family with the help of a linguistic text corpus is a lot of work; this work should be undertaken only if it promises worthwhile results.

On the other hand, the answer to the question whether all of us have to do Computerised Ordinary Language Philosophy is: yes, at least those of us who claim that ordinary language is important. Wittgenstein famously warned:

> Eine Hauptursache philosophischer Krankheiten – einseitige Diät: man nährt sein Denken mit nur einer Art von Beispielen. [A main cause of philosophical disease—a one-sided diet: one nourishes one's thinking with only one kind of example.] (*PU* 593/*PI* 593)[12]

Relying on one's own intuition in linguistic analyses is very likely to result in such a one-sided diet. One way to make the diet more balanced is to avoid making up examples of how language *allegedly* is used and to confront oneself with language as it is used *in fact*. But if the importance of the actual use of ordinary language is stressed in this way, the use of linguistic corpora is hardly avoidable.

**Roland Bluhm**

Technische Universität Dortmund
Roland.Bluhm@tu-dortmund.de

## References

Bergenholtz, H., and B. Schaeder 1985: 'Deskriptive Lexikographie', in L. Zgusta (ed.): *Probleme des Wörterbuchs*, Darmstadt: Wissenschaftliche Buchgesellschaft, 277–319.

Bergh, G., and E. Zanchetta 2008: 'Web Linguistics', in A. Lüdeling and M. Kytö (eds.) 2008–2009, Vol. 1: 309–327.

Bluhm, R. 2012: *Selbsttäuscherische Hoffnung*, Münster: mentis.

[11] Cf. Bruininks and Malle 2008: 348f.

[12] The title of this paper is, of course, also an allusion to Wittgenstein. In this case, to his famous injunction "Wie gesagt: denk nicht, sondern schau! [To repeat: don't think, but look!]" (*PU* 66/*PI* 66).

Bruininks, P., and B. Malle 2005: 'Distinguishing Hope from Optimism and Related Affective States', *Motivation and Emotion* 29, 327–355.

Hundt, M. 2008: 'Text Corpora', in A. Lüdeling and M. Kytö (eds.) 2008–2009, Vol. 1: 168–187.

Kennedy, G. 1998: *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff, A. 2007: 'Googleology Is Bad Science', *Computational Linguistics* 1, 1–5.

Kilgarriff, A., and G. Grefenstette 2003: 'Introduction to the Special Issue on the Web as Corpus', *Computational Linguistics* 29, 333–347.

Lee, D. Y. W. 2010: 'What Corpora Are Available?', in M. McCarthy and A. O'Keeffe (eds.): *Corpus Linguistics*, London: Routledge, 107–21.

Leech, G., P. Rayson, and A. Wilson 2001: *Word Frequencies in Written and Spoken English*. London: Longman.

Lüdeling, A., and M. Kytö (eds.) 2008–2009: *Corpus Linguistics*, 2 vols., Berlin: de Gruyter.

*PI* = Wittgenstein, L. 1968: *Philosophical Investigations*, transl. by G. E. M. Anscombe, 3rd ed., Oxford: Blackwell.

*PU* = Wittgenstein, L. 1960: *Philosophische Untersuchungen*. Frankfurt am Main: Suhrkamp.

Wiegand, H. E. 1989: 'Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch', in F. J. Hausmann, O. Reichmann, H. E. Wiegand, and L. Zgusta (eds.) 1989–1991: *Wörterbücher*, 3 vols., Berlin: Walter de Gruyter, Vol. 1: 530–588.

Xiao, R. 2008: 'Well-Known and Influential Corpora', in: A. Lüdeling and M. Kytö (eds.) 2008, 383–457.

# Rede über fiktive Kontexte

David B. Blumenthal

Der Aufsatz zielt darauf ab, ein angemessenes Verständnis der Semantik von Äußerungen über fiktive Kontexte (kurz: AFKs) zu entwickeln. Der systematische Ausgangspunkt der Arbeit besteht dabei in einem pragmatistischen Inferentialismus à la Robert Brandom. Für diese theoretische Weichenstellung wird nicht gesondert argumentiert, sondern vielmehr geltend gemacht, dass aus ihr zwei Forderungen an eine angemessene Theorie der Semantik von AFKs ableitbar sind. Erstens muss eine solche Theorie den inferentiellen Beziehungen Rechnung tragen, in denen von AFKs gemachte Aussagen de facto stehen. Zweitens darf sie nur auf solche Gegenstände zurückgreifen, die insofern ontologisch unschuldig sind, als sie vollständig und individuierbar sind. Aus diesen Forderungen ergibt sich, dass klassische Theorien der Semantik von AFKs unbefriedigend sind: Weder können AFKs mit Bertrand Russell als Äußerungen über das inferentielle Potenzial bestimmter Mengen fiktionaler Medien betrachtet, noch mit Searle, van Inwagen oder Parsons als Äußerungen über fiktive Gegenstände verstanden werden. Im Anschluss an diese kritische Auseinandersetzung wird ein eigener Vorschlag entwickelt, dessen Kerngedanken darin bestehen, AFKs als Äußerungen über bestimmte Werke zu betrachten, die wesentlich auf Interpretation beruhen. Dabei werden Werke als Äquivalenzklassen fiktionaler Medien verstanden, die logische Feinstruktur von AFKs gemachter Aussagen nach dem Modell von De-dicto-Zuschreibungen erläutert und deren Funktionsweise wiederum inferentialistisch gefasst.

## 1. Gegenstand, Vorhaben, Vorgehensweise

In der vorliegenden Arbeit geht es mir um die Semantik von Äußerungen, die wir benötigen, um uns über fiktive Kontexte zu unterhalten. Unter fiktiven Kontexten verstehe ich erzählte Welten, wie sie durch Romane, Spielfilme, Comics, Sagen oder mündliche Erzählungen konstituiert werden. Beispiele für diejenigen Äußerungen, um die es mir in der vorliegenden Arbeit geht, sind also meine Äußerung der Aussage „Asterix ist ein Gallier“ und deine Bekundung, Herkules sei der Sohn des Zeus. Ich werde diese Äußerungen „Äußerungen über fiktive Kontexte“ oder kurz „AFKs“ nennen. Ziel der Arbeit wird es sein, den lokutionären Bestandteil von AFKs zu analysieren, auf welchen ich mich unter Verwendung des Ausdrucks „von einer AFK gemachte Aussage“ beziehen werde. Die Kernfrage meiner Arbeit lässt sich dann so stellen: Was ist eine angemessene Analyse der Semantik derjenigen Aussagen, die wir machen, wenn wir uns über fiktive Kontexte äußern? Oder kürzer: Was sagen wir, wenn wir eine AFK gebrauchen.[1]

Um das Gegenstandsgebiet meiner Untersuchung genauer zu umreißen, möchte ich klarstellen, dass ich mich nicht um fiktionale Aussagen kümmern werde. Unter fiktionalen Aussagen verstehe ich Aussagen, die in Spielfilmen, Romanen, Comics oder Sagen enthalten sind und fiktive Kontexte erzeugen. Eine Analyse fiktionaler Aussagen könnte beispielsweise danach fragen, was fiktionale Aussagen von nicht-fiktionalen unterscheidet, und wie es fiktionalen Aussagen gelingt, fiktive Welten ins Leben zu rufen. Solche Fragen sind nicht die Fragen dieser Arbeit. Mir geht es darum aufzuklären, wie diejenigen Äußerungen funktionieren, mit denen wir uns über bereits bestehende fiktive Kontexte unterhalten.

[1] Um umständliche Formulierungen zu vermeiden, werde ich beizeiten kurz von der Bedeutung von AFKs anstelle der Bedeutung von AFKs gemachter Aussagen reden. Dies ist insofern unproblematisch, als der nicht-lokutionäre Anteil von AFKs nur insofern von Belang sein wird, als er festlegt, welche Aussage eine gegebene AFK macht.

Eine Auseinandersetzung mit AFKs ist deshalb relevant, weil zwischen unserer alltäglich vollkommenen unproblematischen Verwendung von AFKs und einer ersten, oberflächlichen Analyse derselben eine eigentümliche Spannung zu bestehen scheint. So ist zum einen unsere Praxis der Verwendung von AFKs davon geprägt, dass wir AFKs – wie andere affirmative Äußerungen auch – als wahrheitswertdifferente Äußerungen gebrauchen. Wir sagen, einige AFKs seien wahr, während andere falsch seien, und setzen uns beizeiten darüber auseinander, ob in Bezug auf eine konkrete AFK der erste oder der zweite Fall vorliegt. Zum anderen erwecken AFKs jedoch prima facie den Anschein, als beziehe man sich mit ihnen auf fiktive Gegenstände und Ereignisse. Daher scheinen alle Partizipierenden an einer Praxis der Verwendung von AFKs wie der unseren, der zufolge AFKs wahr sein können, darauf festgelegt zu sein, die Existenz fiktiver Gegenstände anzuerkennen. Eine solche Anerkennung jedoch führt unmittelbar in große ontologische Probleme. Denn angenommen, fiktive Gegenstände existieren: Existieren sie dann auf die gleiche Art und Weise wie nicht-fiktive Gegenstände? Oder gibt es einen eigenen Seins-Modus des Als-fiktiver-Gegenstand-Existierens? Und falls Letzteres der Fall ist: Haben wir uns dann nicht einen Begriff der Existenz eingehandelt, der ganz und gar unverständlich ist?

Mein Vorhaben in dieser Arbeit ist es, eine Theorie der Semantik von AFKs zu entwickeln, die diese Spannungen auf nicht-reformistische und ontologisch sparsame Art und Weise auflöst. Dazu werde ich zunächst (Abschnitt 2) die bedeutungstheoretischen Voraussetzungen dieser Arbeit offenlegen und aus ihnen zwei Forderungen an eine akzeptable Theorie der Semantik von AFKs ableiten. Als Nächstes (Abschnitt 3) werde ich zwei klassische Theorien vorstellen und dafür argumentieren, dass keine von ihnen beiden Forderungen gerecht wird. Abschließend (Abschnitt 4) werde ich meinen eigenen Vorschlag entwickeln und geltend machen, dass dieser beiden Forderungen Rechnung trägt und somit den zuvor diskutierten Ansätzen überlegen ist.

## 2. Bedeutungstheoretische Voraussetzungen und zwei Forderungen

### *2.1 Bedeutungstheoretische Voraussetzungen*

Bevor ich den eigentlichen Gegenstand der Arbeit in den Blick nehmen kann, gilt es, die Frage zu beantworten, wie eine Theorie der Semantik von AFKs überhaupt der Form nach aussehen müsste. Dies ist eine allgemeine sprachphilosophische Frage, denn sie fragt letztlich danach, was es überhaupt heißt, eine Bedeutungstheorie für irgendeine Klasse von Äußerungen zu liefern. Folglich verlangt sie nach einer durch allgemeine Sprachphilosophie informierten Antwort – nach einer Antwort also, die von einer These in Bezug darauf ausgeht, worin die Bedeutsamkeit sprachlicher Ausdrücke überhaupt besteht. Eine solche These zu entwickeln und zu verteidigen, geht über diese Arbeit hinaus. Es bleibt mir also nichts anderes übrig, als von derjenigen Bedeutungstheorie auszugehen, die ich für angemessen halte. Hierbei handelt es sich um eine inferentialistische Semantik à la Robert Brandom, welche im folgenden Absatz ganz kurz vorgestellt werden soll.

Die Grundthese inferentialistischer Bedeutungstheorien jedweder Couleur besagt, dass ein sprachlicher Ausdruck seine Bedeutung durch diejenigen inferentiellen Beziehungen gewinnt, in denen er zu anderen sprachlichen Ausdrücken steht. Innerhalb des Inferentialismus lassen sich weiterhin zwei Strömungen unterscheiden – eine formalistische und eine anti-formalistische. (Bertram u.a. 2008: 77–80) Charakteristisch für einen formalistischen Inferentialismus ist, dass die Konstitution der bedeutungskonstitutiven, inferentiellen

Beziehungen ihrerseits rein innersprachlich gefasst wird.[2] Demgegenüber machen Vertreter eines anti-formalistischen Inferentialismus geltend, dass hierfür wesentlich auch außersprachliche Praktiken in den Blick genommen werden muss. Bei diesen Praktiken handelt es sich im Ansatz von Robert Brandom um soziale Praktiken des Begründens und Rechtfertigens.[3] (Brandom 1994) Damit ist gemeint, dass sich die Bedeutung einer Aussage *p* unter anderem dadurch konstituiert, dass die Mitglieder einer Sprachgemeinschaft einer Sprecherin den Schluss von *p* auf *q* durchgehen lassen, oder sie als darauf verpflichtet ansehen, auch *r* zu vertreten, wenn sie *p* vertritt.

Nimmt man diese bedeutungstheoretischen Überzeugung zum Ausgangspunkt, dann wird einsichtig, dass eine Theorie der Semantik einer bestimmten Klasse von Aussagen *K* der Form nach folgendermaßen muss: Sie muss verständlich machen, warum *K*-Aussagen in denjenigen inferentiellen Beziehungen stehen, die in der sozialen Praxis des Begründens und Rechtfertigens faktisch etabliert sind. Oder anders ausgedrückt: Sie muss in Begriffen in ihrem Funktionieren bereits besser verstandener *J*-Aussagen das inferentielle Netz, in welches *K*-Aussagen eingebettet sind, explizieren und offenlegen.

### *2.2 Zwei Forderungen*

Aus der soeben vorgenommen allgemeinen Charakterisierung der Form einer Theorie der Semantik von *K*-Aussagen ergibt sich direkt, dass eine angemessene Theorie der Semantik von AFKs folgender Forderung gerecht werden muss:

**IF – Inferentialistische Forderung**

Eine angemessen Theorie der Semantik von AFKs muss den spezifischen, praktisch konstituierten inferentiellen Beziehungen, in denen durch AFKs gemachte Aussagen stehen, erstens gerecht werden und sie zweitens verständlich und explizit machen.

Der Gedanke hinter IF ist der, dass erstens einer Theorie, die behauptet, von AFKs gemachte Aussagen stünden in Wahrheit in anderen inferentiellen Beziehungen als den praktisch etablierten, vor dem Hintergrund einer alistischen Semantik attestiert werden muss, dass sie ihr Thema verfehlt. Denn es sind ja gerade diese praktisch etablierten, inferentiellen Beziehungen, die laut einer alistischen Semantik die Bedeutung ebenjener Aussagen – und damit den intendierten Gegenstand der Theorie – ausmachen. Zweitens muss eine angemessene Theorie die von AFKs gemachten Aussagen in eine Form überführen, in der man ihnen ihre inferentielle Rolle gewissermaßen direkt ansieht. Denn nur von einer Theorie, der dies gelingt, kann man sagen, dass sie der Form nach eine Theorie ist, d.h. dass sie es schafft, die Bedeutung von AFKs offenzulegen.

Nach dem bisher Gesagten stellt sich die Frage, was es über IF hinaus noch zu fordern gibt. Denn wird eine Theorie IF gerecht, so erfüllt sie ja bereits die oben entwickelten Formkriterien einer Theorie der Semantik von AFKs gemachter Aussagen. Die nun folgende ontologische Forderung ist daher nicht als eigenständiges Desiderat, sondern als Folgerung aus IF zu verstehen:

**OF – Ontologische Forderung**

Eine angemessene Theorie der Semantik von AFKs darf nur auf Entitäten zurückgreifen, die sowohl vollständig als auch individuierbar sind.

---

[2] Als ein Vertreter des formalistischen Inferentialismus ist vor allem Wilfrid Sellars zu nennen. Vgl. insbesondere (Sellers 1954) und (Sellers 1997: 64–68).

[3] Der zweite prominente Vertreter des anti-formalistischen Inferentialismus ist Donald Davidson, welcher bei den für Bedeutung konstitutiven außersprachlichen Praktiken in erster Linie an intersubjektive Praktiken der Verständigung und wechselseitigen Interpretation in einer geteilten Welt denkt. (Davidson 1984, 2001)

An dieser Stelle stellen sich freilich sofort die Fragen, was es überhaupt heißt, dass ein Gegenstand unvollständig oder nicht-individuierbar ist, und warum solche Gegenstände vor dem Hintergrund eines semantischen Inferentialismus problematisch sind. Um sie beantworten zu können, muss ich auf zwei Begriffe zurückgreifen, die bislang noch nicht eingeführt wurden: die Begriffe der Determinablen und der Determinaten.[4]

Die Begriffe „Determinable“ und „Determinate“ gehen zurück auf den Logiker W. E. Johnson, welcher sie folgendermaßen einführt: „I propose to call such terms as colour and shape determinables in relation to such terms as red and circular which will be called determinates.“ (Johnson 1921, 171) Ohne auf umstrittene Detailfragen einzugehen, kann ich die Grundidee der von Johnson eingeführten Unterscheidung so erläutern:[5] Ein Prädikat $F$ ist eine Determinable, wenn es Prädikate $F_i$ gibt, welche spezifizieren, inwiefern ein $F$-Gegenstand $F$ ist. Ist dies der Fall, so heißen die $F_i$ „Determinaten von $F$“. Das Prädikat „ist farbig“ ist somit eine Determinable, weil es eine Reihe anderer Prädikate gibt – „ist rot“, „ist blau“ etc. – welche spezifizieren, inwiefern ein farbiger Gegenstand farbig ist. Diese Farbprädikate wiederum sind Determinaten von „ist farbig“. Die im Kontext dieser Arbeit entscheidende Eigenschaft von Determinablen besteht nun darin, dass es keinen Gegenstand gibt, welcher unter eine Determinable fällt, ohne zugleich auch unter eine ihrer Determinaten zu fallen. (Funkhouser 2006: 549) So sind farbige Gegenstände immer auf eine bestimmte Art und Weise farbig – sie sind rot, blau oder gelb. Es gibt keine schlichtweg farbigen Gegenstände.

Um die Begriffe der Determinablen und der Determinaten für eine Ausbuchstabierung und Begründung von OF auf eine Art und Weise fruchtbar machen zu können, die mit dem inferentialistischen Setting der Arbeit im Einklang steht, muss ich explizit anerkennen, dass es Determinablen im eben eingeführten Sinne innerhalb unserer Sprachpraxis tatsächlich gibt. Oder besser: Ich muss die Prämisse unterschreiben, dass es innerhalb unserer Begründungspraxis Prädikate $F$ gibt, für die erstens gilt, dass auf die Behauptung „$x$ ist $F$“ hin die Frage „Inwiefern ist $x$ $F$?“ stets legitim ist, und die zweitens eine Menge anderer Prädikate $F_i$ mit sich bringen, welche zusammengenommen alle Antwortmöglichkeiten auf diese Inwiefern-Frage bereitstellen. Diese Prämisse werde ich – wie die Bezeichnung schon nahelegt – nicht ausführlich begründen. Stattdessen muss der Hinweis genügen, dass so alltägliche Prädikate wie „ist farbig“, „ist ein Vieleck“ aber auch „ist behaart“ über die erforderlichen Eigenschaften verfügen.

Es ist mir nun möglich, zu explizieren, was ich unter Vollständigkeit und Individuierbarkeit verstehe:

**Definition – „Vollständigkeit“**

Ein Gegenstand $x$ ist genau dann vollständig, wenn für jede Determinable $F$, unter die $x$ fällt, gilt, dass für alle Determinaten $F_i$ von $F$ die Aussage „$x$ ist $F_i$“ entweder wahr oder falsch ist.

**Definition – „Individuierbarkeit“**

Ein Gegenstand $x$ ist genau dann individuierbar, wenn für jede Determinable $F$, unter die $x$ fällt, die Frage sinnvoll ist, unter welche der Determinaten $F_i$ von $F$ $x$ fällt.

---

[4] Ich verstehe die folgenden Erläuterungen als Ausbuchstabierung der Quine’schen Formel „There is no entity without identity.“ (Quine 1981: 102) Vgl. auch seine Argumentation gegen die Existenz möglicher Gegenstände. (Quine 1963: 4)

[5] Bei diesen Detailfragen handelt es sich beispielsweise um die folgenden: Wie ist das Verhältnis der Unterscheidung zwischen Determinable und Determinaten zu der zwischen Genus und Spezies? Sind Determinaten notwendigerweise (nicht-)disjunkt? Ist die Relation „ist Determinate von“ transitiv, d.h. ist „ist hellrot“ Determinate von „ist farbig“? Vgl. (Sanford 2011) für eine ausführliche Diskussion dieser und verwandter Fragestellungen.

Da der Begriff der Determinablen gerade so eingeführt wurde, dass es keine nicht-vollständigen oder nicht-individuierbaren Gegenstände gibt, folgt OF direkt aus der unproblematischen Prämisse, dass eine akzeptable Theorie irgendeiner Klasse von Aussagen nicht auf Gegenstände zurückgreifen sollte, die es aus begrifflichen Gründen nicht geben kann. Gegeben die Prämisse, dass es tatsächlich Prädikate gibt, welche die inferentielle Rolle von Determinablen ausfüllen, ist OF außerdem nicht-trivial.

## 3. Alternative Theorien der Semantik von Äußerungen über fiktive Kontexte

In diesem Abschnitt möchte ich die beiden aus meiner Sicht wichtigsten Theorien der Semantik von AFKs vorstellen und dafür argumentieren, dass keine sowohl IF als auch OF gerecht wird. Dabei beschränke ich mich auf knappe Darstellungen der grundlegendsten Aspekte der verschiedenen Positionen und verzichte im Wesentlichen auf eine Erläuterung der dahinter stehenden Motivationen. Dennoch bin ich der Ansicht, in diesen kurzen Zusammenfassungen genug Material präsentieren zu können, um verständlich zu machen, warum die diskutierten Ansätze im Lichte der Forderungen nicht haltbar sind.

### *3.1 Bertrand Russell*

Die erste Auseinandersetzung mit von AFKs gemachten Aussagen findet sich in Bertrand Russells Theorie der Funktionsweise von Eigennamen – der sogenannten „Theorie der Kennzeichnungen“. (Russell 1905) Aus den gerade genannten Gründen werde ich auf diese jedoch nicht eingehen, um mich stattdessen direkt seiner Auseinandersetzung mit AFKs zuzuwenden.[6]

Zuvor führe ich noch etwas Terminologie ein, welche ich im Rest der Arbeit verwenden werde. Sei *a* die von Sprecherin *s* geäußerte AFK, dass *p*. Dann erscheint es zunächst naheliegend, dass die von *a* gemachte Aussage, d.h. das, was *s* mit *a* sagt und worauf sie sich festlegt, einfach *p* ist. Dementsprechend nenne ich *p* die „oberflächliche Erscheinung der von *a* gemachten Aussage“. Ferner nenne ich eine Theorie der Semantik von AFKs genau dann „naiv“, wenn sie sich die Position zu eigen macht, die oberflächliche Erscheinung einer AFK sei ihr lokutionärer Bestandteil.

Naiven Theorien der Semantik von AFKs zufolge legen wir uns, indem wir AFKs verwenden, somit auf Aussagen fest, die ganz analog zu der folgenden sind:

(1) Asterix ist ein Gallier.

Russells Theorie der Semantik von AFKs lässt sich nun als eine Kritik naiver Theorien der Semantik von AFKs verstehen. Denn nehmen wir einmal an, diese seien korrekt. Dann folgt aus der Tatsache, dass einige AFKs – wie beispielsweise meine Äußerung, Asterix sei ein Gallier – wahr sind, dass Aussagen wie (1) wahr sind. Genau das aber bestreitet Russell, und zwar aus zwei Gründen. Erstens fasst er den Begriff der Existenz rein raum-zeitlich und ist daher der Ansicht, dass fiktive Gegenstände wie Asterix nicht existieren. Zweitens folgt aus Russells Theorie der Kennzeichnungen, dass Aussagen wie (1) anzuerkennen darauf verpflichtet, anzuerkennen, dass fiktive Gegenstände wie Asterix existieren. Somit kommt er dazu, einen naiven Ansatz zu verwerfen. Stattdessen schlägt er für die von *a* gemachte Aussage die folgende Analyse vor:

[6] In meiner Interpretation von Russells Theorie der Semantik von AFKs folge ich weitestgehend der Darstellung von (Rorty 1983).

**Russells Theorie der Semantik von AFKs**

Mit *a* legt sich *s* darauf fest, dass es eine bestimmte Menge fiktionaler Medien *W* gibt, die entweder *p* enthält oder andere Aussagen, aus denen *p* folgt.

Aufgrund des eng gefassten Russell'schen Existenzbegriffs verstehe ich seine Theorie so, dass er unter fiktionalen Medien Tokens und nicht Types versteht, den Ausdruck also so fasst, dass mein Exemplar von Büchners „Lenz" ein anderes fiktionales Medium ist als Deines. Somit wird Russells Analyse offensichtlich OF gerecht, da nur auf vollkommen unproblematische Gegenstände rekurriert wird – nämlich auf Dinge wie beispielsweise konkrete mündliche Erzählungen und einzelne Exemplare bestimmter Bücher. Außerdem trägt sie – wenngleich, wie in der Folge klar werden wird, auf unglückliche Art und Weise – dem Umstand Rechnung, dass zwischen den von AFKs gemachten Aussagen und Aussagen über fiktionale Medien in der Tat enge inferentielle Beziehungen bestehen.

Dennoch ist Russells Theorie unbefriedigend. Der wohl berühmteste Einwand geht von der Beobachtung aus, dass wir ihr zufolge mit einer AFK behaupten, dass bestimmte fiktionale Medien existieren. Tun sie dies nicht, wird die AFK nach Russells Analyse falsch. Diese Beschreibung, so geht der Einwand weiter, ist jedoch unangemessen. Denn in Wahrheit behauptet eine AFK gerade nicht, dass bestimmte fiktionale Medien existieren. Sie setzt dies vielmehr voraus, und zwar insofern, als sie misslingt und unverständlich wird, falls deren Existenz nicht gegeben ist. Oder anders ausgedrückt: Russells Analyse wird IF nicht gerecht, da ihr zufolge AFKs in anderen inferentiellen Beziehungen stehen, als sie das faktisch tun. (Searle 1979: 160)

Russells Theorie verfehlt IF aber noch auf andere, interessantere Art und Weise. Man betrachte dazu die folgende kleine Geschichte:

**Die Geschichte vom Hans**

Hans starrte aus dem Fenster. Draußen schien die Sonne, aber das tat sie schon seit Tagen. Er ging zum Kühlschrank, öffnete ihn, ließ seinen Blick dreißig Sekunden lang sinnlos zwischen Joghurtbechern und Milchflaschen hin und her schweifen und machte den Kühlschrank wieder zu. Er starrte wieder aus dem Fenster. Die Sonne schien immer noch.

Nun ist folgende AFK klarerweise wahr:

(2) Lisa: „Dem Hans war langweilig."

Gemäß Russells Theorie ist die von Lisa durch (2) gemachte Aussage aber die folgende:

(3) Es gibt die Geschichte vom Hans, die entweder die Aussage „Dem Hans war langweilig" enthält oder andere Aussagen, aus denen „Dem Hans war langweilig" folgt.

(3) ist jedoch falsch, denn „Dem Hans war langweilig" folgt aus keinen der in der Geschichte vom Hans enthaltenen Aussagen. Russells Theorie charakterisiert somit einige AFKs als falsch, obwohl sie tatsächlich wahr sind. Andererseits können nur Menschen ihren Blick sinnlos zwischen Joghurtbechern und Milchflaschen hin und her schweifen lassen, und man ist somit gemäß Russells Analyse darauf verpflichtet, eine AFK als wahr anzuerkennen, die behauptet, Hans sei ein Mensch. Eine solche AFK ist jedoch nicht wahr, denn die kleine Geschichte sagt in Bezug auf Hans' Mensch- oder Nicht-Mensch-Sein schlicht überhaupt nichts. Er könnte ebenso ein Insekt oder ein Alien sein – wir befinden uns schließlich in einer fiktiven Welt.

Russells Theorie scheitert also insofern, als sie IF dadurch auf eklatante Art und Weise verletzt, dass sie einigen AFKs Wahrheitswerte zuweist, die nicht denjenigen entsprechen, die sie tatsächlich haben. Der Grund für dieses Scheitern ist nun über die Diskussion des

Russell'schen Ansatzes hinaus von Interesse. Er besteht darin, dass fiktionale Medien nicht an gültige Inferenzen gebunden sind. Im Gegenteil scheint es mir gerade ein charakteristisches Merkmal fiktionaler Medien zu sein, dass sie diese teilweise aufkündigen und somit unsere nicht-fiktive Welt verlassen: Selbst der Satz vom ausgeschlossenen Widerspruch ist im Kontext fiktiver Welten nicht heilig. (Everett 2005: 633–4) Bei der Formulierung einer zufriedenstellenden Theorie der Semantik von AFKs wird also darauf zu achten sein, dieser Opazität fiktiver Kontexte gerecht zu werden.

### *3.2 John Searle, Peter van Inwagen, Terence Parsons*

Im letzten Paragraphen habe ich dafür argumentiert, dass Russells Analyse deshalb nicht haltbar ist, weil sie AFKs eine andere inferentielle Rolle zuweist, als ihnen tatsächlich zukommt. Eine mögliche Strategie, diesen Befund zu erklären, besteht darin zu sagen, Russell habe sich durch die Verabschiedung einer naiven Theorie der Semantik von AFKs zu weit vom gesunden Menschenverstand entfernt. Statt wie er zu versuchen, umständliche Paraphrasen für die von AFKs gemachten Aussagen zu finden, gelte es vielmehr, an einer naiven Theorie der Semantik von AFKs festzuhalten und sie weiter auszubuchstabieren.

In diesem Abschnitt wird es mir darum gehen, einige Positionen in den Blick zu nehmen, denen genau diese These gemein ist. Die Bezeichnungen des vorherigen Abschnitts übernehmend, kann ich die Kernthese dieser Positionen dann so zusammenfassen:

**Naive Theorie der Semantik von AFKs**

Mit *a* legt sich *s* darauf fest, dass *p*.

Naiven Theorien der Semantik von AFKs zufolge sagen AFKs einfach das, was sie zu sagen scheinen. Eine weitergehende Analyse von AFKs ist laut ihnen nicht notwendig. Die grundlegende Herausforderung und Aufgabe, die sich naiven Theorien der Semantik von AFKs stellt, besteht daher gerade nicht darin, eine Paraphrase für die von AFKs gemachten Aussagen zu liefern, sondern vielmehr darin, die folgenden Fragen zu beantworten: Was folgt daraus, dass wir uns mit AFKs auf Aussagen wie (1) festlegen? Wie können wir verständlich machen, dass einige AFKs wahr sind? Wovon handeln Aussagen wie (1)?

Ein einflussreicher Versuch, Antworten auf diese Fragen zu liefern, ist derjenige John Searles. Sein argumentativer Ausgangspunkt ist dabei die Auseinandersetzung damit, ob die folgende Aussage wahr oder falsch ist oder vielleicht gar keinen Wahrheitswert hat:

> There never existed a Mrs. Sherlock Holmes because Holmes never got married, but there did exist a Mrs. Watson because Watson did get married, though Mrs. Watson died not long after their marriage. (Searle 1975: 329)

Seine Antwort lautet wie folgt:

> But taken as a piece of discourse about fiction, the above statement is true because it accurately reports the marital histories of the two fictional characters Holmes and Watson. [...] Holmes and Watson never existed at all, which is not of course to deny that they exist in fiction and can be talked about as such. [...] Because the author has created these fictional characters, we on the other hand can make true statements about them as fictional characters. (Searle 1975: 329)

Searle gibt auf die Fragen, wovon AFKs handeln und wie wir es verständlich machen können, dass einige AFKs wahr sind, also die folgenden Antworten:

(i) AFKs handeln von fiktiven Gegenständen.

(ii) Fiktive Gegenstände können Gegenstand wahrer Aussagen sein.

(iii) AFKs sind insofern wahr, als sie die Beziehungen, in denen fiktive Gegenstände zueinander stehen, korrekt wiedergeben.

(iv) Fiktive Gegenstände werden von den Autoren fiktionaler Werke erschaffen.

(v) Fiktive Gegenstände existieren nicht wirklich, sondern nur als-fiktive-Gegenstände.

Ein naheliegender Einwand besteht nun darin zu sagen, dass diese Auskünfte unvollständig sind. Denn es bleibt ja zunächst ganz unklar, wie es zu verstehen ist, dass fiktive Gegenstände zwar nicht existieren, aber sehr wohl als-fiktive-Gegenstände-existieren und von Autoren fiktionaler Werke erschaffen werden. Ich verstehe die im Rest dieses Abschnitts zu diskutierenden Positionen als Versuche, diese Unklarheiten zu beseitigen, ohne dabei die Grundidee des Searle'schen Ansatzes mit über Bord zu werfen. Konkreter heißt das, dass (i)-(iii) beibehalten, (iv) und (v) hingegen verworfen oder modifiziert werden.

Peter van Inwagens Versuch einer Präzisierung der Überlegungen Searles besteht darin, fiktive Gegenstände als theoretische Gegenstände zu fassen, auf deren Existenz wir uns durch unsere Verwendung von AFKs festlegen. (Van Inwagen 1983) Das Argument, welches er zur Stützung dieser These vorbringt, geht von der Voraussetzung aus, dass naive Theorien der Semantik von AFKs angemessen sind. Genau wie Russell in seiner Kritik solcher Theorien sieht nun auch van Inwagen, dass diese zusammen mit der Prämisse, dass einige AFKs wahr sind, darauf verpflichten anzuerkennen, dass fiktive Gegenstände existieren. Im Gegensatz zu Russell hat van Inwagen jedoch keinerlei Skrupel, diese Verpflichtung einzugehen. Der Grund hierfür ist, dass er anders als Russell keinen engen, raum-zeitlichen Existenzbegriff vertritt, sondern im Gegenteil denselben im Sinne des in (Quine 1963) entworfenen ontologischen Relativismus versteht.

Grob zusammengefasst, besagt Quines ontologischer Relativismus, dass die Frage, welche Gegenstände existieren, nicht durch „objektive" Untersuchungen der raum-zeitlichen Welt beantwortet werden und der Begriff der Existenz somit nicht rein raum-zeitlich verstanden werden kann. Stattdessen ist die Frage „Was gibt es?" Quine zufolge gleichbedeutend mit der folgenden: Welche Aussagen und Theorien erkennen wir an und welche ontologische Verpflichtungen handeln wir uns dadurch ein? Van Inwagen interpretiert Quine nun so, dass damit alle spezifisch ontologischen Erwägungen durch eine allgemeine Meta-Ontologie ersetzt wurden. Gemäß seiner Lesart besagt Quines ontologischer Relativismus, dass wir immer zuerst und ohne Ansehung ontologischer Argumente entscheiden, welche Aussagen und Theorien wir anerkennen. Danach klären wir in einem zweiten Schritt, auf die Existenz welcher Gegenstände wir uns dadurch verpflichtet haben. Genau diese Gegenstände existieren, und ihr ontologischer Status ist der eines theoretischen Gegenstandes der jeweiligen Theorie oder Aussage. Indem van Inwagen diese Meta-Ontologie auf AFKs anwendet, kann er dann einfach sagen, fiktive Gegenstände seien „theoretical entit[ies] of literary criticism" (Van Inwagen 1983: 75) auf genau dieselbe Art und Weise, wie die leere Menge ein theoretischer Gegenstand der Zermelo-Fraenkel-Mengenlehre ist und Häuser und Steine theoretische Gegenstände unseres alltäglichen Redens über Häuser und Steine sind. Damit ist es ihm scheinbar gelungen, an den Searle'schen Kerneinsichten (i)–(iii) festzuhalten, ohne dabei auch die unterbestimmten Thesen (iv) und (v) zu unterschreiben.

Doch die Unklarheiten bleiben an Bord, obgleich vielleicht auf verstecktere und weniger offensichtliche Art und Weise als bei Searle. Denn anders als Mengen, Häuser und Steine haben fiktive Gegenstände im Sinne van Inwagens eine unangenehme Eigenschaft: Theorien, die ihre Existenz anerkennen, werden OF nicht gerecht.[7] Um zu sehen warum, betrachte man zunächst van Inwagens Antizipation dieses Einwands:

[7] Dieser Befund, den ich gleich begründen werde, ist vielleicht insofern überraschend, als ich mich in meiner Argumentation für OF genau wie van Inwagen auf (Quine 1963) bezogen habe. Er erklärt sich jedoch, wenn man sich klar macht, dass van Inwagen Quine insofern missdeutet, als Quine eben gerade nicht vorschlägt, Ontologie komplett durch Meta-Ontologie zu ersetzen. Bestimmte spezisch ontologische Erwägungen bleiben weiterhin relevant, wenn es darum geht zu entscheiden, welche Gegenstände existieren und welche Theorien wir anerkennen sollen.

> Consider the famous question, How many children had Lady Macbeth? One traditional line of thought runs as follows: *Any* definite answer to this question would be wrong. ('None' would be wrong, 'One' would be wrong, 'Two' would be wrong, and so on.) But, according to the rules of logic that we apply to ordinary, nonfictional beings, some definite answer would have to be right. (Van Inwagen 1983: 75)

Mit der Begrifflichkeit, die ich in Abschnitt 2 entwickelt habe, lässt sich dieser Einwand auf zwei verschiedene Weise reformulieren. In der ersten Form besagt er, dass Lady Macbeth ein unvollständiger Gegenstand ist. Das Argument geht dann so: Shakespeares „Macbeth" zufolge fällt Lady Macbeth unter das Prädikat „ist eine Frau". Da „ist eine Frau" eine Determinable hinsichtlich der Determinaten „hat 0 Kinder", „hat 1 Kind", „hat 2 Kinder" usw. ist, ist Lady Macbeth nur dann vollständig, wenn für alle natürlichen Zahlen *n* „Lady Macbeth hat *n* Kinder" entweder wahr oder falsch ist. Man nehme also an, dass Lady Macbeth vollständig ist. Dann ist der Satz „Lady Macbeth hat 0 Kinder" wahr oder falsch. Da er offensichtlich nicht wahr ist, ist er somit falsch und es existiert daher eine natürliche Zahl *n* größer 0, sodass „Lady Macbeth hat *n* Kinder" wahr ist. Dies ist jedoch nicht der Fall und es folgt somit ein Widerspruch. Also war die Annahme falsch und Lady Macbeth ist kein vollständiger Gegenstand.[8]

Nun könnte man freilich erwidern, dieses Argument ließe sich blockieren, indem man einfach festlegt, dass es eine natürliche Zahl *n* gibt, sodass Lady Macbeth *n* Kinder hat. Da Shakespeares Geschichte – so die Erwiderung weiter – keinerlei Auskunft über die Anzahl ihrer Nachkömmlinge gebe, sei diese Festlegung vollkommen harmlos. Doch auch diese Erwiderung hilft nicht weiter, denn obschon sie Lady Macbeth vor der Unvollständigkeit bewahrt, gelingt es ihr doch nicht, sie auch zu einem individuierbaren Gegenstand zu machen. Zwar gibt es nun per Definition eine natürliche Zahl *n*, sodass „Lady Macbeth hat *n* Kinder" wahr ist. Aber noch immer gilt, dass für jede natürliche Zahl *n* die Frage, ob Lady Macbeth *n* Kinder hat, insofern sinnlos ist, als die angemessene Antwort stets lautet: „Das ist doch total egal! Wenn es Dich glücklich macht zu sagen, sie habe *n* Kinder, dann darfst Du das gerne tun. Aber ebenso gut kannst Du sagen, dass sie *m* Kinder hat."

Ich möchte nun noch ganz kurz die Positionen Terence Parsons' vorstellen, die ich so interpretiere, dass sie explizit einräumt, dass fiktive Gegenstände unvollständig sind. Parsons zufolge haben fiktive Gegenstände genau diejenigen Eigenschaften, die ihnen gemäß der fiktionalen Medien zukommen, in denen sie erwähnt werden. Fiktive Gegenstände seien somit in der Regel unvollständig, „for the body of literature in question will not determine all of their properties". (Parsons 1974: 74) Zum Beispiel habe Sherlock Holmes genau diejenigen Eigenschaften, die ihm laut Conan Doyles Geschichten zukommen. Da es diesen Geschichten zufolge aber weder wahr noch falsch ist, dass Holmes ein Muttermal an seinem linken Bein hat, sei Holmes in Bezug auf die Eigenschaft unvollständig, ein Muttermal am linken Bein zu haben. Diese Auskunft Parsons' ist freilich das offene Eingeständnis, dass seine Theorie OF nicht gerecht wird. Sherlock Holmes als einen Gegenstand zu betrachten, der zwar unter die Determinable „hat ein linkes Bein" fällt, von dem es aber weder wahr noch falsch ist, dass er unter die zugehörige Determinate „hat ein Muttermal am linken Bein" fällt, ist schlicht unverständlich.

Ich kann die Ergebnisse der letzten Absätze zusammenfassen, indem ich sage, dass die vorgestellten naiven Theorien der Semantik von AFKs insofern hinter Russells Ansatz zurückbleiben, als sie einen bewahrenswerten Aspekt seiner Theorie aufgeben – nämlich den, nur auf Gegenstände zurückzugreifen, die vollständig und individuierbar sind. Oder anders

[8] Van Inwagen versucht dieses Problem zu lösen, indem er behauptet, dass erstens fiktive Gegenstände Eigenschaften nicht haben sondern halten und zweitens „[no] principle of logic says anything about what properties an object must *hold*". (Van Inwagen 1983: 76) Ich betrachte diese Auskunft jedoch eher als exzentrische Reformulierung denn als Lösung des Problems und werde daher nicht weiter auf sie eingehen.

formuliert: Ihnen gelingt es nicht, auf verständliche Art und Weise auszubuchstabieren, was ein fiktiver Gegenstand überhaupt ist. Die den Thesen (iv) und (v) innewohnende Unbestimmtheit bleibt in verschiedener Form immer an Bord.

In den verbleibenden Absätzen dieses Abschnitts möchte ich naive Theorien der Semantik von AFKs nun auf noch fundamentalere Art und Weise kritisieren. Das Argument, dass ich vorbringen werde, zielt darauf ab zu zeigen, dass solche Theorien auch IF nicht gerecht werden. Der Grund hierfür ist, dass es ihnen nicht gelingt, diejenigen inferentiellen Beziehungen explizit und verständlich zu machen, die zwischen den von AFKs gemachten Aussagen und Aussagen über fiktionale Medien bestehen. Man betrachte dazu zunächst die folgenden beiden Äußerungen, die Lisa äußert, nachdem sie durch ihres Vaters Ausgabe von „Asterix bei den Briten“ zum ersten Mal mit Asterix-Comics in Berührung kam:

(4) Lisa: „Asterix ist ein Gallier.“

(5) Lisa: „Laut Peters Ausgabe von ‚Asterix bei den Briten‘ ist Asterix ein Gallier.“

Naiven Theorien der Semantik von AFKs zufolge ist die von (4) gemachte Aussage (1). Die von (5) gemachte Aussage ist hingegen offensichtlich einfach eine Aussage, über Peters Asterix-Ausgabe – nämlich gerade der in Anführungszeichen stehende Satz:

(6) Laut Peters Ausgabe von „Asterix bei den Briten“ ist Asterix ein Gallier.

An dieser Stelle gilt es nun zu bemerken, dass wir Lisa so behandeln, dass sie sich durch (4) darauf verpflichtet, den Inhalt von (5) anzuerkennen und umgekehrt. Wir ließen es ihr nicht durchgehen, zu sagen, Asterix sei ein Gallier, laut Peters Ausgabe sei dies jedoch nicht der Fall. Ebenso wenig würden wir es ihr gestatten, die These zu vertreten, Peters Ausgabe sage zwar, dass Asterix ein Gallier ist, in Wahrheit lägen die Dinge jedoch anders. IF besagt nun, dass eine angemessene Analyse der Semantik von AFKs diese inferentiellen Beziehungen verständlich und explizit machen muss. Anders ausgedrückt heißt das, dass eine solche Analyse zeigen muss, warum die von (4) und (5) gemachten Aussagen auseinander folgen. Genau dies leisten naive Theorien der Semantik von AFKs jedoch nicht, da weder (1) aus (6) noch (6) aus (1) folgt. Während (1) eine Aussage über einen fiktiven Gegenstand ist, ist (6) eine Aussage über ein fiktionales Medium, und naive Theorien der Semantik von AFKs geben keinerlei Auskunft darüber, wie Aussagen der ersten mit Aussagen der zweiten Art systematisch zusammenhängen.

## 4. Äußerungen über fiktive Kontexte als Quasi-de-dicto-Zuschreibungen an Werke

Im abschließenden vierten Abschnitt möchte ich eine Theorie der Semantik von AFKs formulieren, welche die den genannten Theorien eigentümlichen Probleme vermeidet und sowohl IF als auch OF gerecht wird. In den ersten beiden Paragraphen werde ich die Theorie präsentieren. Im dritten Paragraph gilt es dann zu zeigen, dass sie diesen Ansprüchen gerecht wird.

### *4.1 Der Gegenstand von Äußerungen über fiktive Kontexte*

Die erste Frage, die es bei der Formulierung einer Theorie der Semantik von AFKs zu beantworten gilt, ist diese: Wovon handeln AFKs? Die Auseinandersetzung mit naiven Theorien der Semantik von AFKs hat gezeigt, dass der Begriff des fiktiven Gegenstands zu erheblichen Problemen führt. Auf ihn gilt es also zu verzichten. Es bleibt der Ansatz Russells. Wie oben vorgestellt, identifiziert dieser die Referenz von AFKs mit bestimmten Mengen fiktionaler Texte. Mir geht es in diesem Paragraphen darum, diesen grundsätzlich

vielversprechenden Ansatz auf eine Art und Weise auszubuchstabieren, die die für Russell charakteristischen Probleme vermeidet.

Zunächst einmal gilt es zu fragen, auf welcher Menge diejenigen Teilmengen definiert werden sollen, die schließlich als Referenz von AFKs fungieren. Ich wähle für diese grundlegende Menge einfach die Menge aller fiktionalen Medien – um einen Namen zu haben, nenne ich diese Menge *M*. Dabei ist wichtig, dass ich den Begriff der fiktionalen Medien so verwende, dass es sich bei diesen um Tokens und nicht um Types handelt. Nun ist es prima facie natürlich alles andere als klar, welche Medien als fiktional und welche als nicht-fiktional gelten. Aber diese Frage zu beantworten – und damit die Extension von *M* zu bestimmen – ist Aufgabe einer Theorie fiktionaler Medien und nicht Gegenstand dieser Arbeit. Ich setzte die Extension von *M* daher als gegeben voraus.

Nun führe man auf *M* die Relation *xRy* ein, sodass *xRy* genau dann, wenn *x* zum gleichen Werk gehört wie *y*. Wie schon bei der Bestimmung der Extension von *M*, gilt auch hier, dass eine Ausbuchstabierung dessen, was es heißt, dass zwei fiktionale Medien zum gleichen Werk gehören, weitestgehend Aufgabe einer Theorie fiktionaler Medien ist. Dennoch kann die Konstruktion von *R* im Gegensatz zur Bestimmung der Extension von *M* nicht vollständig der Theoretikerin fiktionaler Medien überlassen werden. Aus Gründen, die im letzten Paragraphen dieses Abschnitts klar werden, muss *R* vielmehr so bestimmt werden, dass sie der Forderung genügt, dass zwei fiktionale Medien *x* und *y* nur dann zum gleichen Werk gehören, wenn *x* und *y* das Gleiche sagen.

Wichtig ist, dass die Relation *R*, wie auch immer sie genau bestimmt wird, symmetrisch, reflexiv und transitiv ist und somit eine Äquivalenzrelation auf *M* darstellt. Folglich liefert *R* eine Partition auf *M*, d.h. sie teilt *M* disjunkt in Äquivalenzklassen auf, die aus fiktionalen Medien bestehen, die alle zum gleichen Werk gehören. Diese Äquivalenzklassen werde ich „Werke“ nennen. Sei nun *a* eine von *s* geäußerte AFK, *p* die oberflächlichen Erscheinung der von *a* gemachten Aussage, $x_0$ ein bestimmtes fiktionales Medium und $[x_0]_R$ die zu $x_0$ gehörige Äquivalenzklasse, d.i. dasjenige Werk, welches aus genau denjenigen fiktionalen Medien besteht, die zum gleichen Werk gehören wie $x_0$. Dann schlage ich Folgendes als eine erste Annäherung an eine Theorie der Semantik von AFKs vor:

**Vorläufige Theorie der Semantik von AFKs**

Mit *a* legt sich *s* darauf fest, dass für alle *x* aus $[x_0]_R$ gilt: gemäß *x p*.

Diese vorläufige Theorie ist insofern ein erster Schritt, als nun klar ist, wovon AFKs handeln: Sie handeln von genau denjenigen fiktionalen Medien, die Element eines bestimmten Werkes sind.[9] Oder anders ausgedrückt: Sie handeln von genau denjenigen fiktionalen Medien, die zum selben Werke gehören wie ein bestimmtes fiktionales Medium. Um umständliche Formulierungen zu vermeiden, werde ich mich in der Folge auf dieses Ergebnis beziehen, indem ich sagen werde: AFKs handeln von bestimmten Werken.

Um welches Werk es sich dabei handelt, hängt vom Kontext ab. Wenn ein Kind zum ersten Mal einen Asterix-und-Obelix-Comic in die Hände bekommt und dann verkündet: „Asterix ist ein Gallier“, dann redet das Kind von dem zu seinem Exemplar des Asterix-und-Obelix-Comics gehörigen Werk. Manchmal ist die Lage freilich etwas komplizierter. Wenn ich beispielsweise ganz aus dem Blauen heraus sage, dass Innstetten Crampas zum Duell fordert, dann ist zunächst wohl nicht ganz klar, ob ich von dem zu meinem Exemplar von Fontanes

[9] An dieser Stelle könnte es so scheinen, als sei es mir gelungen, komplett ohne einen auf der Type-Ebene angesiedelten Werkbegriff auszukommen. Dieser Eindruck täuscht jedoch, da das Verfügen über einen solchen Begriff insofern eine Bedingung der Möglichkeit der genauen Bestimmung von *R* ist, als eine Diskussion darüber, wann zwei fiktionale Medien zum gleichen Werk gehören, nicht geführt werden kann, wenn keine die einzelnen Medien transzendierende Kategorie zur Verfügung steht. Zwar überlasse ich in der vorliegenden Arbeit die genaue Bestimmung von *R* der Theoretikerin fiktionaler Medien, ihre Verpflichtung auf die Sinnhaftigkeit einer solchen Kategorie erbe ich jedoch.

„Effi Briest“ gehörigen Werk spreche oder von demjenigen Werk, welches von Fassbinders Film induziert wird, den ich irgendwann einmal gesehen habe. Aber diese Unklarheit lässt sich durch Nachfragen beseitigen.

Dennoch ist die vorläufige Theorie nur ein erster Schritt, denn noch ist ganz unklar, was der Ausdruck „gemäß *x p*“ überhaupt besagt. Noch ist die Frage unbeantwortet, was es heißt, dass eine Aussage gemäß eines fiktionalen Mediums wahr ist. Diese Frage werde ich im nächsten Paragraphen zu beantwortet versuchen, indem ich die These vertreten werde, dass AFKs nach dem Modell von De-dicto-Zuschreibungen zu verstehen sind.

### *4.2 Die logische Feinstruktur von AFKs*

Unter De-dicto-Zuschreibungen verstehe ich Äußerungen wie die folgende:

(7) Lisa: „Georg Ratzinger sagt, dass homosexuelle Beziehungen gegen den Willen Gottes verstoßen.“

Die charakteristische Eigenschaft von Äußerungen wie (7) besteht darin, dass innerhalb der Reichweite des Sagt-dass-Operators koreferenzielle Ausdrücke nicht ohne Weiteres gegeneinander ausgetauscht werden können, ohne dass sich dadurch der Wahrheitswert der Äußerungen ändern würde. Der Sagt-dass-Operator erzeugt einen opaken Kontext. Spätestens seit Gottlob Frege steht die Aufgabe, eine angemessene Analyse derartiger Äußerungen bzw. der zugehörigen Aussagen zu liefern, auf der philosophischen Agenda. (Frege 1962, 1966) Auf diese reichhaltige Diskussion hier auch nur ansatzweise einzugehen, würde den Rahmen des Aufsatzes sprengen. Stattdessen werde ich mich darauf beschränken zu erläutern, wie De-dicto-Zuschreibungen im Rahmen des in dieser Arbeit favorisierten semantischen Inferentialismus gefasst werden.

Nach dem in Abschnitt 2 Gesagten ist klar, was eine inferentialistische Theorie von De-dicto-Zuschreibungen leisten muss: Sie muss verständlich machen, in welchen praktisch konstituierten inferentiellen Beziehungen die von De-dicto-Zuschreibungen gemachten Aussagen stehen. Der Ansatz Robert Brandoms schlägt in diesem Sinne vor, De-dicto-Zuschreibungen als Äußerungen über den Überzeugungshaushalt des Interpretierten zu analysieren, die auf die folgende Art und Weise inferentiell eingebettet sind: (Brandom 1994: 8.1.2–3)

(vi) Wir verpflichten uns durch De-dicto-Zuschreibungen darauf, dass ein Interpretierter auf die-und-die Aussagen verpflichtet ist.

(vii) Wir erwerben durch eine De-dicto-Zuschreibung die Berechtigung zu weiteren De-dicto-Zuschreibungen derart, dass der Interpretierte die Inferenz von der ersten zur zweiten zugeschriebenen Aussage billigt.

(viii) Wir verpflichten uns durch eine De-dicto-Zuschreibung ontologisch auf die Existenz des Interpretierten, nicht aber auf die Existenz derjenigen Dinge, auf die uns eine eigenständige Äußerung der zugeschriebenen Aussagen verpflichten würde.

Diese Auskünfte kann ich am Beispiel von (7) verdeutlichen. Ontologisch verpflichtet sich Lisa einzig und allein darauf, dass Georg Ratzinger existiert. Die Sinnhaftigkeit von (7) leidet keineswegs darunter, dass sie Atheistin ist. Die Frage, auf welche Aussage sich Lisa verpflichtet, ist ebenso leicht beantwortbar: Sie verpflichtet sich darauf, dass sich Georg Ratzinger darauf verpflichtet, dass homosexuelle Beziehungen gegen den Willen Gottes verstoßen. Interessanter ist die Frage, zu welchen weiteren Aussagen Lisa durch ihre Behauptung berechtigt ist. Man betrachte dazu folgende Aussagen:

(8) Georg Ratzinger sagt, dass Klaus Wowereits Beziehungsform gegen den Willen Gottes verstößt.

(9) Georg Ratzinger sagt, dass es vollkommen legitime Beziehungsformen gibt, die gegen den Willen Gottes verstoßen.

Intuitiv ist klar, dass (7) Lisa dazu berechtigt, (8) zu vertreten, aber nicht dazu, (9) zu vertreten. Diese Intuition lässt sich durch die soeben gegebenen Auskünfte leicht einholen: Während Lisa dazu berechtigt ist, Ratzinger die Inferenz von „Homosexuelle Beziehungen verstoßen gegen den Willen Gottes“ auf „Klaus Wowereits Beziehungsform verstößt gegen den Willen Gottes“ zuzuschreiben, muss sie davon ausgehen, dass er den Schluss von „Homosexuelle Beziehungen verstoßen gegen den Willen Gottes“ auf „Es gibt vollkommen legitime Beziehungsformen, die gegen den Willen Gottes verstoßen“ ablehnt. Lisa darf annehmen, dass Ratzinger grundlegend über deutsche Landespolitik Bescheid weiß, aber sie darf nicht annehmen, dass er homosexuelle Beziehungen für legitim hält.

Wenn sich die inferentielle Rolle von De-dicto-Zuschreibungen durch (vi)–(viii) beschreiben lässt, stellt sich die Frage, ob sie dadurch bereits vollständig charakterisiert ist. Die Antwort auf diese Frage muss aus meiner Sicht „Nein!“ lauten.[10] Denn tatsächlich schreiben wir mit De-dicto-Zuschreibungen den Interpretierten nicht nur bestimmte Aussagen, sondern vielmehr auch eine eigenständige Perspektive auf diese Aussagen zu. Lisa ist durch die Äußerung von (7) nicht nur darauf verpflichtet, dass Ratzinger eine bestimmte Position hinsichtlich homosexueller Beziehungen vertritt, sondern vielmehr auch darauf, dass Ratzinger sich Bezug auf die von ihm vertretene Position eigenständig artikulieren kann.

Dieser Aspekt der inferentiellen Rolle von De-dicto-Zuschreibungen wird von (vi)–(viii) nicht erfasst. Bei Äußerungen, die durch (vi)–(viii) vollständig charakterisiert sind, handelt es sich somit nicht um De-dicto-Zuschreibungen im vollen Sinne. Um einen Ausdruck zu haben, werde ich solche Äußerungen daher „Quasi-de-dicto-Zuschreibungen“ nennen. Die Hauptthese meiner inferentialistischen Therie der Semantik von AFKs ist, dass es sich bei AFKs um Quasi-de-dicto-Zuschreibungen an ein bestimmtes Werk handelt. Oder anders ausgedrückt: Bei AFKs handelt es sich um Äußerungen, die von bestimmten Werken handeln und in ihrem Funktionieren von (vi)–(viii) vollständig beschrieben werden.

**Inferentialistische Theorie der Semantik von AFKs**

Mit *a* legt sich *s* darauf fest, dass für alle *x* aus $[x_0]_R$ gilt: *x* sagt, dass *p*.

Die inferentialistische Theorie vervollständig die vorläufige, indem der Platzhalter „gemäß *x p*“ durch die quasi-de-dicto zu lesende Wendung „*x* sagt, dass *p*“ ersetzt wird. Vor dem Hintergrund von (vi)–(viii) können nun die folgenden drei Fragen beantwortet und somit das Vorhaben eingelöst werden, die von *a* gemachte Aussage in eine Form zu überführen, in der ihre inferentielle Rolle und damit ihre Bedeutung offen zu Tage liegt:

(vi') *Frage:* Auf welche Aussage verpflichten wir uns durch *a*? *Antwort:* Darauf, dass die *x* aus $[x_0]_R$ auf *p* verpflichtet sind.

(vii') *Frage:* Welche weiteren Aussagen der Form „Für alle *x* aus $[x_0]_R$ gilt: *x* sagt, dass *q*“ dürfen wir auf Grundlage von *a* vertreten? *Antwort:* Genau diejenigen, für die gilt, dass wir dazu berechtigt sind, den *x* aus $[x_0]_R$ „*x* sagt, dass wenn *p*, dann *q*“ zuzuschreiben.

(viii') *Frage:* Auf die Existenz welcher Gegenstände verpflichten wir uns durch *a*? *Antwort:* Auf die Existenz von $x_0$, d.i. die Existenz eines bestimmten fiktionalen Mediums.

[10] Für den Hinweis, dass (vi)–(viii) bei Weitem nicht alles für De-dicto-Zuschreibungen Charakteristische erfassen, bin ich Georg W. Bertram zu Dank verpflichtet.

### *4.3 Begründung der Theorie im Lichte der Forderungen*

In diesem abschließenden Paragraphen gilt es nun, dafür zu argumentieren, dass meine inferentialistische Theorie der Semantik von AFKs den zuvor präsentierten Ansätzen insofern überlegen ist, als sie IF und OF gerecht wird. Der Nachweis, dass sie OF gerecht wird, ist schnell erbracht, da wir uns – wie ich im letzten Abschnitt herausgestellt habe – durch die Verwendung von AFKs ontologisch lediglich auf die Existenz eines bestimmten fiktionalen Mediums verpflichten. Fiktionale Medien sind jedoch einfach Tokens wie beispielsweise meine Ausgabe von Fontanes „Effi Briest“ und damit ontologisch genauso harmlos wie Tische und Stühle.

Der Nachweis, dass meine Theorie auch IF gerecht wird, ist deutlich schwieriger zu führen, und ich habe offen gestanden keine Ahnung, wie er in vollständiger Allgemeinheit aussehen könnte. Ich werde mich in meiner Argumentation daher darauf beschränken zu zeigen, dass diejenigen Argumente, die ich ausgehend von IF gegen die in Abschnitt 3 vorgestellten Positionen angeführt habe, bei meiner Theorie keinen Ansatzpunkt finden.

Der erste Einwand, den ich ausgehend von IF mit (Searle 1979) gegen Russells Theorie der Semantik von AFKs erhoben habe, ist der, dass AFKs voraussetzen und nicht behaupten, dass bestimmte fiktionale Medien existieren. Offensichtlich kann dieser Einwand nicht gegen meine Theorie erhoben werden, da der Ausdruck „$[x_0]_R$“ nur dann sinnvoll ist, wenn $x_0$ existiert. So legt sich Lisa mit (4) beispielsweise darauf fest, dass folgende Aussage wahr ist:

(1o) Für alle fiktionalen Medien *x*, die zum gleichen Werk gehören wie Lisas Vaters Ausgabe von „Asterix bei den Briten“ gilt: *x* sagt, dass Asterix ein Gallier ist.

Diese Aussage wird jedoch, genau wie von IF gefordert, sinnlos und nicht falsch, wenn Lisas Vaters (oder – je nach Kontext – irgendeine andere) Ausgabe von „Asterix bei den Briten“ nicht existiert.

Der zweite auf IF basierende Einwand gegen Russells Theorie nahm Überlegungen von (Everett 2006) auf und bestand darin, dass diese der Tatsache, dass fiktionale Medien nicht an gültige Inferenzen gebunden sind, nicht Rechnung trägt, und folglich einigen AFKs eklatant falsche Wahrheitswerte zuweist. Auch dieser Einwand kann gegen meine Theorie nicht erhoben werden. Der Grund hierfür ist, dass diese durch die Verwendung des opaken Sagt-dass-Operators dem Umstand gerecht wird, dass die Frage, ob eine AFK wahr oder falsch ist, über die Frage hinausgeht, ob eine Aussage aus einer Menge anderer Aussagen folgt. Sie zu beantworten verlangt vielmehr, ein fiktionales Medium zu interpretieren, und das ist ein subtileres und vielschichtigeres Unterfangen als das bloße Anwenden inferentieller Muster. Tatsächlich ist es nach meiner Theorie überhaupt nicht überraschend, dass es passieren kann, dass innerhalb fiktionaler Welten einige gültige Inferenzen scheitern. Dies liegt daran, dass wir auf Grundlage der AFK *a* immer nur zu denjenigen Aussagen der Form „Für alle *x* aus $[x_0]_R$ gilt: *x* sagt, dass *q*“ berechtigt sind, für die gilt, dass wir dazu berechtigt sind, den *x* aus $[x_0]_R$ „*x* sagt, dass wenn *p*, dann *q*“ zuzuschreiben.

Es bleibt zu zeigen, dass es der inferentialistischen Theorie gelingt, auch denjenigen Einwand zu blockieren, den ich ausgehend von IF gegen naive Theorien der Semantik von AFKs vorgebracht habe. Dieser Einwand bestand darin, dass naive Theorien der Semantik von AFKs diejenigen inferentiellen Beziehungen nicht auf zufriedenstellende Art und Weise explizieren, die zwischen den von AFKs gemachten Aussagen und Aussagen über fiktionale Medien bestehen. So gelingt es naiven Theorien beispielsweise nicht, verständlich zu machen, warum Lisa durch (4) auf den Inhalt von (5) verpflichtet ist und umgekehrt. Um zu sehen, wie meiner Theorie gerade dies gelingt, bemerke man, dass sich Lisa ihr zufolge mit (4) und (5) auf (10) bzw. die folgende Aussage (11) festlegt:

(11) Peters Ausgabe von „Asterix bei den Briten“ sagt, dass Asterix ein Gallier ist.

Da Peters und Lisas Vaters Ausgaben von „Asterix bei den Briten“ zum gleichen Werk gehören, folgt (11) aus (10).[11] Da *R* jedoch als Äquivalenzrelation gewählt ist, von der gilt, dass *xRy* nur dann, wenn *x* und *y* das Gleiche sagen, gilt auch die umgekehrte Implikation. Damit ist der gegen naive Theorien der Semantik von AFKs erhobene Einwand blockiert.

Abschließend kann ich somit sagen, dass es meiner inferentialistischen Theorie der Semantik von AFKs gelingt, all diejenigen Einwände zu blockieren, die ich ausgehend von IF gegen alternative Theorien der Semantik von AFKs vorgebracht habe. Außerdem wird sie OF gerecht. Damit meine ich gezeigt zu haben, dass ich eine Theorie der Semantik von AFKs präsentiert habe, die vor dem Hintergrund eines semantischen Inferentialismus zumindest allen anderen in dieser Arbeit vorgestellten Theorien vorzuziehen ist.

**David B. Blumenthal**

Freie Universität Berlin
david.b.blumenthal@fu-berlin.de

## Literatur

Bertram, G., D. Lauer, J. Liptow und M. Seel 2008: *In der Welt der Sprache*. Frankfurt am Main: Suhrkamp.

Brandom, R. 1994: *Making It Explicit*. Cambridge: Cambridge University Press.

Davidson, D. 1984: „Radical Interpretation“, in *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press, 125–40.

— 2001: „Rational Animals“, in *Subjective, Intersubjective, Objective*. Oxford: Oxford University Press, 95–105.

Everett, A. 2005: „Against Fictional Realism“, in *The Journal of Philosophy* 102, 624–49.

Frege, G. 1962: „Über Sinn und Bedeutung“, in *Funktion, Begriff, Bedeutung*. Göttingen: Vandenhoeck & Ruprecht, 40–65.

— 1966: „Der Gedanke“, in *Logische Untersuchungen*. Göttingen: Vandenhoeck & Ruprecht, 30–53.

Funkhouser, E. 2006: „The Determinable-Determinate Relation“, in *Noûs* 40, 548–69.

Johnson, W. 1921: *Logic, Vol 1*. Cambridge: Cambridge University Press.

Parsons, T. 1974: „A Meinongian Analysis of Fictional Objects“, in *Grazer Philosophische Studien* 1, 73–86.

Quine, W. 1963: „On What There Is“, in *From a Logical Point of View*. New York: Harper Torchbook, 20–46.

— 1981: *Theories and Things*. Cambridge: Harvard University Press.

Russell, B. 1905: „On Denoting“, in *Mind* 14, 479–93.

Rorty, R. 1983: „Is There a Problem About Fictional Discourse?“, in D. Henrich und W. Iser (Hrg.): *Funktionen des Fiktiven*, München: Wilhelm Fink, 67–93.

Sanford, D. 2011: „Determinates vs. Determinables“, in E. Zalta (Hrg.): *The Stanford Encyclopedia of Philosophy*.

Searle, J. 1979: *Expression and Meaning*, Cambridge: Cambridge University Press.

[11] Aber was – so könnte man fragen – passiert, wenn Peters und Lisas Vaters Ausgabe nicht zum gleichen Werk gehören, z.B. weil Peter eine überarbeitete Neuausgabe von „Asterix bei den Briten“ gekauft hat, die sich in weiten Teilen vom Original unterscheidet? Dann folgt (11) laut der inferentialistischen Theorie nicht mehr aus (10), und dies ist auch genau das, was wir erwarten würden.

Sellars, W. 1954: „Some Reflections on Language Games“, in *Philosophy of Science* 21, 204–28.

— 1997: *Empiricism and the Philosophy of Mind*. Cambridge: Cambridge University Press.

Van Inwagen, P. 1983: „Fiction and Metaphysics“, in *Philosophy and Literature* 7, 67–77

# The Ineliminability of Non-Nominal Quantification

David Dolby

Objectual interpretations of non-nominal quantification seems to offer a non-substitutional treatment of quantification which respects differences of grammatical category in the object language whilst only employing nominal quantification in the metalanguage. I argue that the satisfaction conditions of such interpretations makes use concepts that must themselves be explained through non-nominal quantification. As a result, the interpretation misrepresents the structure of non-nominal quantification and the relationship between nominal and non-nominal forms of generality.

## 1. Introduction

Natural language appears to contain expressions of generality in various syntactic categories: English has not only the (typically) nominal 'something', but also adverbial forms such as 'somehow' and 'somewhere'. Nevertheless, many philosophers, following Quine, have claimed that all genuine quantification is nominal: quantifiers and variables range over a domain of objects which are named by their substituends. Since supposed non-nominal quantifications quantify into positions which do not name, there is no domain of objects for the quantifier to range over: they are at best substitutional quantifications—conveniently abbreviated nominal quantifications over linguistic items—and not generalisations about the extra-linguistic world (Quine 1970: 91–4; Inwagen 2004: 124). The most significant opponent to this position was Arthur Prior, who argued that predicate and sentence generalisations are genuinely quantificational although they involve no domain of objects: indeed, he suggested that any attempt to reduce non-nominal quantification to objectual quantification must fail, since in any interpretation of a quantifier we cannot avoid employing the very form of quantification we seek to explain (1971: 31–47). Positions similar to his have been developed by Christopher Williams (1981), Philip Hugly and Charles Sayward (1996), and Timothy Williamson (1999). Recently, philosophers have become more willing to accept the legitimacy of non-nominal quantification. However, it is often proposed *pace* Prior that the quantifiers should be given an interpretation in terms of a domain of objects, such as properties or propositions—those associated with the substituends by some semantic relation other than naming (Strawson 1997a; Soames 1999, Künne 2003, 2008a; Glock 2003; Rosefeldt 2008). I shall claim that such an account will inevitably appeal to non-nominal generalisation, either explicitly or implicitly, in its statement of truth-conditions, and thereby obscures the relationships that hold between non-nominal quantified sentences and their parts and between nominal and non-nominal forms of generality.

## 2. Quantification and Reference

Questions about quantification are closely related to questions about reference and syntax. Indeed, Quine argued that nominal positions in a sentence are precisely those which are open to quantification. He thought that to quantify into predicate or sentence position would be to

treat a predicate or sentence as if it were a name referring to an object (1966). Prior agreed with Quine that objectual quantification into predicate or name position would misrepresent predicates and sentences as names, but whereas Quine thought all quantification must be objectual Prior thought that non-objectual forms of quantification were possible: quantification into nominal position is objectual, since nominal quantifiers generalise the function of names, which is to refer to objects; but since predicates and sentences do not name objects, predicate and sentential quantifiers do not quantify over objects (1971: 35). Quine and Prior were both motivated in part by a suspicion of abstract objects, a suspicion that was not shared by Peter Strawson. Like Prior, Strawson upheld the possibility of quantification into non-nominal positions, but held that names and general terms alike introduce or specify objects: the abstract noun 'redness' and the general term 'red' introduce the same property of redness despite the difference in their grammatical categories (1997: 4–6; 85–91). Quantification into general term position can therefore be regarded as objectual along with quantification into the position of the abstract noun.[1]

This, then, is one of the motivations for the objectual account of quantification into non-nominal positions: it offers a non-substitutional treatment of non-nominal quantification that respects differences of grammatical category whilst recognising that sentences and general terms have something importantly in common with their nominalizations. Thus, Fraser MacBride has defended the possibility of objectual quantification into predicate position on the basis that the function of predicates may involve reference as well as description (2006a; see also his 2006b). Wolfgang Künne has developed in detail an account of objectual quantification into general term and sentence positions, both in his account of abstract objects and in his quantificational theory of truth, although he has recently adopted an account of quantification more in line with the argument of this paper (2003: 350–7; 2005; 2008a; 2008b; for his current position see his 2010). The theory of truth provides an additional motivation for the objectual interpretation of non-nominal quantification, since it would allow us to give a sententially quantified definition of the truth-predicate without employing substitutional quantification, which many have argued is problematic and would lead to circularity. Scott Soames has also remarked upon the advantages an objectual account of sentential quantification would have for the theory of truth (1999: 41–8).

## 3. Objectual Non-Nominal Quantification

According to the approach under consideration we should understand non-nominal quantification as quantification over objects. Quantification into general term position, for example, may be regarded as quantification over properties. These properties are not, however, referred to by the substituends of the variable but are instead ascribed by them. Glock explains the position as follows:

> By contrast to substitutional quantification, the variables have not just substituends but also *values*, a range of objects with which they are associated, namely properties. By contrast to [nominal] objectual quantification, the substituends of predicate variables do not *name* these values (attributes), they *ascribe* them. (Glock 2003: 58)

Likewise, quantification into sentence position may be regarded as quantification over the propositions which sentences might express. The truth-conditions of a quantification are then given in terms of the values in the domain of the quantifier meeting certain satisfaction conditions. Since Künne's exposition remains the clearest and fullest I shall follow his account.

[1] Note that Strawson also defended the possibility of non-objectual quantification, which he thought was the correct interpretation for adverbial generalisations. See his 1997b.

> So whether '∃F (Ann is F & Ben is F)' expresses a truth depends on whether there is an object within the range of the variable—that is to say, a property (of being somehow)—which satisfies the condition signified by the open sentence 'Ann is F & Ben is F'. A property (of being somehow) meets this condition if and only if it is exemplified by Ann and by Ben. (Künne 2003: 363)

Quantification into sentence position is explained in a similar fashion ('[*P*]' is read as 'the proposition that *P*'):

> If the sentential quantifier subserves higher-order quantification over propositions, it is objectual. Hence whether '∃*P* (The Pythagorean Theorem = [*P*] & *P*)' expresses a truth depends on whether there is an object within the range of the variable, a proposition, that is, which satisfies the condition signified by the open sentence 'The Pythagorean Theorem = [*P*] & *P*'. A proposition meets this condition if and only if it is identical with the Pythagorean Theorem and true. Unsurprisingly, at this point we cannot avoid employing the concept of truth. (Künne: 2003: 363. I have capitalised the sentential variables to avoid confusion.)

The objectual account thus explained seems to offer a way to remove the traditional restriction of quantification to nominal positions while offering a way of understanding non-nominal quantifications as generalisations about the extra-linguistic world (properties and propositions) and not merely about linguistic items (predicates and sentences).

## 4. Implicit Non-Nominal Quantification

According to Prior it is not possible to account for the truth-conditions of non-nominal quantifications in a way that does not rely on non-nominal generality. The difficulty in explaining general term and sentential quantification in terms of properties and propositions arises from the fact that what is relevant to the meaning of a general term quantification over properties is *how* something is said to be if one ascribes a particular property to it, just as what is relevant to a sentential quantification over propositions is *how* things are said to be if one asserts a particular proposition. Having effectively nominalized an expression to refer to an intensional entity we need to denominalize in order to recover the sense of the original non-nominal expression. Nevertheless, the account of quantification given in the previous section appears to provide a systematic account of the satisfaction conditions for sentences with unbound variables in non-nominal positions in a way that does not appeal to non-nominal generality in the metalanguage. If Prior is right, then appearances must deceive.

The first thing to note about the explanations of satisfaction conditions given above is that they involve concepts such as exemplification and truth, concepts which are most plausibly explained in non-nominally quantified terms:

$\forall x \forall y$ ($x$ exemplifies $y \leftrightarrow \exists F$ ($y$ = the property of being $F$ & $x$ is $F$))

Indeed, Künne himself explains truth quantificationally:

$\forall x$ ($x$ is true $\leftrightarrow \exists P$ ($P$ & $x$ = the proposition that $P$)) (2003: 337)

The appearance in the satisfaction conditions of concepts which seem to call for a non-nominal explanation should give us pause. Künne argued that the employment of truth in the satisfaction conditions was unproblematic since the interpretation is a codification of our ordinary understanding of idiomatic non-nominal quantification in natural language. He said moreover that the appearance of truth in such a codification is inevitable (2003: 363–4). It is not, however, inevitable that one appeal to truth in the satisfaction conditions for sentential quantification; that is, not unless one wishes to avoid the explicit appearance of sentential quantification in the satisfaction conditions. For one could easily eliminate truth from the

interpretation of the analysis of truth by employing sentential quantification in the satisfaction conditions as well:

> A proposition $\alpha$ satisfies the open sentence 'The Pythagorean Theorem = [*P*] & *P*' if and only if the Pythagorean Theorem is identical with the proposition that *things are as* $\alpha$ *says they are* and *things are as* $\alpha$ *says they are*.[2] Alternatively: A proposition $\alpha$ satisfies the open sentence 'The Pythagorean Theorem = [*P*] & *P*' if and only if, for some *S*, $\alpha$ says that *S*, where the Pythagorean Theorem is the proposition that *S*, and *S*.

However, if one is willing to tolerate non-nominal quantification in the satisfaction conditions then one might feel there is no need to appeal to a domain of objects in the first place. For one can avoid mention of properties and propositions and account for any form of quantification in the object language by using the same form of quantification in the metalanguage. Hugly and Sayward (1996: 303–16) and Williamson (1999) have developed this approach.

Of course, one may think that truth and exemplification need not be explained in terms of non-nominal generality: one might prefer to give an alternative account in terms of sets, for example. A set-theoretic account, however, is not suitable for all purposes. Neo-Fregeans employ non-nominal quantification in their attempted reduction of arithmetic to logic: if, however, the semantics of non-nominal quantification is set-theoretic then the mathematics is in the logic from the start.[3] The second problem with set-theoretic accounts would be that one of the goals of a semantics for non-nominal quantification is to reveal the structure of our language and thereby tell us something about our linguistic capacities. An account in which the semantic values of predicates are given extensionally will not explain what speakers of a language know when they understand a predicate. For speakers must understand that what the objects in the extension of a predicate '*F*' have in common is that that they are all *F*.

The difficulty of avoiding implicit non-nominal quantification in the satisfaction conditions becomes more apparent when we consider intensional contexts. Note that in the satisfaction conditions Künne gave for sentential quantification the appearance of a sentential variable in the context of a that-clause was treated differently to its appearance outside of a that-clause: a proposition satisfies '*a* = [*P*]' if and only if it is identical with the referent of '*a*'; a proposition satisfies '*P*' if and only if it is true. Now, there is nothing wrong with this in itself, but it illustrates the fact that while we might be able to explain truth in terms of non-nominal quantification, we cannot explain non-nominal quantification in terms of truth. If both appearances of the variable were treated the same way then '*a* = [*P*]' would be satisfied by a proposition $\alpha$ if and only if the referent of '*a*' is identical with *the proposition that* $\alpha$ *is true*. Clearly this is not what is intended: if the referent of '*a*' is not a proposition about a proposition then no proposition in the domain will satisfy the open sentence. It is for that reason that the semantics must give the additional rule for appearances of a sentential variable in the context of a that-clause: a proposition satisfies '*a* = [*P*]' if and only if it is identical with the referent of '*a*'.

However, this rule suffices only for the occurrence of a single sentential variable in a that-clause. Were we to extend our language further, for instance to accommodate expressions such as 'the proposition that *p* and *q*', we should need additional recourse to non-nominal quantification in order to define new relations in the metalanguage. '[*p* & *q*]' should refer to the proposition expressed by '*p* & *q*' but we must define what proposition this is in terms of the semantic values of '*p*' and '*q*'. One systematic way to do this would be to give designation conditions to supplement the satisfaction conditions: the referent of '[$\phi$ & $\varphi$]' is the proposition which results from the conjunction of the proposition expressed by '$\phi$' with the

[2] Like Künne I take the expressions here italicised to function as prosentences.

[3] See Wright 2007 for a discussion of this point as well as an interesting suggestion as to how we should understand interpretations of the quantifiers.

proposition expressed by 'φ'. This would be perfectly feasible. The problem is simply that knowledge of the satisfaction conditions for a sentence involving such an expression would only suffice for understanding if one knew what it is for a proposition to be the conjunction of two other propositions. That is, one would need to know that:

$\forall x \forall P \forall Q$ ($x$ = the conjunction of [$P$] and [$Q$] $\leftrightarrow$ $x$ = [$P$ & $Q$])

In the notion of the conjunction of two propositions the satisfaction conditions would therefore be employing a further concept which must be explained in terms of non-nominal generalisation. Similar notions will need to be introduced for any other extension of the language giving rise to new intensional contexts.

## 5. Implications

One consequence of the explicit or implicit appearance of non-nominal quantification in the satisfaction conditions is that the statement of truth-conditions for a non-nominal quantification will involve greater quantification than the sentence with which we started. Consider the standard interpretation of nominal quantification: this tells us that an existential quantification is true if and only if some object in the domain satisfies a certain condition specified by the open sentence. This condition will be specified with one fewer quantifiers than were contained in the original quantification. For instance, 'Bill kicked *some*one' will be true if and only if *some* object in the domain satisfied the condition specified by 'Bill kicked $x$', and this condition is met by someone if Bill kicked them. The end result is therefore quantified to precisely the same degree as the original sentence.

The situation with objectual non-nominal quantification is different: for the interpretation effectively transforms the non-nominal quantifier (things are thus and so) in the object language into an objectual quantifier (some proposition) in the metalanguage, but is then forced to introduce additional implicit non-nominal quantification into the condition to be met by the values of the variable. If this implicit quantification is made explicit then the interpretation involves more quantifiers than the original quantification.

Quantificational logic holds out the prospect of articulating the relationship between the truth of a quantification and the satisfaction of more basic conditions, of relating a complex sentence to its parts. Complexity may, of course, remain after the analysis: we may translate the predicate 'is married' as '$Fx$' in which case one could argue that the axiom for the satisfaction of the predicate, and therefore the resulting analysis, will be implicitly quantified, since to be married is to be married to someone. But implicit quantification of this sort is something that we can in principle make explicit by translating 'is married' as '$\exists y Rxy$' instead. And while it may not be possible or even make sense to give an exhaustive analysis of a language into its elements, we can make explicit any given implicit quantification, and it might be fruitful to do so. However, the non-nominal quantification implicit in the objectual interpretation's use of the terms 'exemplifies' and 'true', in contrast, could not be made explicit in this way since these terms are introduced by the interpretation itself. Any attempt to use the objectual interpretation to make this explicit would merely generate further occurrences of the very same implicitly quantified concepts.

Of course, how troublesome any of this is will depend on what we take the function of an interpretation of quantification to be. If our interest is simply in providing a workable semantics for the purposes of displaying relationships of implication between sentences of a language with certain expressive capacities, then there is no objection to giving an objectual interpretation of non-nominal quantification. However, the interpretation of quantification is generally taken to have greater significance than this. It is often argued, for instance, that the standard substitutional interpretation is inappropriate for the interpretation of a

quantificational definition of truth, on the basis that truth appears in the truth conditions (Davidson 1996: 273; Platts 1997: 14–5; Künne 2003: 357–60; but see Soames 1999: 42). Likewise, the account of quantifications as potentially infinite conjunctions or disjunctions of substitution instances has been criticised on the basis that the resulting sentences would outstrip the comprehension of speakers (Inwagen 2002: 214–6). Moreover, the interpretation of the quantifiers is often thought to reveal the ontological commitments of someone holding sentences of the object language to be true. The significance given to the interpretation of quantification can only be justified if it in some way reflects what we mean when we make quantified assertions.

If the objectual interpretation of non-nominal quantification is taken to reveal what our understanding consists in, then one might conclude that at some level all quantification is nominal quantification, quantification over objects, or even that it might be possible to communicate understanding of non-nominal quantification in an explanation that only employs nominal quantification. But this would be a mistake, since, as we have seen, the objectual interpretation only suffices to capture our understanding of non-nominal generality by employing concepts whose implicit non-nominal generality it does not make explicit. An interpretation of non-nominal quantification which itself employs non-nominal quantification, as proposed by Hugly and Sayward (1996) or Williamson (1999), would have the merit of displaying the fact that an understanding of one form of generality cannot be reduced to any other: it is in this sense that each form of generality is ineliminable.[4]

**David Dolby**
University of Zurich
dolby@philos.uzh.ch

## References

Davidson, D. 1996: 'The Folly of Trying to Define Truth', *The Journal of Philosophy* 93.6, 263–78.

Glock, H.-J. 2003: *Quine and Davidson on Language, Thought and Reality*. Cambridge: Cambridge University Press.

Hugly, P. & C. Sayward. 1996: *Intentionality and Truth*. Dordrecht: Kluwer.

Inwagen, P. van. 2002: 'Generalizations of Homophonic Truth-Sentences', in *What is Truth?* ed. R. Schantz. Berlin: De Gruyter, 205–222.

— 2004: 'A Theory of Properties', in D. Zimmerman (ed.): *Oxford Studies in Metaphysics* Vol. 1. Oxford: Oxford University Press, 107–138.

Künne, W. 2003: *Conceptions of Truth*. Oxford: Oxford University Press.

— 2005: 'The Modest Account of Truth Reconsidered: with a Postscript on Metaphysical Categories', *Dialogue* 44, 563-96.

— 2008a: 'The Modest, or Quantificational, Account of Truth', *Studia Philosophica Estonica* 1, 122-168.

— 2008b: 'Précis of Conceptions of Truth and Replies to Commentators', *Dialectica* 62 (3), 355–7 & 385-401.

— 2010: 'Replies to Paul Boghossian and Kevin Mulligan', *Dialectica* 64:4, 585–615.

[4] I am grateful to Kai Büttner, Max de Gaynesford, Hanjo Glock, Christoph Pfisterer, Constantine Sandis and Severin Schroeder for helpful discussion.

MacBride, F. 2006a: 'Predicates and Properties: an Examination of P.K. Sen's Theory of Universals', in P. F. Strawson, and A. Chakrabarti (eds.): *Universals, Concepts and Qualities: New Essays on the Meaning of Predicates*. Aldershot: Ashgate, 67–90.

— 2006b: 'Predicate Reference', in B. Smith and E. Lepore (eds.): *The Oxford Handbook of Philosophy of Language*. Oxford: Oxford University Press, 422-74.

Platts, M. 1997: *Ways of Meaning*. 2nd Ed. Cambridge (MA): MIT Press.

Prior, A. 1971: *Objects of Thought*. Oxford: Oxford University Press.

Quine, W.V.O. 1966: 'A Logistical Approach to the Ontological Problem', in *Ways of Paradox and Other Essays*. Cambridge, MA: Harvard University Press.

— 1970:*Philosophy of Logic*. Cambridge (MA): Harvard University Press.

Rosefelt, T. 2008: '"That"-Clauses and Non-Nominal Quantification', *Philosophical Studies* 137, 301–33.

Soames, S. 1999: *Understanding Truth*. Oxford: Oxford University Press.

Strawson, P.F. 1997a: 'Introduction', in *Entity and Identity, and Other Essays*. Oxford: Oxford University Press, 1–19.

— 1997b: 'Concepts and Properties', in *Entity and Identity, and Other Essays*. Oxford: Oxford University Press, 85–91.

Williams, C. 1981: *What is Existence?* Oxford: Oxford University Press.

Williamson, T. 1999: 'Truthmakers and the converse Barcan formula', *Dialectica* 53 (3-4), 253–70.

# Primitive Normativität als Antwort auf den Regelfolgen-Skeptiker

Nadja El Kassar

Ursprünglich auf ihre Theorie der Wahrnehmung beschränkt, bildet *primitive Normativität* in Hannah Ginsborgs jüngstem Aufsatz das Fundament für ihre teil-reduktionistische, naturalistische Antwort auf Saul Kripkes Regelfolgen-Skeptiker. Der vorliegende Beitrag diskutiert und verwirft diese alternative Antwort auf den Regelfolgen-Skeptiker. Die Argumentation erfolgt in drei Schritten. Im ersten Schritt werden die Herausforderung durch den Kripke'schen Skeptiker sowie zwei Standardreaktionen, Dispositionalismus und Anti-Reduktionismus, dargestellt. Der zweite Schritt stellt Hannah Ginsborgs Antwort, insbesondere ihren Begriff der primitiven Normativität, vor. Im abschließenden dritten Schritt werden drei Argumente präsentiert, die nahelegen, dass Ginsborgs Antwort abzulehnen ist. Weder gelingt es ihr eine systematisch kohärente Position zu entwickeln, noch vermag sie, die von Saul Kripke diskutierten Bedingungen für eine akzeptable Replik auf den Regelfolgen-Skeptiker zu erfüllen. Das Modell der primitiven Normativität scheitert sowohl als eigenständige Theorie als auch als Antwort auf Kripkes Regelfolgen-Skeptiker.

## 1. Einleitung

Hannah Ginsborg führt in ihren Beiträgen zur Philosophie der Wahrnehmung den Begriff *primitive Normativität* [*primitive normativity*] ein, der dazu dienen soll eine stabile Mittelposition zwischen entgegenstehenden Theorien aufzustellen: Primitive Normativität ist der Schlüssel zur Vermittlung zwischen Empiristen und Konzeptualisten (Ginsborg 2006b) sowie zwischen Repräsentationalisten und Relationisten (Ginsborg 2011a). In ihren jüngsten Veröffentlichungen erweitert Ginsborg den Anwendungsbereich auf Saul Kripkes ‚Regelfolgen-Skeptiker' und die Frage nach der Normativität von Bedeutung (Ginsborg 2011b; Ginsborg 2012). Dieser Artikel wird sich jedoch ausschließlich mit Ginsborgs Antwort auf den Regelfolgen-Skeptiker auseinandersetzen.[1] Das Ziel ist es, zu zeigen, dass Ginsborgs Konzeption keine zufriedenstellende Antwort auf den Skeptiker bietet. Die Argumentation erfolgt in drei Schritten. Im ersten Schritt wird die Herausforderung durch den Kripke'schen Skeptiker sowie zwei Standardreaktionen auf die Herausforderung dargestellt. Anschließend wird Ginsborgs Antwort, insbesondere ihr Konzept der primitiven Normativität, vorgestellt. Im dritten Schritt präsentiere ich drei Argumente, die nahelegen, dass Ginsborgs Antwort abzulehnen ist: Der Begriff primitive Normativität kann nicht widerspruchsfrei expliziert werden und ermöglicht ferner keine akzeptable Replik auf den Regelfolgen-Skeptiker. Der Artikel nimmt also keinen Vergleich vor zwischen Ginsborgs Antwort auf den Regelfolgen-Skeptiker und anderen Antworten auf den Regelfolgen-Skeptiker. Die hier entwickelte Kritik ist Ginsborg-intern und beschränkt sich auf Einwände gegen ihre Konzeption, sowie auf die sich anschließende Frage, *ob* ihre Antwort dem Regelfolgen-Skeptiker überhaupt angemessen ist.

[1] Da das Regelfolgen-Problem und die Normativität von Bedeutung eng verbunden sind, ist es natürlich unumgänglich, dass die Thematik der Normativität von Bedeutung auch behandelt wird. Für eine ausführliche Diskussion von Ginsborgs Ausführungen zur Normativität von Bedeutung siehe jedoch (Haddock 2012).

## 2. Das Regelfolgen-Problem nach Saul Kripke

Im Regelfolgen-Problem nach Kripkes Interpretation (Kripke 1982) stellt der Regelfolgen-Skeptiker unter anderem heraus, dass die Bedeutung, die durch eine Regel erfasst wird, arbiträr ist. Regeln sind beispielsweise in der korrekten Begriffsverwendung und der korrekten Fortsetzung einer Zahlenfolge manifest. Nach Ansicht des Skeptikers kann ein Akteur nie rechtfertigende Gründe dafür anführen, dass er den Begriff *plus* in seinen bisherigen Verwendungen tatsächlich als „Addition bedeutend" verwendet hat. Es gibt keine Tatsache, die beweist, dass *plus* „Addition" heißen sollte, und nicht vielmehr „Quaddition". Das Fehlen einer Begründungsgrundlage manifestiert sich an dem folgenden Problem: Jede vergangene Addition kann ebenso mit einer anderen Regel, die nicht die Plus-Regel ist, in Einklang gebracht werden; jede vergangene Addition könnte genauso erfolgreich durch die Quus-Regel erklärt werden.[2] Der Akteur hätte *a fortiori* keinen Grund dafür den Begriff *plus* normativ bindend mit „Addition" zu verknüpfen. Jede Formulierung einer Regel, die die Bedeutung von *plus* erfasste, würde durch den Skeptiker in Zweifel gezogen werden können. Wer dem Skeptiker entkommen möchte, muss demnach zwei Aufgaben lösen: Erstens muss auf eine Tatsache verwiesen werden, die die Bedeutung der Regel ‚*plus* bedeutet Addition' konstituiert und mögliche Alternativen wie ‚*plus* bedeutet Quaddition' ausschließt. Zweitens muss gezeigt werden, dass das Subjekt eine Rechtfertigung für das Bestehen der Tatsache ‚*plus* bedeutet Addition' geben kann.

> First, [the skeptic, N.E.] questions whether there is any *fact* that I meant plus, not quus, that will answer his sceptical challenge. Second, he questions whether I have any reason to be so confident that now I should answer '125' rather than '5'. The two forms of the challenge are related. I am confident that I should answer '125' because I am confident that this answer also accords with what I meant. (Kripke 1982: 11, Hervorhebung im Original, N.E.)

In Antwort auf das Regelfolgen-Problem werden zwei Standardpositionen vertreten, ein Dispositionalismus sowie ein Anti-Reduktionismus. Dispositionalisten wollen das Verhältnis zwischen Begriff und Bedeutung, zwischen Regel und Befolgen der Regel ausschließlich durch Dispositionen erklären. Das Subjekt meinte ‚plus' und nicht ‚quus', weil es die Disposition besitzt die Summe von zwei Zahlen anzugeben und nicht die ‚Qumme'. Anti-Reduktionisten betrachten Bedeutung und Regel-Befolgung als *sui generis*, als nicht reduzierbar auf Zustände oder Dispositionen. Beide Positionen werden aus verschiedenen, allgemein bekannten Gründen abgelehnt. Hier sollen nur die zwei für Ginsborg wichtigsten Gründe genannt werden. Dispositionalisten wollen die Normativität von Regeln und Bedeutungen nicht erklären, da etwa Bedeutung nicht normativ ist (z.B. (Hattiangadi 2006)). Laut Anti-Reduktionisten hingegen sind Regeln und Bedeutungen zwar normativ, aber die Theorien können die genaue Konstitution von Regeln nicht präzisieren und können damit u.a. auch nicht erklären, warum sie kausal wirkungsvoll sind (Ginsborg 2011b: 229f.).

## 3. Hannah Ginsborgs Analyse des Regelfolgen-Problem

In dieses Problemfeld tritt nun Ginsborg. Sie schlägt vor, dass die beiden Aufgaben getrennt und in umgekehrter Reihenfolge zu behandeln sind. Zuerst solle gezeigt werden, dass die Frage nach einer Rechtfertigung des Subjekts unmotiviert ist. Dann könne mit denselben Mitteln, die für die zweite Aufgabe verwendet wurden, die erste Aufgabe erfüllt werden.

---

[2] Ich lasse das mit dem Regelfolgen ebenfalls verbundene Regress-Problem unbetrachtet, da es Ginsborg primär um den „Normativitäts-Aspekt" (Esfeld 2003: 129) beim Regelfolgen geht und nicht um den „Infinitäts-Aspekt" (ibid.). Für Kritik an Ginsborgs Konzeption auf Basis des Regress-Problems siehe (Haddock 2012).

Zur Bearbeitung der zweiten Aufgabe führt Ginsborg das folgende Wittgenstein-inspirierte Beispiel ein: Man stelle sich ein Kind vor, das gelernt hat die Addiere-zwei-Reihe zu vervollständigen. Der Lernprozess erfolgte durch Beobachtung und nicht durch explizite Instruktion. Im Szenario ist festgelegt, dass das Kind die Regel „Addiere-zwei" nicht beherrscht, da ihm die dafür erforderlichen Begriffe fehlen. Nehmen wir nun an, dass das Kind die Reihe ‚... 36, 38, 40' mit ‚42' fortsetzen möchte. Wenn wir das Kind unterbrechen und fragen, warum ‚42' die nächste Zahl ist, wird es keinen Grund angeben können. Dennoch, so betont Ginsborg, wird es den möglichen Vorschlag ‚43' ablehnen und auf der Fortsetzung mit ‚42' insistieren, denn dieses ist seinem Empfinden nach die angemessene [„appropriate" (Ginsborg 2011b: 234)] Fortsetzung. ‚42' ist die Zahl, die es nennen sollte [„what she ought to say" (ibid.)]. Ginsborg nennt dieses Empfinden *feeling*, *awareness*, oder auch *consciousness* und behandelt die Ausdrücke gleichbedeutend (Ginsborg 2006b; Ginsborg 2011a; Ginsborg 2011b). Die Normativität, die sich im Insistieren manifestieren würde, ist laut Ginsborg eine *primitive* Normativität, denn dem Kind mangelt es an den Begriffen, die für die entwickelte [*sophisticated*] Normativität notwendig wären. Aus demselben Grund kann das Verhalten des Kindes auch nicht durch Regelbefolgung erklärt werden. Das Verhalten (des Kindes) muss vielmehr folgendermaßen erklärt werden: Es besitzt die Disposition die Zahlenreihe mit der Zahl ‚42' fortzusetzen. Diese Disposition ist jedoch normativ aufgeladen, da das Kind das Bewusstsein hat, dass ‚42' folgen sollte und ‚43' nicht folgen sollte. Diese normative Aufladung ohne Regelverstehen oder Begriffsbesitz wird ausgedrückt durch primitive Normativität: „[Primitive normativity is] normativity which does not depend on conformity to an antecedently recognized rule." (Ginsborg 2011b: 233) *Das* ist die Antwort auf die zweite Aufgabe. Das Kind ist gerechtfertigt und besteht auf der Fortsetzung durch ‚42', weil es das primitive Bewusstsein hat, dass diese Fortsetzung der Reihe einfach angemessen ist. Damit ist es schlicht irrelevant, ob das Kind auf dem Weg durch die Zahlenreihe eine von ihm verstandene Regel befolgt hat und ob die Fortführung auf einer nicht-ambigen Tatsache beruht.

Wie kann nun mit Hilfe von primitiver Normativität auf die erste Aufgabe geantwortet werden? Die Aufgabe war ein Faktum anzugeben, das festlegt, dass in der Addiere-zwei-Reihe etwa tatsächlich ‚42' folgen sollte und nicht ‚43'. Wir brauchen, laut Ginsborg, jedoch gar nicht auf ein Faktum zu verweisen, da das primitive Bewusstsein des Kindes ausreicht: Für das Kind soll ‚42' folgen, weil nur diese Fortsetzung mit dem Angemessenheits-Empfinden des Kindes einher geht. Dieses Angemessenheits-Empfinden ist bezogen auf Angemessenheit *simpliciter*, unabhängig davon, ob das Subjekt die Regel für die Addiere-Zwei-Reihe bereits erfasst hat (ibid.: 234).

Im Regelfolgen-Problem ist primitive Normativität also der Schlüssel, der laut Ginsborg die richtige Antwort auf Kripkes Skeptiker eröffnet. Ihre Antwort ist, wie sie es nennt, „partly reductionist" und „naturalistic" (ibid.: 230, 237): Das Fortsetzen von Reihen, aber auch Zählen und Sortier-Tätigkeiten, werden auf Dispositionen reduziert, die stets ‚begleitet' sind von primitiver Normativität. Ginsborg erklärt:

> The normative proviso builds into your disposition the feature that every response you are disposed to give involves a claim to its own appropriateness to the context in which you give it. (ibid.: 244)

Beispiele für solche Dispositionen sind bei Menschen Sortier-Dispositionen (z.B. Grünes zu Grünem) und Zähl-Dispositionen (z.B. Reihen von geraden Zahlen). Darüber steht auf zweiter Ebene die Disposition auf eine bestimmte Weise auf „Training" zu reagieren (ibid.: 236). Diese Dispositionen teilen Menschen mit nicht-menschlichen Lebewesen. Auch sie haben die angeborene Disposition auf Training in einer gewissen Weise zu reagieren; z.B. Tauben, die trainiert wurden Perlen farblich zu sortieren und im richtigen Kontext dies auch tun. Die Dispositionen gehören zur Natur der jeweiligen Spezies, ob Taube oder Mensch (ibid: 237). Dennoch unterscheiden sich die Dispositionen von nicht-menschlichen

Lebewesen auf fundamentale Weise von den Dispositionen von Menschen, denn laut Ginsborg sind erstere nicht mit primitiver Normativität verbunden. Wenn eine Taube Grünes zu Grünem sortiert, geht das nicht mit dem Bewusstsein einher, dass sie angemessen handelt oder so handelt, wie sie handeln sollte.

Primitive Normativität findet sich nach Ginsborg nicht nur in der Fortsetzung von Reihen, sondern auch in Begriffs- und Spracherwerbsprozessen (ibid.: 235); sie geht noch weiter und betont, dass das Angemessenheits-Bewusstsein eine Bedingung für die Aneignung von Begriffen sei. Ein Kind könne die Begriffe *grün* oder auch *Pyramide* nur deswegen empirisch aneignen, weil es von primitiver Normativität begleitete Dispositionen, etwa zum Sortieren von einem grünen Stein zu anderen grünen Steinen, besitze (Ginsborg 2006a: 420). Genau darin liegt einer der zwei Vorteile, die Ginsborg in ihrer eigenen Konzeption im Vergleich zum Dispositionalismus und Anti-Reduktionismus erkennt: Erstens biete sie eine wirkliche Lösung für Kripkes Skeptiker, und zweitens könne sie erklären, wie Kinder Begriffe und Regeln lernen, nämlich auf Grundlage von primitiver Normativität.

Ginsborgs Beispiele beschränken sich auf Kinder, aber sie betont, dass auch das Verhalten vollkompetenter Sprecher in primitiver Normativität gründet. Primitive Normativität und entwickelte [*sophisticated*] Normativität, die kompetenten Begriffsbesitz sowie Regelverstehen umfasst, koexistieren: Eine Erwachsene, die die Addiere-zwei-Regel beherrscht, betrachtet ihre Fortführung durch ‚42' nach ‚40' auch als angemessen *simpliciter*, unabhängig davon, ob sie in den vorherigen Teilen der Reihe die Regel befolgt hat (Ginsborg 2011b: 234). Die These, dass auch die Erwachsene ein Angemessenheits-Bewusstsein hat, wenn sie die Zahlen-Reihe mit ‚42' fortführt, ist laut Ginsborg eine anthropologische These (ibid.: 240), die auf einer von uns allen geteilten „pretheoretical intuition" basiere (ibid.).[3]

## 4. Einwände gegen Ginsborgs Theorie primitiver Normativität

In diesem vierten Abschnitt werde ich nun drei Einwände hervorbringen, die Ginsborgs Theorie in Hinblick auf ihre innere Konsistenz sowie ihre Angemessenheit als Antwort auf den Regelfolgen-Skeptiker in Frage stellen.

Der erste Einwand beginnt mit der Frage, was *primitive Normativität* überhaupt meint. Ginsborg definiert primitive Normativität als ein primitives Bewusstsein, ein Empfinden, das vorhergehendes Regel-Verstehen nicht voraussetzt (Ginsborg 2011b: 233). Genau diese besondere Eigenschaft, die Unabhängigkeit von einem vorausgehenden Regel- oder Begriffsverständnis, ist essentiell für eine der Funktionen, die laut Ginsborg ihre Theorie so hervorragend macht: Primitive Normativität ist nicht-begrifflich, da sie keinen Begriffsbesitz voraussetzt und ist damit die beste Grundlage für eine nicht-zirkuläre Erklärung der Aneignung von empirischen Begriffen und Regeln (ibid.: 238). Diese Facette von primitiver Normativität steht jedoch ganz offenbar in Spannung mit einer anderen Eigenschaft von primitiver Normativität: Ginsborg betont, dass nur Menschen für primitive Normativität empfänglich sind. Die Handlungen von Tieren sind ‚blind', ohne Angemessenheits-Bewusstsein (Ginsborg 2006a: 420; Ginsborg 2006b: 36of.; Ginsborg 2011a: 152; Ginsborg 2011b: 237). Genauer betrachtet ist diese Ausgrenzung von Tieren jedoch unverständlich: Wenn Ginsborgs Aussagen zur Relation zwischen primitiver Normativität und Begriffsbesitz zu Ende gedacht werden, stellt sich die Frage, warum nicht-begriffsbegabte Tiere nicht genauso wie noch nicht begriffsbegabte Kinder empfänglich für primitive Normativität sind. Gibt es ein Zuschreibungskriterium, das Kinder erfüllen, Tiere aber nicht?

[3] Ginsborg fügt hinzu: Nur Philosophen kämen auf die Idee die Fortführung durch ‚43' als gleichermaßen mit einem Bewusstsein von Angemessenheit, nämlich angemessen zu einer Quaddiere-zwei-Regel, verbunden vorzustellen (Ginsborg 2006a: 426).

Zur Beantwortung dieser Frage müssen wir zurück zu Ginsborgs Ausführungen gehen. Wie erkennt man, dass und ob ein Lebewesen empfänglich für primitive Normativität ist? Ginsborgs Ausführungen zum Regelfolgen-Problem enthalten keinerlei Feststellungen dazu, doch können ihre diesbezüglichen Erklärungen aus ihrer Wahrnehmungskonzeption übernommen werden. Über primitive Normativität in der Wahrnehmung behauptet Ginsborg, dass das beobachtete Sortierverhalten eines Kindes Rückschlüsse auf die Wahrnehmungsprozesse zulasse. Ein Kind, das Pyramiden in einer Schachtel sammelt und sich weigert Dreiecke in diese Schachtel werfen zu lassen, beschreibt Ginsborg folgendermaßen: Das Kind habe das Bewusstsein, dass die Pyramiden zusammengehören, dass es angemessen ist eine Pyramide zu Pyramiden zu sortieren, aber nicht angemessen ist ein Dreieck zu Pyramiden zu sortieren. Gleiches soll für das Fortsetzen der Addiere-zwei-Reihe gelten. Das Verhalten des Kindes lässt nach Ginsborg direkte Rückschlüsse auf interne Prozesse zu (Ginsborg 2006a: 364; Ginsborg 2006b: 419).

Diese Methode der Zuschreibung eines Bewusstseins von primitiver Normativität bei einem Kind schließt jedoch weder die Möglichkeit, noch die Korrektheit der Zuschreibung von primitiver Normativität zu nicht-begriffsbegabten Tieren aus. Damit ihre Argumente Tiere von der Empfänglichkeit für primitive Normativität ausschließen, müsste Ginsborg beispielsweise zusätzlich zeigen, dass eine Taube, die Farben sortieren kann, etwa blaue Perlen nicht aus der Schachtel mit grünen Perlen entfernen würde. Mit der aktuellen Diagnose-Methode allein kann Ginsborg nicht gegen die Zuschreibung von primitiver Normativität zu nicht-menschlichen Tieren argumentieren.[4]

Was bedeutet das für Ginsborgs Konzeption? Zuvorderst deuten diese Punkte auf eine problematische Unschärfe des Begriffs *primitive Normativität*: Es gibt Widersprüche zwischen der Gruppe der Lebewesen, denen Ginsborg primitive Normativität zuschreiben möchte, und der Gruppe der Lebewesen, denen primitive Normativität auf Basis von Ginsborgs Ausführungen zugeschrieben werden kann. Ein derart unscharfer Begriff, dessen Darstellung inkompatible Festlegungen enthält, kann nicht die definitive Grundlage für ein so kontrovers diskutiertes Problem wie das Regelfolgen-Problem bieten.

Ginsborg könnte versuchen diesen Einwand mit Verweis auf eine Endnote in einem ihrer Aufsätze zu entkräften. Dort merkt sie an, dass sie nicht darauf verpflichtet sei primitive Normativität für Tiere auszuschließen (Ginsborg 2006b: 435, fn.32). Diese Antwort ist jedoch klarerweise *ad hoc*. Erstens erläutert Ginsborg nicht, warum diese Verpflichtung nicht besteht. Zweitens wird damit unverständlich, warum Ginsborg in ihren Haupttexten stets Tiere von Kindern in Hinsicht auf primitive Normativität unterscheidet (Ginsborg 2006a; Ginsborg 2006b; Ginsborg 2006c; Ginsborg 2011a; Ginsborg 2011b). Auch die Beschreibung ihrer Thesen über primitive Normativität als „anthropological claim“ (Ginsborg 2011b: 240) wirkt damit maximal fehlleitend und mindestens unmotiviert.[5]

---

[4] Versuche, das Bestehen von primitiver Normativität in anderen Kennzeichen zu verankern, drohen auf Einwände des Regelfolgen-Skeptikers zu treffen. Vgl. (Haddock 2012: 150)

[5] Zwei weitere Reaktionen stehen Ginsborg offen. Erstens könnte sie korrigierend anmerken, dass sie gar nicht behaupte, dass primitive Normativität nicht-begrifflich sei. Sie behaupte nur, dass eine Theorie der Aneignung von Begriffen oder Regeln aus Erfahrung nicht schon den Besitz ebendieser voraussetzen dürfe, da sonst die Erklärung zirkulär sei (Ginsborg 2006a: 407f.). Damit sei nicht ausgeschlossen, dass das Subjekt nicht schon bestimmte Begriffe besitzt oder Regeln versteht. Eine solche teil-begriffliche Rekonstruktion würde sich allerdings weiteren eigenen Einwänden gegenüber sehen, vgl. dazu (Haddock 2012). Zweitens könnten Unterstützer Ginsborgs einwenden, dass einige Tiere durchaus empfänglich für primitive Normativität sind; bei ihnen ist primitive Normativität nur weniger komplex. Sie besitzen normativ aufgeladene Dispositionen, die oft im Kontext von anderen Dispositionen erfolgreich aktualisiert werden. Diese zweite Reaktion kann durchaus vorgebracht werden, doch sie wäre klarerweise nicht mehr in Übereinstimmung mit der Position, die in Ginsborgs Schriften entwickelt wird (s.o.).

Ginsborgs „anthropological claim“ (Ginsborg 2011b: 240) behauptet auch die Zuschreibung von primitiver Normativität zu Erwachsenen: primitive Normativität und entwickelte Normativität koexistieren bei Erwachsenen. Ginsborg argumentiert also für folgendes Bild: Wenn eine Erwachsene die Addiere-Zwei-Reihe mit ‚42' vervollständigt, dann tut sie das, weil sie die Regel *Addiere-Zwei* versteht und weil sie das primitive Bewusstsein hat, dass sie so handeln sollte. Dieses Bild führt jedoch zu meinem zweiten Einwand: Was ist bei Erwachsenen überhaupt noch die Rolle von primitiver Normativität?

Ginsborg ist sich bewusst, dass primitive Normativität bei Erwachsenen leicht zu übersehen ist, bietet aber sogar Erklärungen dafür an, wie es dazu kommen kann, dass man sie übersieht (ibid.: 429f.). Sie fügt hinzu, dass man primitive Normativität allerdings sichtbar machen könne, indem man sich die Aneignung von Begriffen und Regeln aus empirischer Erfahrung sowie ihre eigene Interpretation der kantischen Konzeption ästhetischer Urteile anschaue. Ginsborgs Kant-Interpretation kann hier nicht angemessen entwickelt, geschweige denn bewertet werden, doch die Grundidee soll knapp skizziert werden, da sich an dieser Argumentationsweise ein grundlegendes Problem für Ginsborgs Theorie primitiver Normativität manifestiert.

Das von Kant diskutierte Standardproblem ästhetischer Urteile (z.B. ‚Dieses Bild ist schön.') ist, dass sie einerseits die Zustimmung anderer Beobachter verlangen, andererseits aber auf keine Eigenschaft hinweisen können, die die Zuschreibung der Schönheit rechtfertigen könnte. Die Lösung liegt laut Ginsborgs Kant-Interpretation in der Einführung einer primitiven Normativität, die nicht auf objektiv erkennbaren Objekt-Eigenschaften basiert. Das Subjekt hat das primitive Bewusstsein, dass sie das Objekt als schön bewerten sollte. Ginsborg folgert hieraus zweierlei: Erstens, das Bestehen primitiver Normativität in ästhetischen Urteilen; zweitens, die Koexistenz von entwickelter und primitiver Normativität bei Erwachsenen (Ginsborg 2006a).

Jedoch kann die Position durch diese Ergänzung nicht rehabilitiert werden. Vielmehr schwächt sie die Koexistenz-These: Ginsborgs Argumente für die Koexistenz bestehen aus weiteren Theorien, die mit guten Gründen bezweifelbar sind. Sowohl Kants Analyse, als auch Ginsborg Interpretation ebendieser sind nicht unumstritten.[6] Gleiches gilt auch für ihre Konzeption der Aneignung von Begriffen aus der Wahrnehmung.[7] Der Verweis auf weitere Theorien schwächt die Koexistenz-These und auch den Begriff *primitive Normativität* an sich, da die Stärke beider essentiell von der Gültigkeit der zugrundeliegenden Theorien abhängt und keine unabhängige Plausibilität hat. Primitive Normativität wird damit zu einem bloßen theoretischen Postulat.[8]

Die ersten beiden Einwände bezogen sich auf den Begriff der primitiven Normativität. Der dritte Einwand kehrt nun zurück zur Funktion des Begriffs *primitive Normativität* in Ginsborgs Konzeption. Ginsborg führt den Begriff ein, um auf Kripkes Regelfolgen-Skeptiker zu antworten. Das regelgemäße Verhalten, das noch nicht auf dem Verständnis der Regel basiert, ist durch Dispositionen gerechtfertigt, die von einem Bewusstsein primitiver Normativität begleitet sind. Die Dispositionen gehören zur Natur der Spezies Mensch. Die

[6] Siehe z.B. (Allison 2001).

[7] Für einen Einblick siehe z.B. (Peacocke 2009).

[8] Hinter diesem Einwand steht die folgende allgemeinere Diagnose: Ginsborgs Konzeption ist grundlegend falsch, da sie ein *layer-cake model* der Normativitätsentwicklung entfaltet: Die erste Schicht *primitive Normativität* wird im Laufe der Entwicklung des Individuums durch eine weitere Schicht, *entwickelte Normativität,* ergänzt. Die Gegenthese, die in diesem Rahmen nicht weiter begründet werden kann, lautet, dass Normativität gemäß eines *Ersetzungsbildes* verstanden werden muss; die Aneignung von entwickelter Normativität geht mit der Ersetzung von primitiven Vorstufen einher. Weitere Implikationen, die sich aus unserer Verwendung des Begriffs *layer-cake model* ergeben könnten, müssen unbeachtet bleiben, da es hier nur um die strukturelle Ähnlichkeit zwischen den so kategorisierten Theorien geht. Für Kritik an verschiedenen Arten von *layer-cake models,* siehe z.B. (Sellars 1973) und (Lauer 2012).

begleitende primitive Normativität jedoch ist *sui generis*. Ginsborgs Einführung von „primitiver Normativität“ ähnelt damit einer Strategie, die Kripke in seiner Diskussion des Regelfolgen-Skeptikers diskutiert und umgehend verwirft:

> Perhaps we may try to recoup [the state of meaning addition by ‘plus’, N.E.], by arguing that meaning addition by ‘plus’ is a state even more *sui generis* than we have argued before. Perhaps it is simply a primitive state, not to be assimilated to sensations or headaches or any 'qualitative' states, nor to be assimilated to dispositions, but a state of a unique kind of its own. (Kripke 1982: 51)

Kripke bezeichnet diese teil-reduktionistische Position als unwiderlegbar, aber „verzweifelt“ (ibid.): Der postulierte *primitive state* soll introspektiv nicht erkennbar sein, aber dennoch sollen wir uns seiner bewusst sein, da wir uns ja sicher sind, dass *plus* Addition meint und dieses Bewusstsein der Sicherheit auf diesem *primitive state* basiert. Dieser *primitive state* bleibt jedoch undurchsichtig und kann damit keinerlei wirkliche explanatorische Funktion erfüllen (ibid.).[9] Gleiches gilt für die Theorie primitiver Normativität: Ginsborg versucht zwar zu zeigen, dass primitive Normativität nicht auf problematische Weise *sui generis* ist (Ginsborg, 2011b: 228), doch unsere unbeantwortet gebliebenen Nachfragen in den Einwänden 1 und 2 haben gezeigt, dass dieser Versuch scheitert.[10] Sie zeigen, dass auch primitive Normativität undurchsichtig ist und keine explanatorische Funktion erfüllen kann. Ginsborgs Theorie fällt auf die obige Antwort auf den Regelfolgen-Skeptiker zurück, die Kripke bereits verworfen hatte, und verliert damit Bedeutung und Gewicht in der Auseinandersetzung mit dem Regelfolgen-Skeptiker.

## 5. Fazit

Es bleibt abschließend festzuhalten, dass Ginsborgs mit Hilfe primitiver Normativität entwickelte Theorie keine zufriedenstellende Antwort auf Kripkes Regelfolgen-Skeptiker bietet. Die dem Begriff der primitiven Normativität inhärenten Probleme stellen seine Kohärenz und seine Funktionalität grundsätzlich in Frage. Zudem ist die Konzeption Ginsborgs anscheinend identisch mit einer von Kripke verworfenen teil-reduktionistischen Antwort auf den Regelfolgen-Skeptiker und scheitert damit auch im breiteren Problemfeld des Regelfolgens. Es ist zweifellos denkbar, dass die Verwendung des Begriffs der primitiven Normativität für Wahrnehmungstheorien rehabilitiert werden kann, doch den Herausforderungen durch Kripkes Regelfolgen-Skeptiker ist der Begriff nicht gewachsen.[11]

**Nadja El Kassar**
Universität Potsdam
nadja.el.kassar@uni-potsdam.de

[9] Da ich das Regress-Problem hier nicht diskutiert habe, werde ich Kripkes „wichtigeres“ Argument Wittgensteins, das die Position ultimativ in den Regress zurückfallen sieht (Kripke 1982: 51ff.), ignorieren.

[10] Ginsborg könnte einwenden, dass sie aber doch mehr zur ‚Geschichte‘ von primitiver Normativität sagt und damit keine *black box* einfügt, doch hier greift wieder die Anmerkung zum zweiten Einwand aus Fußnote 8: Ein *layer-cake model* von Normativität ist falsch.

[11] Für hilfreiche Diskussionen und Anmerkungen danke ich Logi Gunnarsson, David Löwenstein, Luz Christopher Seiberth und Thomas Jussuf Spiegel.

## Literatur

Allison, H. E. 2001: *Kant's Theory of Taste: A Reading of the Critique of Aesthetic Judgment*. Cambridge University Press.

Esfeld, M. 2003: „Regelfolgen 20 Jahre nach Kripkes Wittgenstein“, *Zeitschrift für philosophische Forschung* 57, 128-138.

Ginsborg, H. 2006a: „Aesthetic Judgment and Perceptual Normativity.“ *Inquiry* 49(5), 403-437.

— 2006b: „Empirical Concepts and the Content of Experience.“ *European Journal of Philosophy* 14(3), 349-372.

— 2006c: „Kant and the Problem of Experience.“ *Philosophical Topics* 34(1-2), 59-106.

— 2011a: „Perception, Generality and Reasons“, in Reisner, A. und A. Steglich-Petersen (Hrg.), 131-157.

— 2011b: „Primitive Normativity and Skepticism About Rules“, *Journal of Philosophy* 108(5), 227-254.

— 2012: „Meaning, Understanding and Normativity“, *Aristotelian Society Supplementary Volume* 86(1), 127-146.

Haddock, A. 2012: „Meaning, Justification, and 'Primitive Normativity'“, *Aristotelian Society Supplementary Volume* 86(1), 147-174.

Hattiangadi, A. 2006: „Is Meaning Normative?“, *Mind and Language* 21(2), 220-240.

Kripke, S. A. 1982: *Wittgenstein on Rules and Private Language*. Cambridge (MA): Harvard University Press.

Lauer, D. 2012: „Expressivism and the Layer Cake Picture of Discursive Practice”, *Philosophia* 40, 55–73

McLaughlin, B. und A. Beckermann (Hrg.) 2009: *The Oxford Handbook of Philosophy of Mind*. Oxford: Oxford University Press.

Peacocke, C. 2009: „Concepts and Possession Conditions“, in B. McLaughlin and A. Beckermann (Hrg.), 437-456.

Reisner, A. und A. Steglich-Petersen (Hrg.) 2011: *Reasons for Belief*. Cambridge: Cambridge University Press

Sellars, W. 1973: „Reply to Marras“, *Canadian Journal of Philosophy* 2, 485-493.

# Relativism and Superassertibility

Manfred Harth

In this paper, I shall explore the prospects of a shape of relativism in ethics that is supposed to be an alternative to the relativist account of truth that recently emerged in semantics, which relativizes the truth predicate by adding an extra parameter for a perspective, a context of assessment or the like. This alternative is based on an epistemic account of ethical truth as superassertibility; and the straightforward road to relativism then is to hold that two contradictory propositions may be both stably assertible relative to divergent starting points of information. Yet this sort of relativism requires a relativization of the truth predicate – which was to be avoided from the outset. I'll discuss the following response to this problem: limiting relativism to epistemic relativism conjoint with an account of ethical truth as monadic superassertibility – thereby denying the possibility that two contradictory propositions may be both stably assertible – and a restriction to intuitionistic logic. I'll conclude that this account, which I call Anti-realist Epistemic Relativism, yields a promising approach to ethical relativism that presents an alternative to semantic truth-relativism.

## 1. The Role(s) of Faultless Disagreement

**1.** The idea that in some area of thought and language there may be genuine disagreement in which nobody can be accused of having failed or of having done something wrong or of being mistaken etc. – for which the term "faultless disagreement" has established in the literature – plays a crucial role in the philosophical debate about relativism. Yet there are at least two different roles to play for faultless disagreement depending on the region of discourse one is concerned with. For discourse about matters of taste, humour and the like it has a *motivational* role. That is, the possibility of disagreement in which nobody needs to be wrong is a widespread intuition, or arguably the pre-theoretic view, about matters of taste or humour that should be explained by a semantic theory. For other regions, notably for ethical discourse, with which I'm concerned here, the possibility of faultless disagreement plays a different role: it doesn't seem to be the common pre-theoretic view or a widespread intuition that has to be explained by semanticists – quite the contrary, many ordinary people seem to be inclined to think that in a disagreement about morals one party has to be wrong, although it might be hard or in some cases even impossible to find out which one is. Nevertheless, some people are inclined to think otherwise, and we may call them *relativists*. So one way to be an ethical relativist is to claim that for moral matters faultless disagreement is possible – this is the second role of the possibility of faultless disagreement: it *defines* a relativist position. That is, ethical relativists of that shape maintain that there may be ethical questions, e.g. the question whether or not abortion is always morally wrong, to which more than one correct answer can be given: Alice might belief that abortion is always wrong and Bob might belief that it is not, and neither of them needs to be at fault. However, simple considerations show that faultless disagreement thus conceived is not possible; it gives rise to contradiction, which is proved by the following *Simple Deduction* (cf. Wright 2002: 105):

| | | |
|---|---|---|
| (1) | A believes that P and B believes that not-P | Disagreement |
| (2) | Neither A nor B has made a mistake | Faultlessness |
| (3) | ∀X: If X believes that P, and not-P, X has made a mistake | Error Principle |

| | | |
|---|---|---|
| (4) | Not (A believes that P, and not-P) | (2), (3), MTT |
| (5) | Not (B believes that not-P, and not-not-P) | (2), (3), MTT |
| (6) | Not-not-P | (1), (4), MPT |
| (7) | Not-P | (1), (5), MPT, TNE |

So in order to be relativist of the shape in question one has to block the Simple Deduction and so to save the possibility of faultless disagreement. The straightforward road to this shape of relativism, then, is blocking the Simple Deduction by *relativizing the truth predicate*. This solution is much discussed in the recent debate under the labels "genuine relativism", "relativism about truth" or "truth-relativism" (Egan 2007, 2009; Kölbel 2002, 2004a, b, 2007, 2009; Lasersohn 2005, 2009; MacFarlane 2003, 2005, 2007, 2011, MS). So to claim that faultless disagreement, in the sense of our two assumptions *Disagreement* and *Faultlessness*, is possible implies a relativization of the truth predicate for propositions to some extra parameter over and above the "relativization" to possible worlds. Since it blocks the Simple Deduction – its new conclusion is, roughly, that P is true *for A* and not-P is true *for B* – it seems to be a coherent relativist position. But there remain doubts: doubts concerning a truth-relativist account of the role and purpose of assertion, the so-called *Evans' challenge* (Evans 1985: 349-50, and Garcia-Carpintero 2008: 141-142), doubts concerning the conceptual connection between assertion, belief and relative truth (Harth 2013a), doubts as regards the Equivalence Schema for the meta-language truth-predicate (Harth 2013a), doubts concerning relative truth conditions and their role in constituting shared contents on which two thinkers disagree (Capps, Lynch and Massey 2009), and, finally, doubts concerning the truth-relativist explanation of faultless disagreement (cf. Binderup 2008; Coliva and Morruzzi 2012; Francén 2010; Harth 2013b; Moruzzi 2008; Rosenkranz 2008 and Stojanovic 2007). Of course, there have been efforts to dispel (most of) these doubts, notably by John MacFarlane (2003, 2005, 2007, MS) and Max Kölbel (2002, 2004, 2008). Yet, on my view, some of them pose serious problems for truth-relativism. Moreover, in addition to these general problems there is a specific one concerning truth-relativism in *ethics*: in contrast to linguistic practices within regions of discourse such as discourse about matters of taste our linguistic practices within *ethical* discourse do *not* provide evidence for a relativist semantics. So it is not only that the possibility of faultless disagreement is not the prevailing intuition, or the ordinary pre-theoretical view, that has to be explained in ethics by a semantic theory, but also that, according to our linguistic practices within ethics, which is characterized by a sort of objectivity, truth-relativism does not seem to provide the correct semantics for moral language (see also Coliva and Morruzzi 2012: 52).

In face of the general problems and the specific problem for truth-relativism in ethics, it may seem to be advisable for philosophers with relativistic inclinations to search for an alternative approach to relativism in ethics – and as an alternative it *nolens volens* must deny the *possibility of faultless disagreement*. However, the denial of this possibility seems to contradict the most basic idea of any interesting shape of relativism – by the following consideration. If faultless disagreement isn't possible, any disagreement between two thinkers A and B necessarily involves *some* mistake. Even if *both* A and B made a mistake, not both of their beliefs (contents) can be false since they constitute a contradiction. So just one of them can be false and the other has to be true – and any question has just one correct answer, viz. the content of the true belief.[1] So the big challenge is to reconcile the concession that

[1] This conclusion is compatible with *indexical relativism/contextualism* – in short, the view that the *content* of an assertion or belief is relative to a moral framework or the like – which is the more traditional approach to relativism in ethics (cf. Harman 1975, 1978; Harman 1996; and more recently Dreier 1990, 2006 and Wong 2006). For indexical relativists/contextualists accept that any ethical question has just one correct answer, since the question itself can only be sensibly asked *within* a moral framework – and the question as to which framework is the right one is supposed to be meaningless.

faultless disagreement is impossible, i.e. a denial of truth-relativism, with an alternative shape of relativism that deserves its name. In the remainder of this paper, I shall discuss the prospects of such reconciliation and an alternative shape of ethical relativism.

**2.** One thing should be clear: the intended reconciliation can only be achieved by *logical revision*, because the rules of classical logic inevitably lead us from the impossibility of faultless disagreement to the unwelcome conclusion that any disagreement involves some mistake and thus any ethical question has just one correct answer. Since the logical moves that yield this conclusion essentially involve Double Negation Elimination (DNE) – from the negation of Faultlessness to the conclusion that there is some fault in the disagreement – a restriction to intuitionistic or some equally weak logic that doesn't provide DNE precludes the fatal conclusion (cf. Wright 2002, 2006). Yet logical revision is just a *necessary* move; it can only be the first step on the road to relativism of the shape envisaged here, since a restriction to intuitionistic logic has just the power to *preclude* the logical transition from the impossibility of faultless disagreement to the fatal conclusion that any disagreement involves some mistake.[2] But blocking the critical step from the impossibility of faultless disagreement to a denial of relativism of the shape we are seeking arguably is considered by itself not yet relativism. It is just rebutting a disproof of relativism. In order to establish a relativist position we have to go further. So which might be the next step?

## 2. Truth as Superassertibility

**1.** The next step on the road to ethical relativism I shall discuss is a *substantial* account of truth that identifies ethical truth with an *epistemic* property, viz. what Wright (1992) termed *superassertibility*. This is intended as a *local* thesis, i.e. a thesis about truth in *ethics*. So relativism in ethics of the shape envisaged here does not entail relativism in other regions; and in order to avoid global relativism, we have to presuppose a framework called *alethic pluralism* (cf. Edwards 2011, 2012a, b; Pedersen 2010, 2012a, b; Lynch 2009; Wright 1992, 2001). This, in short, is the view that there is just one *concept* of truth – which is a minimal or metaphysically neutral concept governed by so-called platitudes or truisms – but there are possibly different properties of being true, different manifestations of it or different properties in virtue of which propositions are true, depending on the selected region of discourse. In more detail, alethic pluralism comprises the following theses (following Wright (1992: 24-36)).

> The concept of truth is a *minimal* concept. That is to say, the concept of truth is solely determined by a set of very general and intuitive principles, so-called *platitudes*, which are metaphysically neutral and connect truth to other concepts, e.g. the platitude that to assert is to present as true, that a belief is true just in case things are as they are believed to be, that it is true that P if and only if P, and so on. The platitudes build up an analytic theory, or a network analysis, of the concept of truth. Consequently, any correct substantial account of truth that proposes to define the *nature* of truth, or the *property* of being true, has to satisfy the platitudes.
>
> The *nature*, or *property*, of truth may be different depending on the selected region of discourse. That is, over and above the minimal features of truth expressed by the

---

Indexical relativism/contextualism, however, has problems of its own (cf. Boghossian 2006; Kölbel 2004b; MacFarlane MS, Wright 2002). So we are discussing the prospects for an alternative to truth-relativism *and* indexical relativism/contextualism.

[2] The restriction to intuitionistic logic within ethical discourse (and its meta-discourse) has to be motivated independently. It might suffice here to say that intuitionistic logic will turn out to suggest itself for the ethical domain, since it is congenial to the epistemic conception of ethical truth proposed in the following sections of this paper (see also Wright 1992: 42-44).

> platitudes there are extra features determining the *domain-specific* or *local* nature of truth in some targeted region of discourse, which may vary from region to region. There isn't *the* nature of truth across all areas, but possibly different natures depending on the selected region; and any such property is a model of the analytic theory provided it satisfies the platitudes.[3]

Hence, within a pluralist framework, truth in ethics may have a relative, evidentially constraint and mind-dependent nature, whereas truth of statements about physical objects, say, is, or is manifested by, an absolute, evidentially unconstraint and mind-independent property. In particular, truth in ethics may be a somehow idealized *epistemic* property that is constructed out of ordinary epistemic properties or norms such as assertibility, warrant or coherence.

**2.** There are two prominent accounts of ethical truth of this shape, which are closely related: Wright's conception of truth as *superassertibility* (Wright 1992, 2003, 2006) and Lynch's conception of truth as *supercoherence* or, more precisely, *concordance* (Lynch 2009). Superassertibility is assertibility not in some ideal, limiting state of information, but *stable* assertibility, i.e. the property of being assertible in some actually accessible state of information and remaining so no matter what enlargements or improvements are made to it. Supercoherence is not maximal (ideal) coherence, but the property of being coherent with some framework, e.g. moral framework, and remaining so no matter what improvements are made to the initial framework. Concordance in ethics is the property of being supercoherent with some moral framework combined with the truth of the morally relevant non-moral judgements of that framework (Lynch 2009: 176). In what follows, I'll confine myself to the idea of superassertibility, but nothing relevant for our discussion hinges on that choice instead of concordance. Wright introduces a range of concepts of superassertibility, not a single concept (cf. Wright 1992: 68), depending on the (perhaps domain-specific) interpretation of some key concepts or the range of the universal quantifier employed in the definition of superassertibility; and he presents the definition in slightly different versions (cf. Wright 1992: 48, 2003: 199, 2006: 56). For our purposes we distil the following working definition of superassertibility:

> **(S)** P is superassertible iff there is some accessible state of information S such that P is warrantable in S and remains so no matter what enlargements or other forms of improvement are made to S.

Notoriously, two main problems arise for this definition of superassertibility (see e.g. Edwards 1996: 105-108 and Wright 1992: 67-68): the first regards the question as to what is an *accessible state of information* and the second concerns the concept of *enlargement* or *improvement* (of a state of information) and the range of the corresponding states-of-information quantifier. The first problem has a straightforward solution for the ethical domain. A state of information, as mentioned in the definition of ethical superassertibility, can be conceived of as consisting of two components that are interwoven: first, a sufficiently coherent set of *moral beliefs* achievable by some thinker T at time t on the basis of her actual beliefs and, second, a set of morally relevant *non-moral*, notably *empirical information*

[3] The accounts of Wright and Lynch differ in a certain respect that is irrelevant for our discussion: for Wright there is one *concept* of truth but possibly different *properties* of being true such that "true" or "truth" may designate a different property when used in different regions of discourse, whereas, according to Lynch, there is just one generic *property* of being true, which is a *functional* property, but possibly multiple *manifestations* of it, i.e. different domain-specific properties in virtue of which propositions are true. There is also a further pluralist view called alethic disjunctivism (e.g. Edwards 2012b), according to which generic truth is the disjunctive property determined by all the domain-specific properties. For the following discussion, the crucial and common feature of all pluralist accounts is that for any predicate "T" that designates the domain-specific, truth-manifesting or truth-determining property the following holds good: P is true iff P is T.

available to T at t, i.e. a set of non-moral beliefs that would be, in the world as it actually is, generated in T at t by investigating in sufficiently favourable epistemic circumstances. The second problem, however, is more intricate. Yet here is what I think should suffice as a solution for the ethical domain. First, in definition (S), i.e. in the second half of the definiens, it is quantified over *all* enlargements and other forms of improvement of some state of information. So something has to be said about the range of this universal states-of-information quantifier. The following seems to be a plausible restriction: any enlarged or otherwise improved state of information, as any state of information mentioned or quantified over in the definition, is conceived of such that it is accessible to somebody at some time in the world as it actually is. That is to say, enlargements and other forms of improvement are conceived of as, in principle, achievable in this world by human beings. Moreover, we may think of the process of enlargement or improvement as an endless tree the root of which is the state of information available to somebody at some time and the knots are states of information such that any state on the tree is *better* than its predecessors. Since in what follows, I'll consider enlargement to be a special case of improvement, I'll simply talk of improved or better states of information. So, the crucial question is, when is a state of information *better* than another? It is natural to think of improving of one's state of information in terms of increasing one's knowledge or successively approaching the truth. However, as regards the *moral* part of information at least, i.e. the relevant set of moral beliefs, we cannot conceive of improvement as a process that, inter alia, enhances the moral *knowledge* or approaches the ethical *truth*, since then our characterization or grasp of the nature of *ethical* truth in terms of superassertibility would be defective, viz. *circular*: in order to grasp the nature of ethical truth in terms of superassertibility we already need to have a grasp of what constitutes the truth of a moral belief, i.e. a grasp of the nature of ethical truth. Moreover, employing the idea of approaching *the* truth runs counter to the idea of relativism. So in defining the concept of improvement of *moral* "information" we must avoid notions such as of knowledge or truth. A straightforward definition then might be given in terms of *increasing coherence* (cf. Lynch 2009: 172): improving *moral* information, i.e. a set of moral beliefs, is increasing its *coherence*.[4] As regards the *non-moral*, empirical part of information, however, we can make use of the concept of knowledge or truth, i.e. *empirical* truth, provided that empirical truth isn't identified with superassertibility. That is to say, the process of improvement of a state of information is the process of improving the morally relevant *non-moral*, *empirical knowledge*, i.e. acquiring *true* empirical information relevant for one's moral judgements and abandoning false empirical beliefs, while increasing the *coherence* of the overall state of information. Superassertibility thus conceived comes close to Lynch's concept of concordance (Lynch 2009: 176); in fact, concordance is a special case of superassertibility, which is neutral with respect to the question as to what constitutes ethical warrant or what type of epistemology in ethics should be adopted – coherentism or fundamentalism. If one takes, as I do, coherentism to be the preferable epistemology for ethics, ethical superassertibility as defined above *is* concordance.

**3.** So ethical truth may plausibly be conceived of as superassertibility thus defined. But how does relativism come into this epistemic picture of ethical truth? The straightforward answer, which also is the answer of the two main proponents of an epistemic conception of truth in ethics, Lynch (2009) and Wright (2006), runs as follows. To conceive of ethical truth as superassertibility is a form of ethical relativism since the following is a possibility: a proposition P is stably assertible for thinker A, i.e. warrantable in A's currently accessible state of information and remaining so no matter what improvements are made to it, and not-

[4] For the problem how to define coherence see Lynch 2009 (164-168) and his list of "coherence-making features" (167), which are: mutual explanatory support, predictive power, simplicity, completeness and consistency. So a framework grows in coherence iff, on balance, it shows more of these features or some of them to a greater degree (ibid.: 171).

P is stably assertible for another thinker B, i.e. warrantable in B's currently accessible state of information and remaining so no matter what improvements are made to it. That is to say, both P and not-P might be stably assertible relative to divergent starting points of states of information that build up the basis for A's or B's accepting P or not-P. This is the way Wright, at least in one of his papers (Wright 2006), characterizes the position he calls *True Relativism*: "It may also happen that some of the resulting enlarged states of information continue to warrant acceptance of P, and others acceptance of not-P. And once granted to be possible at all, it's difficult to see how to exclude the thought that such a situation might persist indefinitely. In that case superassertibility would be relative to a starting point, an initial basis for acceptance or rejection. [...] That would be a kind of relativity of truth" (Wright 2006: 57) Lynch concedes the same possibility for his conception of ethical truth as concordance: "You and I may have inconsistent but concordant judgements [...]." (Lynch 2009: 183) And in the footnote on the same page he writes, "Won't this entail that concordance is a relative notion? [...] Is it possible that the judgement that p might be concordant relative to one framework but not to another, and hence that moral truth is itself relative? It certainly raises this as a possibility." (Lynch 2009: 183, footnote 24) Thus the conception of truth as superassertibility, or concordance, seems to offer an interesting form of ethical relativism.

However, this proposal obviously has an perhaps unwelcome consequence (cf. Connolly 2012: 139): if ethical truth *is* superassertibility, which implies that P is true if and only if there is some accessible state of information S such that P is stably assertible based on S, it is possible that both P and not-P are *true*. And this, of course, is only possible – leaving dialetheism aside – if the *concept* of ethical truth, or the truth *predicate* applicable to ethical propositions, is *relativized*. Then, and only then, we are able to say that both P and not-P are true, i.e. that P is true relative to A's initial state of information and not-P is true relative to B's initial state of information. Yet a relativity of the *concept* of truth, or the truth *predicate*, is the very assumption we tried to avoid from the outset, since, in the face of the problems for truth-relativism pointed out above, we were looking for an *alternative* to truth-relativism. Moreover, *alethic pluralism* and its alleged *unity* of the *concept of truth* are ruled out when we relativize this concept in certain domains, e.g. in ethics, while retaining its absoluteness in others. So accepting a relativity of truth, i.e. the property of being true or the property in virtue of which true propositions are true, runs counter to alethic pluralism, at least as originally conceived of by Lynch and Wright. Hence we should not accept that both P and not-P might be stably assertible based on divergent starting points of information.

But doesn't the contrary seem to be a possibility? Wright puts the question thus: "[...] can this happen – that *in a single world* one thinker, Hero, is in a position to accept P, and another, Heroine, is in a position to accept not-P, and that each can retain their respective situations no matter what improvements or enlargements are made to their states of information?" (Wright 2006: 56) And even if it is granted that their respective bodies of information allow of *pooling*, his answer is affirmative: "When Hero and Heroine bring their respective bodies of information together, it may be that there is more than one equally rationally defensible way for accommodating the components into a unified state, [...]." (Wright 2006: 56-57) As already mentioned, Lynch (2009: 183) admits the same possibility. So it is difficult to see how to exclude the thought that there may be a moral conflict between two incompatible moral frameworks that remain incompatible no matter how much their coherence is increased and no matter how much non-moral knowledge is acquired. Thus it seems hard to deny the possibility that both P and not-P are such that they are warrantable in divergent states of information and remain so no matter what improvements are made to them. Hence, within our epistemic conception of truth as superassertibility, the possibility that both P and not-P are true cannot be ruled out *a priori* – or so it seems.

However, a proponent of that approach to truth, who seeks for an alternative to semantic truth-relativism, has to *deny* the possibility that both P and not-P are stably assertible based on divergent states of information – maybe justified on grounds of somewhat reconsidering the notion of improvement as employed in our definition of superassertibility, e.g. by widening the range of the general states-of-information quantifier to all (metaphysically) possible improvements instead of humanly possible ones, or on grounds of considerations that are specific for the ethical domain. I cannot dwell on this problem here, but simply assume that there is room for manoeuvre and some way or other to make plausible the denial of the possibility in question.[5]

So let's grant that there is some *a priori* reason to think that no ethical proposition P could be such that both P and not-P are stably warrantable based on divergent states of information. But is this not the end of our relativist story? For, if it is not possible that both P and not-P are superassertible, then necessarily it is not the case that both P and not-P are superassertible, which implies that necessarily P is not superassertible or not-P is not superassertible. And if one of the two propositions is not superassertible, the other is superassertible (since not both can be not superassertible). So P is superassertible or not-P is – again the fatal conclusion. Yet, fortunately, we made essential use of classical logic, viz. DNE in the transition from "Not-P is not superassertible" to "P is superassertible". By intuitionistic rules alone we are solely allowed to conclude: if one of the propositions, e.g. not-P, is not superassertible, then *it is not the case* that the other, viz. P, is *not* superassertible. In other words, the assumption that one of the propositions is not superassertible solely implies that we are *not justified* to suppose that the other proposition is *not* superassertible, which is not to say that we are *justified* to suppose that the other proposition *is* superassertible. Only by applying DNE we arrive at the unwelcome conclusion that if one of the two propositions is not superassertible then the other is. So, within an intuitionistic framework, we are in a position to block the above train of thought that resulted in the fatal conclusion that any ethical question has just one correct answer – and a shape of relativism envisaged here is still a lively option.

However, up to now, it is just that: an *option*. For by adopting an epistemic account of truth and a restriction to intuitionistic logic we again solely *averted a refutation* of the envisaged shape of relativism. We just precluded that our targeted position implies an anti-relativist conclusion. Arguably this isn't yet relativism by itself.

## 3. Anti-realist Epistemic Relativism

It still remains the task to show that by adding a suitable extra ingredient to our epistemic approach to ethical truth conjoint with a restriction to intuitionistic logic we really achieve a satisfactory shape of ethical relativism. So, how could relativism come into our picture of truth and logic in ethics if the possibility is denied that both P and not-P are stably assertible based on divergent states of information? I think that *epistemic* relativism, i.e. something like the idea that there may be intractable ethical disputes no matter what degree of coherence the moral frameworks involved show and no matter what amount of empirical knowledge is gained, is the only candidate for the missing extra ingredient. So how exactly is epistemic relativism to be defined and does it, when properly conceived, really help with achieving a satisfactory shape of ethical relativism?

A straightforward approach to defining epistemic relativism parallels the definition of relativism as mentioned at the beginning of the paper: relativism defined in terms of *faultless disagreement*. Epistemic relativism then is the thesis that *epistemically faultless disagreement* is possible. It is the claim that there may be disagreement such that neither of

[5] For considerations to the effect that this possibility cannot be ruled out by Lynch's conception of ethical truth as concordance see Connolly 2012 (pp. 138-144).

the parties involved has made an *epistemic* mistake, i.e. showed an insufficient or imperfect cognitive *procedure* that generated her belief in question, and not just plain error in the end product, i.e. false belief. However, epistemic relativism thus defined is *not* possible, which is shown, even within intuitionistic limits, by the following deduction:[6]

| | | |
|---|---|---|
| (1) | A believes that P and B believes that not-P | Disagreement |
| (2) | Neither A nor B has made an epistemic mistake | Faultlessness |
| (3) | ∀X: If X does not believe that P, and P is stably assertible for X, X has made an epistemic mistake | Error Principle |
| (4) | Not (A does not believe that not-P, and not-P is stably assertible for A) | (2), (3), MTT |
| (5) | Not (B does not believe that P, and P is stably assertible for B) | (2), (3), MTT |
| (6) | A does not believe that not-P | (1), (2) |
| (7) | B does not believe that P | (1), (2) |
| (8) | Not-P is stably assertible for A | (4), (6), MPT |
| (9) | P is stably assertible for B | (5), (7), MPT |
| (10) | Not-P is superassertible | (8), def. (S) |
| (11) | P is superassertible | (9), def. (S) |
| (12) | Not-P | (10), ES |
| (13) | P | (11), ES |

So we have to deny the possibility of epistemically faultless disagreement, if truth is superassertibility (and so the Equivalence Schema, ES, is accepted for superassertibility, which is used for deriving the last two lines). We must propose an alternative conception of epistemic relativism that in conjunction with our epistemic approach to truth and the restriction to intuitionistic logic may yield a satisfactory shape of ethical relativism.

**4.** Apart from the possibility of epistemically faultless disagreement there is another quite evident way how to conceive of epistemic relativism, namely one in terms of *rationally irresolvable disagreement*. The thought is that there may be intractable disputes that *cannot be resolved by rational means*. Of course, this must not be understood as disagreement in which no rational shortcoming is involved, since this would amount to nothing but epistemically faultless disagreement. Rather, it is, or may plausibly be regarded as, disagreement that is irresolvable by means of *improving the respective states of information involved in the disagreement*. Since arguably improving one's state of information, i.e. increasing its coherence and enhancing one's empirical knowledge, is always a rational endeavour, any improvement of some accessible state of information is achievable by rational means. So a rationally irresolvable disagreement may be defined as such that it cannot be resolved no matter what improvements are made to the states of information involved in it. Hence, first, we define rational resolvability and, then, take rational irresolvability to be its denial. That one, and only one, of the two contradictory beliefs involved in a disagreement remains warrantable seems to be necessary for its rational resolvability: a conflict is rationally

[6] Compare Wright's EC-deduction to the same effect (Wright 2002: 109). It works for discourses, such as arguably ethical discourse, in which it is a priori that truth is *evidentially constrained*, i.e. in which any truth, in principle, is feasible to know. The Error Principle used in the EC-deduction says that it involves a cognitive shortcoming in a procedural sense, i.e. an epistemic mistake, to believe the negation of something that is feasible to know.

resolvable only if one, and only one, of the beliefs is stably assertible based on the respective state of information – and this belief, or its content, would be the resolution of the dispute. So we may define rational resolvability of a disagreement constituted by A's belief that P and B's belief that not-P thus:

**(RR)** The disagreement between A and B *is rationally resolvable* iff either P is stably warrantable based on A's state of information, $S_A$, or not-P is stably warrantable based on B's state of information, $S_B$, but not both propositions are stably warrantable based on the respective states of information $S_A$ or $S_B$.

Now, we define *rational irresolvability* by denying (RR):

**(RI)** The disagreement between A and B is *rationally irresolvable* iff it is not the case that either P is stably warrantable based on $S_A$ or not-P is stably warrantable based on $S_B$ and not both propositions are stably warrantable based on $S_A$ or $S_B$.

So we may define epistemic relativism in terms of the possibility of rationally irresolvable disagreement, i.e. in terms of conflicts that are not resolvable by means of improving the respective states of information involved in the conflict. However, one might object that the idea of rational resolvability or irresolvability is not adequately captured by (RR) or (RI): rational resolvability of a conflict does demand less than resolvability by means of improving the states of information *involved in the disagreement*. The resolvability of the conflict between A and B is not necessarily resolvability *by A and B*, but possibly resolvability solely by some third thinker C, who is in a better epistemic position than A and B are in or may be in after improving their states of information. In other words, if there is a state of information outside the range of improvements of $S_A$ and $S_B$, viz. a state of information $S_C$ accessible to C, in which the question as to whether or not P is *decidable*, so that *either* P is warrantable in $S_C$ *or* not-P is warrantable in $S_C$ (but not both are warrantable in $S_C$), and this decidability of P in $S_C$ is *stable* no matter what improvements are made to $S_C$, then the conflict between A and B arguably is rationally resolvable, though possibly not resolvable by *their* rational capacities alone, i.e. by means of improving *their* states of information, but by means of a third (better) state of information and all of its improvements. And since if a disagreement about whether or not P is rationally resolvable there is *some* state of information based on which P is stably decidable, we get the following definition of rational resolvability:

**(RR*)** A disagreement about whether or not P *is rationally resolvable* iff there is a state of information S such that P is stably decidable based on S, i.e. if *either* P is stably warrantable based on S *or* not-P is stably warrantable based on S but not both are.

Let's call P *superdecidable* just in case there is a state of information S such that P is stably decidable based on S. For rational *irresolvability* we then get the following definition:

**(RI*)** A disagreement about whether or not P is *rationally irresolvable* iff P is not superdecidable, i.e. P is superundecidable: there is no state of information S such that P is stably decidable based on S.

However, superundecidability[7] might still be insufficient for rational irresolvability. For suppose that there is a state of information S* in which one would be warranted in believing that P is superundecidable, then it is quite plausible to say that the conflict about P is in a sense resolved in S*: the answer to the question whether or not P is just that P is superundecidable. That is, the solution of the conflict about P is the insight that such a solution is (rationally) impossible. And if this *solution of no solution* survives all further

[7] Superundecidability is conceived of as the *negation* of superdecidability, i.e. the idea that there is some state of information based on which P is stably decidable. Superundecidability is not constructed out of *stable undecidability* based on a state of information, viz. the idea that *there is a* state of information S such that P is *undecidable* in S and remains so no matter what improvements are made to S.

improvement of S*, i.e. if a warrant for the belief that P is superundecidable survives all improvements made to S*, the disagreement is in this sense stably resolvable based on S*. Hence there is then a *higher-order* rational resolvability.[8] So we may demand of rational irresolvability that there is no such state of information S* or such improvements of it, and improve our definition accordingly:

**(RI+)** A disagreement about whether or not P is *rationally irresolvable* iff P is superundecidable and it is superundecidable whether it is so, i.e. there is no state of information S such that the belief that P is superundecidable is stably decidable based on S.

Obviously, rational irresolvability thus defined is closely related to an epistemic situation that Wright termed a *Quandary*: a situation in which there is uncertainty through and through (Wright 2002: 112-115). More precisely, the idea is, a proposition P presents a Quandary for a thinker T just in case the following conditions are met (Wright 2002: 113):

(i) T does not know whether or not P

(ii) T does not know of any way of knowing whether or not P

(iii) T does not know that there is any way of knowing whether or not P

(iv) T does not know that it is (metaphysically) possible to know whether or not P

but this condition is not met:

(v) T knows that it is (metaphysically) impossible to know whether or not P.

A Quandary is always a Quandary for somebody at some time and the state of information available for her at that time. So we may replace "T does not know ..." and "T knows ..." by "... is not stably warrantable based on T's state of information" and "... is stably warrantable based on T's state of information" in the above conditions (i) to (v), and accordingly define a slightly modified concept of a Quandary – a *Quandary**. Quandary* thus defined is undecidability through and through, with undecidability of P in the abovementioned sense: neither P nor not-P is warrantable in some given state of information. Then we define a *Stable Quandary** for T as follows: P presents a Stable Quandary* for T if and only if P presents a Quandary* for T and this remains so no matter what improvements are made to T's initial state of information. Finally, we define a Superquandary*: P presents a Superquandary* just in case P presents a Stable Quandary* for *any* thinker who would deliberate on P, which is tantamount to say that P is superundecidable and it is superundecidable whether this is so, i.e. rationally irresolvable in the sense of (RI*). Now we are in a position to define *epistemic relativism* as the following thesis:

**(ER)** There may be some proposition P such that P presents a Superquandary*.

So this is the overall package of ethical relativism proposed here:

(a) Conservatism as regards the concept of truth: retaining a *monadic truth predicate*.

(b) Logical revision for ethical discourse: restriction to *intuitionistic logic*.

(c) Epistemic conception of ethical truth: *alethic pluralism*, truth as *superassertibility*.

(d) Epistemic relativism for the ethical domain: *Superquandary** is possible, i.e. *(ER)*.

[8] Think of undecidable mathematical propositions that are proven to be so, e.g. the continuum hypothesis: although the *mathematical* problem, i.e. the problem as to whether or not the continuum hypothesis is true, is thereby not solved, there is meta-mathematical certainty, and any further dispute would be irrational – the dispute is quiesced and in this sense rationally solved.

**5.** In concluding the paper, I'll briefly discuss the question as to whether the relativist proposal constituted by features (a) to (d) really builds up a relativist position in ethics that deserves its name. For, prima facie, it may be objected that, after all, the proposed shape of epistemic relativism is merely that: *epistemic* relativism – and as such an innocent and far too weak position to be properly called relativist. Almost nobody thoroughly concerned with such meta-ethical issues would deny that there might be ethical propositions, maybe e.g. those presented by moral dilemmas, that are stably undecidable through and through for anybody deliberating on it. However, this objection would be serious if, and only if, epistemic relativism would be combined, or combinable, with (the possibility of) *rampant realism*, i.e. the thesis that for any P there is *a matter of fact* about whether or not P, a fact that is represented by a true belief, though possibly a fact that we might be unable to discover. For epistemic relativism conjoint with (the possibility of) rampant realism indeed would be an innocent shape of ethical relativism. Yet, since our epistemic conception of ethical truth conjoint with logical revision entails a kind of *anti-realism*, a proponent of the shape of ethical relativism proposed here – let's call it *Anti-realist Epistemic Relativism* – is in a position to deny rampant realism and thus to rebut the objection. But, again, is it really what we were seeking: ethical *relativism*? How could one tell? After all, when does a position deserve to be called *relativism*? Beyond our minimal constraint – which demands that the account *must not entail* that any ethical question has just one correct answer, by which epistemic relativism conjoint with rampant realism is ruled out as a kind of relativism – there seems to be no general, unprejudiced criterion for relativism. Certainly, if one simply *defines genuine* or *true* relativism as relativism about *truth* – "One is only a relativist if one takes the accuracy of some assertions or beliefs to vary with the context from which they are assessed (the 'context of assessment')" (MacFarlane 2011: 443-444) – or, what amounts to the same thing, as the possibility of faultless disagreement, then our proposal is not *genuine* or *true* relativism. But once we abstain from that preconception the proposal made in this paper may be an interesting approach to ethical relativism. For the question as to whether it *really* is relativism doesn't make sense. The discussion, I think, has shown that this is merely a terminological issue.

Be that as it may, in this paper, I tried to show that for the ethical domain Anti-realist Epistemic Relativism is at least a plausible position and, after all, the *best* we can get as an *alternative* to the other forms of relativism, notably (semantic) truth-relativism. Whether it is good enough to meet one's relativistic demands and suffices to embrace one's relativistic intuitions – that's a question anybody has to answer for herself. Anyway, Anti-realist Epistemic Relativism seems to be a position attractive for those with relativistic inclinations who are nevertheless sceptical as regards the prospects of truth-relativism and contextualism, let alone expressivism and related positions that deny the truth-evaluability of ethical judgements.

**Manfred Harth**

Ludwig-Maximilians-Universität München
Manfred.Harth@lrz.uni-muenchen.de

## References

Binderup, Lars 2008: 'Brogaard's moral contextualism', *The Philosophical Quarterly* 58, 40–415.

Boghossian, Paul 2006: 'What is relativism?' in P. Greenough and M. Lynch (eds.) 2006, 13–37.

Capps, Lynch and Massey 2009: 'A coherent moral relativism', *Synthese* 166, 413–430.

Coliva and Morruzzi 2012: 'Truth Relativists Can't Trump Moral Progress', *Analytic Philosophy* 53, 48–57.

Dreier, James 1990: 'Internalism and Speaker Relativism', *Ethics* 101, 6–26.

Dreier, J. 2006: 'Moral Relativism and Moral Nihilism', in D. Copp (ed.) *The Oxford Handbook of Ethical Theory*. Oxford: OUP, 240–264.

Edwards, Jim 1996: 'Anti-realist Truth and Concepts of Superassertibility', *Synthese* 109, 103–120.

Edwards, Douglas 2011: 'Simplifying Alethic Pluralism', *The Southern Journal of Philosophy* 49, 28–48.

Edwards, Douglas 2012a: 'Alethic vs. Deflationary Functionalism' *International Journal of Philosophical Studies* 20, 115–124.

Edwards, Douglas 2012b: 'On Alethic Disjunctivism', *Dialectica* 66, 200–214.

Egan, Andy 2007: 'Epistemic Modals, Relativism and Assertion', *Philosophical Studies* 133, 1–22.

Egan, Andy 2009: 'Billboards, Bombs, and Shotgun Weddings', *Synthese* 166, 251–279.

Egan, Andy 2010: 'Disputing about Taste', in R. Feldman and T. Warfield (eds.), *Disagreement*, Oxford: OUP, 247–286.

Evans, Gareth 1985: *Collected Papers*. Oxford: OUP.

Francén, Ragnar 2010: 'No Deep Disagreement for New Relativists', *Philosophical Studies* 151, 19–37.

García-Carpintero, Manuel and Kölbel, Max (eds.) 2008: *Relative Truth*. Oxford: OUP.

García-Carpintero 2008: 'Relativism, Vagueness and What Is Said', in M. García-Carpintero and M. Kölbel (eds.) 2008, 129–154.

Greenough, Peter and Lynch, Michael (eds.) 2006: *Truth and Realism*. Oxford: OUP.

Harman, Gilbert 1975: 'Moral Relativism Defended', *Philosophical Review* 84, 3–22.

Harman, G. 1978: 'What Is Moral Relativism?' in A. Goldman and J. Kim (eds.) *Values and Morals*. Dordrecht: Reidel, 143–161.

Harman, G. 1996: 'Moral Relativism' in G. Harman and J. Thomson 1996, 3–64.

Harman, G. and Thomson, J. 1996: *Moral Relativism and Moral Objectivity*. Oxford: OUP.

Harth, Manfred 2013a: 'Is Relative Truth Really Truth?', manuscript.

Harth, M. 2013b: 'Relative Truth, Lost Disagreement and Faultless Contradiction', manuscript.

Kölbel, Max 2002: *Truth without Objectivity*. London.

Kölbel, M. 2004a: 'Faultless Disagreement', *Proceedings of the Aristotelian Society* 104, 53–73.

Kölbel, M. 2004b: 'Indexical Relativism versus Genuine Relativism', *International Journal of Philosophical Studies* 12, 297–313.

Kölbel, M. 2007: 'How to Spell out Genuine Relativism and How to Defend Indexical Relativism', *International Journal of Philosophical Studies* 15, 281–288.

Kölbel, M. 2009: 'The evidence for relativism', *Synthese* 166, 375–395.

Lasersohn, Peter 2005: 'Context Dependence, Disagreement, and Predicates of Personal Taste', *Linguistics and Philosophy* 28, 634–686.

Lasersohn, P. 2009: 'Relative Truth, Speaker Commitment, and Control of Implicit Arguments', *Synthese* 166, 359–374.

Lynch, Michael 2009: *Truth as One and Many*. Oxford. OUP.

MacFarlane, John 2003: 'Future Contingents and Relative Truth', *The Philosophical Quarterly* 53, 321–336.

MacFarlane, John 2005: 'Making Sense of Relative Truth', *Proceedings of the Aristotelian Society* 105, 321–339.

MacFarlane, John 2007: 'Relativism and Disagreement', *Philosophical Studies* 132, 17–31.

MacFarlane, John 2011: 'Simplicity Made Difficult' *Philosophical Studies* 156, 441–448.

MacFarlane, John MS: *Assessment Sensitivity*. Online manuscript, draft of May 10, 2012.

Moruzzi, Sebastiano 2008: 'Assertion, Belief and Disagreement: A Problem for Truth-Relativism', in M. García-Carpintero and M. Kölbel (eds.) 2008, 207–224.

Pedersen, Nikolaj 2010: 'Stabilizing Alethic Pluralism', *The Philosophical Quarterly* 60: 92–108.

Pedersen, N. 2012a: 'True Alethic Functionalism?', *International Journal of Philosophical Studies* 20: 125–133.

Pedersen, N, 2012b: 'Recent Work on Alethic Pluralism', *Analysis* 72: 588–607.

Rosenkranz, Sven 2008: 'Frege, Relativism and Faultless Disagreement', in M. García-Carpintero and M. Kölbel (eds.) 2008, 225–237.

Stojanovic, Isidora 2007: 'Talking about Taste: Disagreement, Implicit Arguments and Relative Truth', *Linguistics and Philosophy* 30, 691–706.

Wong 2006: *Natural Moralities.* Oxford. OUP.

Wright, Crispin 1992: *Truth and Objectivity*. Cambridge. HUP.

Wright, Crispin 2001: 'Minimalism, Deflationism, Pragmatism, Pluralism', in M. Lynch (ed.) *The Nature of Truth*, 751–787.

Wright, C. 2002: 'Relativism and Classical Logic', in A. O'Hear (ed.) *Logic, Thought and Language*. Cambridge, 95–118.

Wright, C. 2003: 'Truth in Ethics', in *Saving the Differences*. Cambridge: HUP, 183–203.

Wright, C. 2006: 'Intuitionism, Realism, Relativism and Rhubarb', in P. Greenough and M. Lynch (eds.) 2006, 38–60.

# Has Vagueness Really No Function in Law?

David Lanius

When the United States Supreme Court used the expression "with all deliberate speed" in the case *Brown v. Board of Education*, it did so presumably because of its vagueness. Many jurists, economists, linguists, and philosophers accordingly assume that vagueness can be strategically used to one's advantage. Roy Sorensen has cast doubt on this assumption by strictly differentiating between vagueness and generality. Indeed, most arguments for the value of vagueness go through only when vagueness is confused with generality. Sorensen claims that vagueness – correctly understood – has no function in law *inter alia* because judges lie systematically when confronted with borderline cases. I argue that both claims are wrong. First, judges do not need to resort to lying when adjudicating borderline cases, and even if they had to, this would not render vagueness useless. Secondly, vagueness has several important functions in law such as the reduction of decision costs and the delegation of power. Although many functions commonly attributed to the vagueness of legal expressions are in fact due to their generality or other semantic properties, vagueness has at least these two functions in law.

## 1. What Is the Problem?

It is widely believed that vagueness plays an important role in law. Many jurists, economists, linguists, and philosophers claim that vagueness can be strategically used to one's advantage. So called strategic vagueness has been studied generally in communication (cf. Franke 2011 and de Jaegher 2011), but also specifically in contracts, verdicts and statutes (cf. Staton 2008 and Choi 2011).

However, it seems that in most (or even all) cases the utility of vague expressions is due to generality rather than vagueness. Roy Sorensen holds that no advantage whatsoever can be gained by using vague expressions due to their vagueness (cf. Sorensen 2001). Indeed, many arguments for the use of vagueness are convincing only if vagueness is (mis-)understood as generality. Thus, if one wants to show that vagueness – correctly understood – has a function, one has to make sure that the effects commonly attributed to vagueness are not in fact due to something else.

The primary aim of this paper is to show that there are cases in which we cannot attribute the intended positive effects of vague expressions to anything else than their vagueness. Also – as a secondary aim – I will argue that judges are not forced to lie because of vagueness.

## 2. What Is Vagueness?

Vagueness is usually characterized by three criteria (cf. Keefe 2000):

(C1) Borderline Cases

(C2) Fuzzy Boundaries

(C3) Sorites Susceptibility

For this paper's purposes, however, we will bracket out criteria (C2) und (C3) and focus on criterion (C1), since only allowing for borderline cases is universally accepted as a necessary

condition for an expression being vague. For instance, Paul Grice influentially defined "vagueness" by way of reference to uncertainty in borderline cases:

> To say that an expression is vague [is] to say that there are cases (actual and possible) in which one just does not know whether to apply the expression or to withhold it, and one's not knowing is not due to ignorance of the facts. (Grice 1989: 177)

Thus, we can say that an expression is vague if it admits borderline cases, where borderline cases are cases in which the expression neither clearly applies nor clearly does not apply. The most common examples of vague expressions given in philosophical contexts are "heap," "bald," "red," and in more legal contexts "vehicle" and "reasonable." An expression which is not vague is precise.

So far, everything said about vagueness is fairly uncontroversial. However, several theories of vagueness have been proposed none of which gained any reasonably wide acceptance in the philosophical community to count as the mainstream view. One theory of vagueness though managed to gain exceptionally little acceptance – this is the epistemic theory of vagueness defended by Roy Sorensen, which claims that vagueness is ignorance and that all borderline x of F are in fact either F or ¬F, but we cannot know. My objections against Sorensen's arguments are independent of any particular theory of vagueness, but his arguments can only properly be understood against the background of his epistemicism.

In contrast to vagueness, we can define generality by saying that an expression is general if it can be applied to a wide variety of cases, where the precise scope of "wide" is relative to some context or purpose. For instance, one could say that the expression "person" is general because it covers children, biology students, retirees, pianists, millionaires and every Tom, Dick and Harry. The opposite of generality is specificity.

In a nutshell, then, one can say that while vagueness is an expression's property of allowing for borderline cases, generality is an expression's property of applying to a wide variety of cases. Consequently, the properties of vagueness and generality are logically independent, that is to say that there are general and precise expressions as well as specific and vague ones. However, most ordinary language expressions are both vague and general, which presumably is one reason why people tend to confuse vagueness and generality so frequently and persistently.

## 3. What Are Borderline Cases?

We said that borderline cases are cases in which a vague expression neither clearly applies nor clearly does not apply. Now, we can distinguish between two kinds of borderline cases. Philosophers are usually interested only in the *possibility* of borderline cases. If some case is conceivable in which the expression in question neither clearly applies nor clearly does not apply, then the expression is vague.

In law, however, we generally want to minimise the occurrence of *actual* borderline cases. It is notable that only then it is meaningful at all to talk about more or less vague expressions. All vague expressions arguably admit indefinitely many possible borderline cases,[1] but they evidently differ considerably with respect to the number of actual borderline cases – real cases which in fact (in the actual world) are unclear. So, an expression can be more or less vague with respect to the number of actual borderline cases it generates, but not with respect to the number of possible borderline cases.

[1] Potential exceptions are expressions like "small natural number," which allow only a limited number of possible borderline cases.

Accordingly, we can distinguish between to kinds of vagueness:[2]

> An expression is **intensionally vague** if it possibly admits borderline cases.
>
> An expression is **extensionally vague** if it actually admits borderline cases.

There is a second very important distinction. According to Sorensen, some borderline cases depend on our epistemological access and can be resolved. These are relative borderline cases. Absolute borderline cases, in contrast, cannot be resolved and give rise to vagueness.

A judge can, for instance, sort all the documents, which are clearly relevant, on one pile, all the documents, which are clearly not relevant, on a second pile, and then leaving a pile of documents with all the unclear cases in the middle. First, she might continue to find one or another relative borderline case in the middle, which can eventually be decided. However, those cases that will remain in the middle after exhausting all epistemological means of the judge are absolute borderline cases.

Relative and absolute borderline cases can be defined as follows:

> **Absolute Borderline Cases**
>
> An individual x is an absolute borderline F iff given any means of answering "Is x an F?" x is borderline.
>
> **Relative Borderline Cases**
>
> An individual x is a relative borderline F iff x is borderline, but can be identified either as F or as not F given some answering resource.

This distinction is crucial for Sorensen, since he maintains that in law relative borderline cases have a function, while absolute borderline cases do not. He has the view that vague expressions can be useful because they – despite their vagueness – allow for (relative borderline) cases that are initially unclear, but can later be resolved in one way only.

In any case, based on Sorensen's distinction, we have to adjust our definition of vagueness such that an expression is vague if it possibly admits *absolute* borderline cases.

## 3. Do Judges Necessarily Lie?

Sorensen not only claims that absolute borderline cases are never useful, they also force judges to make groundless decisions and eventually to lie:

> Judges cannot discover the correct answer to a question about an absolute borderline case because no one can. [...] The judge is not permitted just to confess his ignorance; the judge is obliged to answer. Therefore, he is obliged to answer insincerely. (Sorensen 2001: 400)

According to Sorensen, judges are lying when asserting verdicts on absolute borderline cases, since they cannot be justified in believing the truth of their assertions. Judges must decide cases that are brought before them, some of which are absolute borderline cases. Hence, judges must decide absolute borderline cases, which are by definition *undecidable*.

---

[2] This distinction is due to Kit Fine, who based it on Rudolf Carnap's differentiation between an expression's extension and its intension: "A predicate *F* is extensionally vague if it has borderline cases, intensionally vague if it could have borderline cases." (Fine 1975: 266) Consequently, extensionally vague expressions allow cases that neither clearly belong to their extension nor clearly do not. Intensionally vague expressions, on the other hand, allow cases that neither clearly belong to their intension nor clearly do not.

I will now give a semi-formal reconstruction of his argument consisting of two parts. The first part establishes that necessarily judges cannot know statements about absolute borderline cases:

(P1) x is an absolute borderline F.

(P2) Nobody can know F(x) if x is an absolute borderline F.

(K1) Thus, judge J cannot know that F(x). (from (P1)-(P2))

This seems to be rather unproblematic. Nobody can know whether a statement about an absolute borderline case is true or false, not even a highly qualified judge. This is also plausible if one does not submit to Sorensen's background assumption that vagueness is ignorance, that is, to his epistemicism.

The second more problematic part of the argument seeks to establish that judges are necessarily insincere in absolute borderline cases:

(P3) Judge J's verdict is an assertion that F(x).

(P4) Judge J does not believe that F(x). (from (K1))

(P5) Judge J has the intention that some X ≠ J shall be led to believe that F(x).

(P6) Some Y lies with respect to F(x) if Y asserts that F(x), while not believing that F(x), with the intention that some X ≠ Y shall be led to believe that F(x).

(K2) Thus, judge J lies with respect to F(x). (from (P3)-(P6))

If we accept premises (P3) to (P6), we have to accept Sorensen's claim that judges necessarily lie when adjudicating absolute borderline cases. Thus, if we don't accept the conclusion, we have to dismiss one of the premises. Are all premises (P3) to (P6) plausible?

One might argue that premise (P4) is not entailed by conclusion (K1), since one can believe something without being able to know it. However, a judge who finds out that she cannot know the verdict should not believe it. So, premise (P4) is entailed by conclusion (K1) if it is assumed that the judge has a relative high standard of belief justification with respect to her verdict. It is not necessary to assume that the judge needs to actually know that she cannot know the verdict. The judge must only try hard enough and fail, in order to establish that she is not justified to believe it. Hence, the judge would not believe something she has grounds to think she does not know:

(K1) Judge J cannot know that F(x).

(S1) Judge J does not know that F(x).

(S2) Judge J is not justified to believe that F(x).

(P4) Judge J does not believe that F(x).

Thus, we should accept premise (P4) because judges have a particular high standard of belief justification. Of course, there are actual judges who neglect this standard; probably a lot of judges believe all sorts of things. But, one can hardly say that they do it systematically.

Maybe, then, the definition of lying which Sorensen uses is inadequate. In this case, premise (P6) should be rejected because having the intention to make someone else believe something one doesn't know is not sufficient for lying. Perhaps lying necessarily involves intentional deception. However, the argument can be restated with different premises (P5*) and (P6*) just as well such that:

(P5*) Judge J has the intention to deceive some X ≠ J such that X shall believe that F(x).

(P6*) Some Y lies with respect to F(x) if Y asserts that F(x), while not believing that F(x), with the intention to deceive some X ≠ Y such that X shall believe that F(x).

Then (K2) would still be entailed by the premises and the judge's verdict would count as lying. Yet a different definition of lying has been proposed by Jason Glenn:

> [W]hen one lies one deliberately states a negation of what they know (or think they know) to be the case. (Glenn 2007)

So understood, conclusion (K2) cannot be derived because in an absolute borderline case the judge does not come to knowledge at all. Consequently, she does not state its negation. However, this change in definition seems to suggest that some kind of insincerity on the part of the judge remains, and Glenn concedes that judges *bullshit* when adjudicating absolute borderline cases.[3]

This is a possible objection to Sorensen's argument. On one hand, though, this definition of lying is rather restrictive, ruling out cases in which one is asserting something one has good reason not to believe in order to deceive somebody else. Even if one might normally call this lying, the distinction to bullshitting is a useful one and perhaps one should change how one uses the expression "lying." On the other hand, we then would have to agree that judges systematically bullshit in absolute borderline cases which is only slightly more plausible than having them lying.

Luckily, there is an alternative; namely, one can reject premise (P3) and argue that judicial verdicts are not mere assertions of facts of the matter. They are surely speech acts asserting what the law in a particular case is, but they are also (equally surely) speech acts ascribing legal character to the facts found, declaring the institutional fact of guilt, and (in hard cases) creating new law (cf. Bernal 2007). Thus, verdicts are not true or false in the same way as descriptions or claims about the world are.[4]

In hard cases judicial verdicts are often both assertions of what the law is and advancements to existing law. Even though judge-made law is a controversial phenomenon (especially in Germany), judges are frequently quasi-lawmakers and their verdicts are not directly subject to truth or falsity. What the German Federal Constitutional Court or the United States Supreme Court decides is not only final in the sense that it cannot be appealed to, but also generally binding for future cases.

But also in clear cases, a judge's ruling is not just a literal application of the law. It is argumentation *inter alia* about which aspects of an expression's meaning are relevant in a particular case.[5] But this is neither fixed by semantics nor pragmatics. Judges can justify their decisions by bringing in linguistic and extra-linguistic as well as legal and extra-legal arguments. Vague legal expressions are highly multi-dimensional and it is not determinate which dimension prevails in a particular context.[6] Taking into account possible intentions of

---

[3] Glenn follows Harry Frankfurt in his definition of "bullshitting:" "When one is bullshitting, the only thing that one deceives another about is one's attitude towards the mode of presentation of one's claims about a certain subject matter, and not the subject matter itself." (Glenn 2007; cf. Frankfurt 2005)

[4] This does not mean that verdicts are a kind of assertion without truth-value – as for instance Woozley suggested (cf. Woozley 1979: 30).

[5] Of course, legal expressions can attain quite different meanings than the ones of their everyday counterparts (cf. Poscher 2009). A precise threshold for legal age can be introduced to reduce absolute borderline cases. Because this is stipulated by the lawmaker, it might get in conflict with our vague everyday concept of maturity. Consequently, natural language and legal expressions often diverge in meaning.

[6] For instance, the expression "heap" is often considered to be one-dimensional; its application depends on the number of grains. But actually it is at least two-dimensional, as it depends also on their arrangement. Most legal expressions have indeterminately many dimensions.

the lawmaker or the purpose of the law can enable judges to honestly believe their verdicts to be justified, even when facing absolute borderline cases.

When adjudicating both clear and hard cases, judges usually give all relevant reasons of why a particular decision was reached. Insofar as there is any systematic deception when absolute borderline cases are concerned, a judge might try to implicate that her and only her decision was required by the law (cf. Bix 2003: 105). Hence, judges might decide absolute borderline cases relying on reasons which they actually believe to be conclusive, but objectively are not.

Based on similar considerations, Scott Altman argues that judges should be candid but not introspective:

> By candid, I mean never being consciously duplicitous. Candid opinions do not offer reasons judges know do not persuade them. By introspective, I mean critically examining one's mental states to avoid any self-deception or error. (cf. Altman 1990: 297)

Thus, judges can honestly and sincerely decide absolute borderline cases, even if the law does not require any particular verdict because of its vagueness. The vagueness of the law does not force judges to abandon candor, but it might require a certain abstinence of introspection.

In conclusion, Sorensen's argument for judicial insincerity is unsound because premise (P3) neglects the pragmatic, legal, and argumentative aspects of the judge's speech act when adjudicating a (borderline) case.[7]

## 4. Has Vagueness Really No Function in Law?

Sorensen's main argument against a positive function of vagueness in law can be reconstructed in this very straightforward way:

**(P1)** Vagueness is ignorance.

**(P2)** Ignorance has no (relevant) function in law.

**(K1)** Thus, vagueness has no (relevant) function in law.

Premise **(P1)** is, given the huge community of non-epistemicists, (at least) highly controversial. The vast majority of vagueness theorists is convinced that vagueness is (some kind of) semantic phenomenon. However, vague expressions create uncertainty on whether to apply them in borderline cases, and the view that there is a single right answer to every legal question is not as controversial as ordinary language epistemicism is (cf. Dworkin 1978).

Premise **(P2)** seems to be somewhat problematic as well. Job applications are often anonymously submitted because of unconscious bias. It usually allows for better judgement of the job applicant's abilities not knowing his or her name, age, gender, appearance or place of birth. The judge's judgement of the accused can be similarly biased. Thus, ignorance can facilitate fairer and less biased judgements. However, this kind of ignorance is strikingly different from the kind of irremediable ignorance Sorensen associates with vagueness, since in the former case it is ignorance of certain seemingly relevant, but effectively irrelevant facts that are unknown. Hence, vagueness related ignorance can hardly be said to play any relevant role in these kinds of circumstances.

Based on these considerations, should we accept Sorensen's conclusion that vagueness has no relevant function in law? Can we refute his claim only by proving epistemicism wrong? In

[7] If premise (P3) is rejected, premise (P4) cannot be defended either. Even though judges should have high standards of belief, if verdicts are not simple assertions, they can gather convincing reasons for a particular one such that they come to justifiably believe it to be determined by the law, while in fact it is not. Thus, one could after all dismiss step (S2) of the argument from (K1) to (P4).

fact, I think that one can reject Sorensen's conclusion without committing oneself to any particular view of vagueness. The reason is that not even an epistemicist should accept premise **(P1)** as it is put forward in the argument. Generally the term "ignorance" implies that there is something to know that one is ignorant of. But even if there is a fact of the matter in absolute borderline cases (as epistemicists claim), nobody could possibly know it – according to Sorensen, not even God. Hence, vagueness is not simply ignorance, and premise **(P1)** must be rejected.

I will now positively argue that vagueness has a (relevant) function in law by focusing on an example which Sorensen himself used in his argumentation and Jamie Tappenden originally introduced to the philosophical debate for an argument against epistemicism (cf. Tappenden 1994). If my argument is successful, the contrary conclusion that vagueness has a (relevant) function in law together with premise **(P2)** entails that vagueness is not ignorance. First, however, I will point out what functions of vagueness I have in mind.

Sorensen's argument against the value of vagueness in law focuses primarily on the role of the judge, while neglecting the role of the legislator. The creation of absolute borderline cases can effectively be used by the legislator to reduce decision costs and save time for more important issues (cf. Poscher 2012). The use of vague expressions in law seems to allow some questions to remain open, while giving general guidance. In absolute borderline cases judges and authorities have the requisite discretion to decide either way – precisely because there is nothing to be found or to be known. Lawmakers delegate power to judges and authorities who are often confronted with situations they had no time to consider or simply did not foresee. Even if this brings potential arbitrariness on behalf of the judge when exercising discretion and deciding hard cases, it reduces arbitrariness on behalf of the lawmakers when setting unjustifiable precise thresholds (cf. Endicott 2005).

Vague expressions evidently possess the functions pointed out above. However, it could be the case – as Sorensen pointed out – that these functions are due to other (coincident) properties of vague expressions than their vagueness. Consequently, I will argue now that neither relative borderline cases nor generality could possibly achieve these functions and, hence, it must be absolute borderline cases, i.e. vagueness proper, that provide for them.[8]

Sorensen claims that the "relative borderline cases are doing the work and the absolute borderline cases are epiphenomenal" (Sorensen 2001: 397). The case *Brown v. Board of Education* 349 U.S. 294 (1955), which Sorensen cites to support his claim, was the follow-up of the United States Supreme Court case *Brown v. Board of Education* 347 U.S. 483 (1954), in which the Supreme Court declared state laws unconstitutional that established separate public schools for black and white students.

When the Supreme Court ruled the second *Brown* case in 1955, it could not know how long and difficult the desegregation process would be, so it did not want to decide immediately. That is why it used the infamous expression "with all deliberate speed." Sorensen takes the Court's decision as evidence for the deliberate use of relative borderline cases in the task of power delegation. Future courts are confronted with cases one by one and can gather much more information about each of them than the Supreme Court could generally. Future courts can then easily sort the clear cases and the relative borderline ones and decide them individually, while the absolute borderline cases would be an unavoidable and unwanted by-product. Since what use would it be to facilitate absolute borderline cases in which there is a fact of the matter but no one can ever find out about it?

[8] Of course, there are other potentially functional properties of vague expressions than being general, having relative borderline cases and having absolute borderline cases. However, so far in the philosophical discussion only generality and relative borderline cases have been suggested as functional properties which might be confused with vagueness.

This is where I think Sorensen's epistemicism leads him and us (if we believe him so far) astray. When the Supreme Court decided that schools must be integrated with all deliberate speed, it really could not anticipate how long it would take even progressive schools under the best possible circumstances. For this reason, it used a general and vague expression, deciding on an indeterminate and longish time period. As Sorensen admits, "the Supreme Court coped with borderline cases of 'all deliberate speed' by waiting them out" (Sorensen 2001: 396). While the expression's generality provided wide applicability, its vagueness gave a margin of discretion to schools. This discretion cannot be given by relative borderline cases. Theoretically, the Court could have ascertained all relative borderline cases in advance. Thus, assuming that there were no absolute borderline cases, all subsequent cases would have had only one correct way to decide them and the Court would have known it beforehand. But this is not what the Court had in mind; it wanted to provide some flexibility to account for unforeseen cases, varying circumstances and political hindrances. As Chief Justice Earl Warren expressed it:

> There were so many blocks preventing an immediate solution [...] that the best we could look for would be a progression of action; and to keep it going, in a proper manner, we adopted that phrase, *all deliberate speed*. (Scheiber 2007: 17)

In contrast to generality and relative borderline cases, absolute borderline cases leave some questions actually open. If the ruling would have made use of generality or relative borderline cases alone, the more regressive and conservative schools might have been let to integrate with the objectively slowest allowed speed, which was certainly not what the court wanted to achieve – only absolute borderline cases can account for unforeseen cases and varying circumstances because only they effectively delegate power.

Sorensen's argument can also be presented in a different way. Hrafn Asgeirsson interprets it as claiming that the delegation of power is valuable only if the delegates are in a better position to resolve a borderline case than the delegator (Asgeirsson 2012). But in absolute borderline cases no one is by definition in *any* position to resolve them. Consequently, the delegation of power by way of absolute borderline cases is not valuable.

Asgeirsson rejects Sorensen's claim, however, by arguing that the delegation of power is valuable even if the delegates are not in a better position to resolve a borderline case than the delegator. He agrees with Sorensen that absolute borderline cases prompt judicial discretion and that this discretion is due to an implicit or explicit change in question. In an absolute borderline case one asks not whether some x is F, but whether some x should count as F.

Asgeirsson, then, goes on to argue that this question changing discretion prompted by absolute borderline cases is valuable by pointing out that "being in a better position" does not need to be understood solely epistemically; that the delegator is in a better position to resolve a case than the delegates does not necessarily mean that she has better knowledge than them. Instead, one should understand "being in a better position" as having better tools to find the answer or having lower decision costs. Then it becomes evident that a judge deciding an actual (absolute borderline) case has in fact usually better resources and more information about the particular case than the lawmaker. Hence, it can be sensible for the lawmaker to enact a vague law that reduces her own decision costs and gives discretion to the judges who can decide actual (absolute borderline) cases more individually and less costly.

From this argument it becomes obvious that both Sorensen's argument for the necessary insincerity of judges and the one against the function 0f vagueness in law share the same misleading conception of what verdicts are. In both cases, it is assumed that verdicts are mere assertions of what the law is, while for every case there is a (legal) fact of the matter. Only then it makes sense to talk about judges not being in a better (epistemic) position to decide the case than the lawmaker. Once we have done away with this conception and accepted that

verdicts in absolute borderline cases are real decisions because there is no question to be answered, we will see that vagueness has a function in law.

But even if one is not convinced by my view of what verdicts are and sides with Sorensen on this point, it should have become clear that the United States Supreme Court intentionally chose the vague phrase "with all deliberate speed" *inter alia* to allow judges to decide particular cases at their discretion, and the intentional use of this phrase cannot be explained by any other phenomenon than its vagueness.

## 5. What Is the Lesson?

In summary, judges do not need to resort to lying when ruling absolute borderline cases because the speech act of delivering a judgement is not a mere assertion and is, thus, not directly subject to truth and falsity.

One can argue that judges invoke Dworkinian principles when adjudicating absolute borderline cases, and there is a single right answer, though it is not determined by the relevant laws alone (cf. Dworkin 1978). The vagueness of legal expressions demands then the application of Dworkinian principles, which provide flexibility inasmuch as they presumably change over time, but not by giving judicial discretion. One can also argue, along Hartian lines, that there are various right answers in an absolute borderline case, even though none is the single true one (cf. Hart 1961). In this case, judges have real discretion, since they can use legal and extra-legal principles, welfare considerations, moral and ethical beliefs, and many other reasons. Both positions are compatible with the arguments for the value of vagueness presented in this paper, although a Dworkinian framework would require some adjustments.

The arguments I have made are also compatible with Sorensen's epistemicism as well as with any other theory of vagueness. Independently of one's own approach to vagueness it should have become evident that vagueness has a relevant function in law because it can be (and actually is) used by lawmakers in order to reduce decision costs and to delegate power by giving discretion. This deliberate use of vague language can neither be explained by reference to the generality of the language nor to its allowing merely relative borderline cases.

**David Lanius**

Humboldt Universität zu Berlin
david.lanius@hu-berlin.de

## References

Altman, S. 1990: 'Beyond Candor', in *Michigan Law Review* 89, 296-351.

Asgeirsson, H. 2012: 'Vagueness and Power-Delegation in Law. A Reply to Sorensen', in M. Freeman, and F. Smith (eds.): *Current Legal Issues, Law and Language*, Oxford: Oxford University Press.

Bernal, C. L. 2007: 'A Speech Act Analysis of Judicial Decisions', *European Journal of Legal Studies* 1, 1-24.

Bix, B. 2003: *Law, Language and Legal Determinacy*. Oxford: Clarendon Press.

Choi, A. H., and G. G. Triantis 2010: 'Strategic Vagueness in Contract Design: The Case of Corporate Acquisitions', *Yale Law Journal* 119, 848-924.

Dworkin, R. M. 1978: *Taking Rights Seriously*. Cambridge (MA): Harvard University Press.

Endicott, T. A. O. 2005: 'The Value of Vagueness', in V. K. Bhatia, J. Engberg, M. Gotti, and D. Heller (eds.): *Vagueness in Normative Texts, Linguistic Insights 23*, Bern: Lang, 27-48.

Fine, K. 1975: 'Vagueness, Truth and Logic', *Synthese* 30, 265-300.

Franke, M., G. Jäger, and R. van Rooij 2011: 'Vagueness, Signaling and Bounded Rationality', in T. Onada, D. Bekki, und E. McCready (eds.): *New Frontiers in Artificial Intelligence* 6797, Berlin: Springer (Lecture Notes in Computer Science), 45-59.

Frankfurt, H. G. 2005: *On Bullshit*. Princeton (NY): Princeton University Press.

Glenn, J. 2007: 'May Judges Sometimes Lie? Remarks on Sorensen's Views of Vagueness and Law', *Sorites* 18, 10-16.

Grice, H. P. 1989: *Studies in the Way of Words*, Cambridge: Harvard University Press.

Hart, H. L. A. 1961: *The Concept of Law*, Oxford: Clarendon Press (Clarendon Law Series).

de Jaegher, Kris, and Robert van Rooij 2011. 'Strategic Vagueness, and Appropriate Contexts' in A. Benz, C. Ebert, G. Jäger, and R. van Rooij (eds.): *Language, Games, and Evolution*, Berlin: Springer, 40-59.

Keefe, R. 2000: *Theories of Vagueness*, Cambridge: Cambridge University Press.

Poscher, R. 2009: 'The Hand of Midas: When Concepts Turn Legal', in J. C. Hage, and D. von der Pfordten (eds.): *Concepts in Law* 88, Dordrecht: Springer (Law and Philosophy Library), 99-115.

Poscher, R. 2012: 'Ambiguity and Vagueness in Legal Interpretation', in L. Solan, and P. Tiersma (eds.): *Oxford Handbook on Language and Law*, Oxford: Oxford University Press.

Scheiber, H. N. 2007: *Earl Warren and the Warren Court: The Legacy in American and Foreign Law*, Lanham: Rowman & Littlefield Publishers.

Sorensen, R. 2001: 'Vagueness Has No Function in Law', *Legal Theory* 7, 387-417.

Staton, J. K., and G. Vanberg 2008: 'The Value of Vagueness: Delegation, Defiance, and Judicial Opinions', *American Journal of Political Science* 52, 504-519.

Tappenden, J. 1994: 'Some Remarks on Vagueness and a Dynamic Conception of Language', *Southern Journal of Philosophy* 33, 193-201.

Woozley, A. D. 1979: 'No Right Answer', *Philosophical Quarterly* 29, 25-34.

# A Single-Type Ontology for Natural Language

Kristina Liefke

In (Montague 1970a), Richard Montague defines a formal theory of linguistic meaning which interprets a small fragment of English through the use of *two* basic types of objects: individuals and propositions. In (Partee 2006), Barbara Partee conjectures the possibility of a comparable semantic theory, which only uses *one* basic type of object (hence, *single-type semantics*). This paper supports Partee's conjecture by identifying two suitable single-type candidates. To single out the latter, we first introduce a set of semantic requirements on any single basic type. The application of these requirements to the familiar types from (Montague 1973) reduces this set to two members. The paper closes with a comparison of Partee's preliminary single-type choice and our newly identified single basic types.

## 1. Introduction

Natural languages presuppose a rich semantic ontology. To provide an interpretation for, e.g., English, we require the existence of individuals (e.g. Bill), propositions (Bill walks), first- and higher-order properties (walk, rapidly), relations (find), and many other kinds of objects. Theories of formal linguistic semantics (paradigmatically (Montague 1970a; 1970b; 1973)) tame this ontological 'zoo' by casting its members into a type structure, and generating objects of a more complex type from objects of a simpler type via a variant of Church's type-forming rule (Church 1940), (cf. Gallin 1975):

**Type-forming rule (CT)**

If $\alpha$ and $\beta$ are the types for two (possibly different) kinds of objects, then $<\alpha, \beta>$ is the type for functions from objects of the type $\alpha$ to objects of the type $\beta$.

In this way, Montague (1970a) reduces the referents of the small subset of English from (Montague 1973) to constructions out of two basic types of objects: individuals (or *entities*, type *e*) and propositions (or functions from indices to truth-values, type $<s, t>$). Proper names (e.g. *Bill*) and sentences (*Bill walks*) are then interpreted as entities, respectively propositions, intransitive verbs (*walk*) as functions from entities to propositions (type $<e, <s, t>>$), and transitive verbs (*find*) as functions from entities to functions from entities to propositions (type $<e, <e, <s, t>>>$).

Montague's distinction between entities and propositions (or between entities, indices, and truth-values) has today become standard in formal semantics. This is due to the resulting semantics' modelling power, and the attendant possibility of explaining a wide range of syntactic and semantic phenomena. However, recent findings in language development (Carstairs-McCarthy 1999; Cheney and Seyfarth 1990; Snedeker et al. 2007) suggest the possibility of an even simpler semantic basis for natural language. The latter lies in a single basic type (dubbed '*o*') whose objects encode the semantic content of entities and propositions. From them, objects of a more complex type are constructed via a variant of the rule **CT**:

**Single-type-forming rule (ST)**

If *o* is the single basic type and $\alpha$ and $\beta$ are single-type types, then $<\alpha, \beta>$ is a single-type type.

As a result of the neutrality of the type *o* between Montague's types *e* and <*s*, *t*>, we can identify basic single-type objects with the semantic values of proper names and sentences. Intransitive and transitive verbs can then be interpreted as objects of the types <*o*, *o*>, respectively <*o*, <*o*, *o*>>.

In reflection of the observations from the previous paragraphs, Barbara Partee (2006) has recently made the following claim about the linguistic ontology:

> **Proposition 1 (Single-Type Hypothesis)**
>
> Montague's distinction between entities and propositions is inessential for the construction of a rich linguistic ontology. Natural language can be modelled through the use of a single basic type of object.

Partee supports her hypothesis by identifying a preliminary single-type candidate (i.e. properties of Kratzerian situations (cf. Kratzer 1989)), and describing representations of several Montagovian objects in this type. This procedure suggests the suitability of her basic-type choice for the type-neutral interpretation of proper names and sentences. However, the brevity of her exposition prevents a detailed motivation of the latter. In particular, Partee does not identify properties of situations as the *only* suitable single-type candidate. Her characterization of single-type objects with "ontologically neutral" objects (p. 39) further posits the single basic type outside of the Montagovian type system and, thus, obfuscates the relation between single-type objects and basic Montagovian objects[1].

The present paper attempts to compensate for these shortcomings. Its objective lies in the identification of suitable *Montague* types for the modelling of natural language along the lines of Proposition 1, that stand in a definable relation to all linguistically relevant types. The plan for the paper is as follows: To narrow down our choice of single-type candidates, the first section (Sect. 2) introduces a set of requirements that ensure the type's suitability as a single semantic basis for natural language. The application of these requirements to the simplest Montague types (Sect. 3) reduces the set of possible single-type candidates to the types <*s*, *t*> and <*s*, <*s*, *t*>>. The paper closes with a comparison of Partee's basic-type choice and our newly identified single basic types (Sect. 4).

## 2. Single-Type Requirements

Arguably, not all Montague types which are obtained from the types *e*, *s*, and *t* through the type-forming rule **CT** are equally suitable as a single semantic basis for natural language. To identify the best candidate(s), we demand that they satisfy Properties 0 to 5, below:

(0) **Familiarity:** *The single basic type figures in the formal semantic analysis of some linguistic phenomenon.*

(1) **Booleanness:** *The single-type domain has an algebraic structure.*

(2) **Representability:** *All types of objects can be represented via (objects of) the single basic type.*

(3) **Intensionality:** *Single-type objects have strong identity criteria.*

(4) **Partiality:** *Some single-type objects are not fully defined.*

(5) **Simplicity:** *Given its satisfaction of Properties 0-4, the single basic type is obtained from e, s, and t through the least number of* **CT***-applications.*

[1] For example, Partee describes the single-type correspondent of the semantic value of the pronoun *you* (in a given context) as "[t]he property a situation has if it's a minimal situation containing you" (Partee 2006, 40). However, the containment relation between entities and situations is never formally defined.

In particular, Property 0 ensures the proximity of our single-type system to mainstream formal semantics. Property 1 is required for the interpretation of natural language connectives as algebraic operations. The obtaining of Property 2 allows the bootstrapping of representations of all Montagovian objects from objects of the single basic type. Property 3 ensures that single-type representations of Montagovian objects allow correct predictions about linguistic entailment. Property 4 enables the representation of information growth in single-type semantics. Property 5 warrants the low semantic complexity of single-type objects.

In virtue of their conceptual simplicity, the requirements from Properties 0 and 5 do not demand further exposition. The requirements from Properties 1 to 4 are discussed in detail in Sections 2.1 through 2.4. There, the reader will observe that our advance from one section to the next involves a decrease in the semantic generality of the presented properties. In particular, we will see in Section 3 that the properties of Representability (Sect. 3.2, cf. Sect. 2.2) and Partiality (Sect. 3.4, cf. Sect. 2.4) are intimately connected to the semantics of the single basic type. In this way, our discussion of single-type requirements will give us a feel for what is necessary and what is possible in single-type semantics.

We start with a discussion of algebraicity or Booleanness (Property 1).

### *2.1. Booleanness*

Algebraicity constitutes the most general semantic requirement on any single basic type. This property is demanded by the need to provide semantic counterparts of natural language connectives, and to give a formal basis for the relation of linguistic entailment. Many linguists have suggested that the English words *and, or,* and *not* act as algebraic operators between linguistic expressions of the same category. However, this interpretation of English connectives is only possible if their respective domains exhibit an algebraic structure.

Further, since all single-type candidates are, by definition, the *only* basic type in their associated logic, the algebraic structure of complex single-type domains (i.e. domains of the type $<o, o>$, $<o, <o, o>>$, etc.) depends entirely on the structure of the candidate's base domain. Thus, domains of a complex type only form an algebra if the set of basic-type objects forms an algebra. The latter enables the interpretation of linguistic conjunction, disjunction, and negation as meet, join, and complement in a typed algebraic domain. Entailment between type-identical expressions can then be treated as inclusion of objects of the single-type type.

We next turn to a presentation of the requirement of representability on objects of the single basic type (Property 2).

### *2.2. Representability*

The representability requirement on single-type candidates is a direct consequence of Partee's conjecture from Proposition 1. The latter demands that objects of any suitable single basic type enable us to bootstrap representations of all Montagovian objects. Let *o* be our single basic type. The requirement of representability is then expressed below:

**2.' Representability:**

*Let a Montague type $\alpha$ be related† to some single-type type $\beta$ if there exists, for every type-$\alpha$ object a exactly one type-$\beta$ object b which represents a, such that the objects a and b are one-to-one related. Define $(<\beta_1, <. . . , \beta_n >> \# o)$ as $< \beta_1, <. . . , < \beta_n, o>>>$ for single-type types $\beta_1, . . . , \beta_n$. Then, one of the following holds for all Montague types $\alpha$:*

*(a) The type $\alpha$ is related† to the single basic type o;*

*(b) The type $\alpha$ is related† to some derived type $<< \beta_1, <. . . , \beta_n >>, o>$ or $(<\beta_1, <. . . , \beta_n >> \# o)$, where $\beta_1, . . . , \beta_n$ are single-type types.*

Given the existence of a unique single-type representation for every basic Montagovian object (clause (a)), the rule **ST** ensures the existence of a unique single-type representation for every Montagovian object of a complex type. As a result, it suffices for a demonstration of the satisfaction of the Representability requirement to show that (2a) obtains.

We will give concrete examples of the success and failure of the representability of single-type objects in Sections 3.2 and 3.3, respectively. The next subsection presents the requirement of their intensionality (Property 3).

### *2.3. Intensionality*

The intensionality requirement on single basic types is a response to the granularity problem for linguistic meanings from (Frege 1892). The latter concerns the fact that interpretations of natural language expressions do not individuate semantic objects as finely as speakers' intuitions about strict synonymy, to the effect that there are too few intensions to enable correct predictions about linguistic entailment (Muskens 1995).

Most logics for natural language adopt a version of the axiom scheme of Extensionality from (Ext), where the variables $X$ and $Y$ range over objects of some type $<\alpha_1, <. . . , <\alpha_n, t>>>$ and where the variables $x_1, . . . , x_n$ have the Montague types $\alpha_1, . . . , \alpha_n$:

**Extensionality (Ext)**

$\forall X \, \forall Y. \, \forall \, x_1 \ldots x_n \, (X \, (x_1) \ldots (x_n) \leftrightarrow Y \, (x_1) \ldots (x_n)) \rightarrow X = Y)$

As a result, models of such logics identify all objects of some type $<\alpha_1, <. . . , <\alpha_n, t>>>$ (*or* $<<<\alpha_1, . . . >, \alpha_n,> t>$) that are logically equivalent. For the case of (type-$<s, t>$) propositions, these models identify all propositions which have the same truth-values across all indices. Thus, the proposition 'John seeks a unicorn' will be treated as identical to the propositions 'John seeks a unicorn and $1^3 + 12^3 = 9^3 + 10^3$' and 'John seeks a unicorn and Bill walks or does not walk'.[2]

In an extensional setting (where we are concerned with a description of the actual *physical world*, and assume the availability of *all relevant facts* about this world), the identification of the above propositions is unproblematic. Thus, the inference from (1a) to (1b) or (1c) is intuitively valid, where @ denotes planet Earth on Wednesday 12th December, 2012 at 10:52am.

(1) a. At @, John seeks a unicorn.

b. At @, John seeks a unicorn and $1^3 + 12^3 = 9^3 + 10^3$.

c. At @, John seeks a unicorn and Bill walks or does not walk.

However, in epistemic contexts, the extension of our commitment from a single proposition to the set of its semantic equivalents is much less-warranted. This is due to the fact that a cognitive agent may possess only *partial* information about the physical world, such that (s)he may assume the truth of one, but not of another proposition. The substitution *salva non veritate* of the proposition 'John seeks a unicorn' in the complement of the verb *believes* in (2a) by any of the propositions from (1b) or (1c) (in (2b), resp. (2c)) illustrates the special status of such contexts:

(2) a. Mary believes that John seeks a unicorn.

b. Mary believes that John seeks a unicorn and $1^3 + 12^3 = 9^3 + 10^3$.

[2] We will justify the equivalence of these three propositions in Section 2.4.

c. Mary believes that John seeks a unicorn and Bill walks or does not walk.

To block invalid inferences of the above form, formal semanticists have, in the last 30 years, sought for more strongly intensional notions of proposition, whose objects exhibit more fine-grained identity criteria. In particular, such notions have been developed in the frameworks of 'structured meanings' theories (Cresswell 1985), Situation Semantics (Barwise and Perry 1983), Data Semantics (Landman 1986), Property Theory (Chierchia and Turner 1988), impossible world semantics (Rantala 1982), and partial type theory (Muskens 1995).

Our stipulation of the intensionality requirement on single-type candidates is motivated by the characterization of single-type semantics as a theory of meaning *for natural language*, that therewith also models epistemic statements of the form from (2a) to (2c).

We finish our discussion of single-type requirements with a presentation of the requirement of partiality (Property 4). The latter concerns the possibility of leaving some single-type objects underdefined, such that they can be extended into better-defined, total objects.

### *2.4. Partiality*

In single-type semantics, partiality serves double duty as a strategy for the obtaining of fine-grained linguistic meanings, and for the modelling of information growth. On its former use, the adoption of partial single-type objects constitutes a means of satisfying the intensionality requirement from Section 2.3, that follows the approach of partial type theory. On its latter use, the adoption of partial single-type objects constitutes a means of accommodating the dynamics of linguistic meanings (cf. van Benthem 1991).

We discuss the two rationales for the adoption of partial single-type objects below. To prepare their presentation, we first provide a brief characterization of partial objects (or functions). We then show that the properties of dynamicity and intensionality arise naturally from the property of partiality.

The characterization of single-type objects as *partial* objects relates to their algebraic structure – in particular to the identity of the type-*t* domain. Thus, since the 'ingredient'-type *t* of the type for propositions $<s, t>$ is classically associated with the two truth-values *true* (**T**) and *false* (**F**), objects of some type $<\alpha_1, <. . . , <\alpha_n, <s, t>>>>$ are taken to be *total* (or *Boolean*) functions, that obey the law of Excluded Middle. As a result, one can directly obtain a function's complement from the function itself.

Our description of single-type objects as *partial* objects involves a generalization of the set of truth-values to the set {**T**, **F**, **N**}, where **N** is the truth-valuationally undefined element (*neither-true-nor-false*). As a result of its introduction, some functions of the type $<\alpha_1, <. . . , <\alpha_n, <s, t>>>>$ will send certain arguments to the truth-value **N** and do, thus, fail to satisfy the law of Excluded Middle (Fact 1). The partiality of the set of truth-values further induces a definedness ordering on all partial single-type domains (Fact 2). We will see below that our use of partial single-type objects for the obtaining of fine-grained linguistic meanings uses Fact 1. Our use of partial single-type objects for the modelling of information growth employs Fact 2.

We will demonstrate the *dynamicity* of the partial single-type objects in Section 3.4. Their intensionality (cf. Sect. 2.3) is illustrated below:

The fine-grainedness of partial single-type objects is a consequence of our adoption of the partial set of truth-values {**T**, **F**, **N**}. Our consideration of the propositions 'John seeks a unicorn' and 'John seeks a unicorn and Bill walks or does not walk' from Section 2.3 illustrates this point: The logical equivalence of these propositions is conditional on the adoption of the law of Excluded Middle in the underlying logic, and the attendant possibility of identifying propositions of the form $(p \vee \neg p)$ with universally true propositions. Thus, the proposition 'John seeks a unicorn' is only equivalent to the proposition 'John seeks a unicorn and Bill

walks or does not walk' if the sentence *Bill walks or does not walk* receives the value **T** at every index. The partiality of the propositional domain and the resulting invalidity of the law of Excluded Middle prevent this truth-assignment. The inference from (2a) to (2c) is blocked.

The partial assignment of truth-values to indices (above) suggests the existence of differently well-defined indices, that are ordered with respect to the richness (or *strength*) of their encoded propositional information. Our blocking of the inference from (2a) to (2c) exploits the existence of partial indices at which a given proposition (here, the proposition 'Bill walks') and its complement are both undefined. Our blocking of the inference from (2a) to (2b) adopts this strategy: While the proposition '$1^3 + 12^3 = 9^3 + 10^3$' is true at all total indices, it is undefined at some partial indices. The existence of the latter prevents the identification of the proposition '$1^3 + 12^3 = 9^3 + 10^3$' with the universally true proposition, and avoids the equivalence of the proposition 'John seeks a unicorn' with the proposition 'John seeks a unicorn and '$1^3 + 12^3 = 9^3 + 10^3$'.

This completes our discussion of the semantic requirements on the single basic type. We next show how the latter can be used to identify the most suitable single semantic basis for the modelling of natural language.

## 3. Eliminating Unsuitable Candidates

Section 1 has suggested an eliminative identification of the most promising single-type candidate(s): From the set of Montague types, we successively exclude members on the basis of their failure to satisfy Properties 0 to 5.

Above, we have identified the domain of elimination for Properties 0 to 4 with the set of the *simplest* Montague types, whose members are obtained from Montague's basic types through the least number of applications of the rule **CT**. In particular, we will hereafter consider Montague types that have been obtained through at most *two* applications of this rule. This restriction is justified by the existence of suitable single-type candidates in the resulting set, and by our adoption of the requirement of simplicity from Property 5.

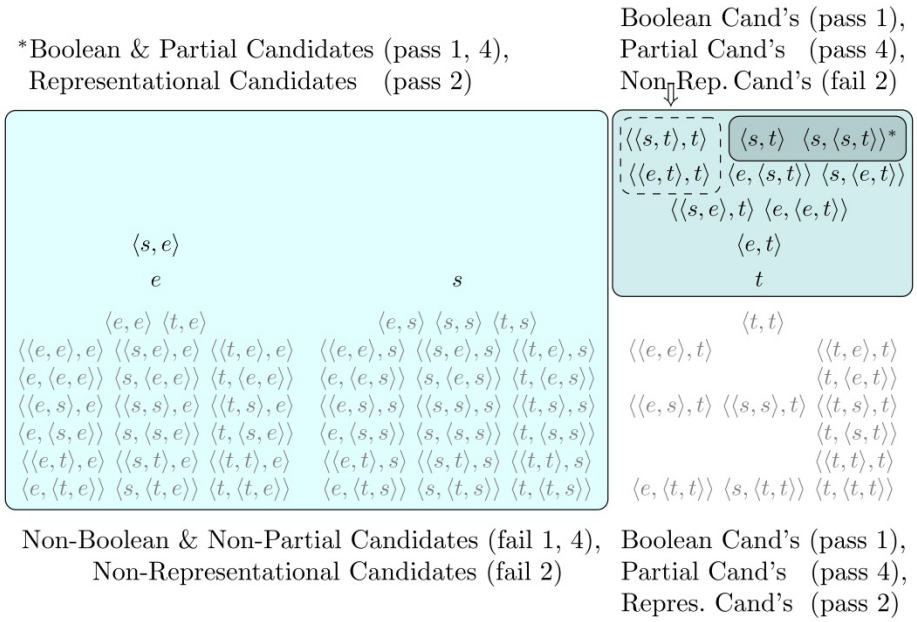



TABLE 1. Single-Type Candidates and their Elimination.

To enable the identification of the simplest single basic type, we replace the types *e* and <*s*, *t*> from the introduction (Sect. 1) by the types *e*, *s*, and *t*. This move corresponds to the adoption

of a streamlined variant of Montague's type theory from (Montague 1970a; 1973), that is due to Gallin (1975). The (at most) double application of **CT** to Gallin's set of basic types yields the set of single-type candidates from Table 1.

Sections 3.1 to 3.4 successively eliminate single-type candidates on the basis of their failure to satisfy Properties 1, 2, and 4. The decorations in Table 1 summarize the reasons for the persistence or drop-out of each candidate. In the table, Montague types that violate the requirement of Familiarity (Property 0) are marked in grey.

### *3.1. Eliminating Non-Boolean Types*

The lack of algebraic structure (Property 1) constitutes one of the most effective criteria for the exclusion of single-type candidates from the set in Table 1. The latter relates to the difficulty of interpreting linguistic connectives in non-algebraic domains, and of analyzing linguistic entailment in the absence of a domination relation. In Sections 2.1 and 2.4, we have already suggested that the domain of the type $t$ has an algebraic structure, and that all domains of some type $<\alpha_1, <. . . , <\alpha_n, <s, t>>>>$ inherit this structure through the lifting of algebraic operations on the set of truth-values. As a result, all candidates from the right-side partitions of Table 1 are suitable single basic types from the point of view of Property 1.

On the basis of Property 1, candidates from the left-side partition of Table 1 *disqualify* as suitable single-type candidates, such that they can*not* serve as the single semantic basis for natural language. This is a result of the absence of an algebraic structure on the domains of entities and indices, and the attendant 'primitiveness' of all domains of some type $<\alpha_1, <. . . , <\alpha_n, e>>>$ or $<\alpha_1, <. . . , <\alpha_n, s>>>$. Since the latter constitute two thirds of the members of the set from Table 1, the algebraicity requirement from Property 1 already enables us to exclude most of the available candidates as strong single-type candidates.

Notably, our elimination of entities from the set of types in Table 1 also excludes a common type for entities and propositions (or for entities and truth-values) as a suitable single semantic basis for natural language. The latter has been proposed by some semanticists[3] as an obvious single-type candidate, and has been motivated with reference to Frege's characterization of truth-values as *Gegenstände* (cf. Frege 1891). On this assumption, Frege's linguistic ontology can be construed as a semantics for natural language that obtains representations of all Montagovian objects from the single basic type $e$. However, given the identification of this type as a non-Boolean type, a common type $e$ for entities and propositions is ruled out as a suitable single-type candidate. We will see in Section 3.3 that a semantics based on the type $<s, t>$ satisfies all requirements.

This completes our elimination of non-Boolean types from the set of single-type candidates. We next investigate the exclusion of non-representational types from the remaining set.

### *3.2. Eliminating Non-Representational Types*

The ability of representing different Montagovian objects is a more elusive criterion for the exclusion of single-type candidates than their algebraicity. This is due to the impossibility of inferring a type's satisfaction of Property 2 from its outer type structure. As a result, we need to check the representability of the remaining single-type candidates from the top right-side partition in Table 1 one-by one.

To this aim, we will first consider single-type candidates (i.e. the types $t$, $<e, t>$, $<s, <e, t>>$, $<e, <s, t>>$, and $<e, <e, t>>$), which fail to provide index-relative (*local*) representations of Montagovian entities and propositions. We will then consider candidates (i.e. the types $<<e, t>, t>$ and $<<s, t>, t>$) which succeed in providing local, but which fail at giving suitable

[3] Proponents include Chierchia and Turner (1988) and Zoltán Gendler Szabó (p.c.).

index-general (*global*) representations. Section 3.3 presents two single-type candidates (i.e. <*s*, *t*> and <*s*, <*s*, *t*>>), which enable the global representation of entities and propositions. We start our elimination of non-representational types with the type for truth-values, *t*.

Arguably, truth-values do not enable the representation of objects of all Montagovian types. This is a result of the small cardinality of the set of truth-values, and the large cardinality of the domains of entities and propositions. As a result, the representation of entities and propositions via truth-values associates different Montagovian objects with the same single-type object. But this violates our assumption of a *correspondence* between Montagovian objects and their single-type representations from the formal definition of Representability (cf. Sect. 2.2). Arguably, the replacement of the set {**T**, **F**} (or {**T**, **F**, **N**}) by a countably infinite set of truth-*degrees* establishes the desired correspondence. However, since there does not exist a principled relation between the members of these two sets, the type *t* is ruled out as a suitable single basic type.

On the basis of their inability to give a suitable representation of propositions, extensional (type-<*e*, *t*>) and intensional properties of entities (type <*s*, <*e*, *t*>> or <*e*, <*s*, *t*>>) and binary relations between entities (type <*e*, <*e*, *t*>>) are also ruled out as single semantic bases for natural language. Consider the case of extensional properties (i.e. functions from entities to truth-values, or sets of entities): Their adoption enables the representation of entities *a* via their singleton sets {*a*}. However, there does not exist a comparable strategy for the representation of propositions. This results from the fact that entities typically carry more than one property (Fact 1), and that some propositions encode information about more than one entity (Fact 2).

As a consequence of Fact 1, a proposition's representation in the type <*e*, *t*> (e.g. the representation of some proposition *Fa* via the set {*a*}) may be ambiguous between different propositions (which carry information about *a*), such that their representation is not injective. As a consequence of Fact 2, a single proposition (e.g. the proposition 'John loves Mary') can have different representations in the type <*e*, *t*> (here, via the set of entities which John loves or via the set of entities which love Mary), such that their representation is also not functional. The described relation between propositions and their single-type representations undermines the correspondence assumption from Section 2.2.

The representations of propositions via intensional properties (i.e. via functions from entities to propositions, or sets of entity-index pairs) or via relations between entities (i.e. via functions from entities to functions from entities to truth-values, or sets of ordered pairs of entities) face a similar problem. As a result, the types <*s*, <*e*, *t*>>, <*e*, <*s*, *t*>>, and <*e*, <*e*, *t*>> also disqualify as single basic types.

Our discussion of the exclusion of the type <*e*, *t*> from the set of single-type candidates has suggested the possibility of representing all Montagovian objects via sets of objects of the type <*e*, *t*> (i.e. via (type-<<*e*, *t*>, *t*>) functions from extensional properties of entities to truth-values, or via generalized quantifiers over entities). The latter enable the representation of entities via the singletons of their singleton sets (such that every entity *a* is represented by the set {{*a*}}), or via the set of their extensional properties at the current index @. A given entity $\alpha$ is then represented by the set of its extensional properties from (1), where variables are subscripted with their logical type:

(1) $\{P_{<e,\,t>} \mid a \in P \text{ at } @\}$

On the basis of the above, propositions can be represented via the union of the representation of their type-*e* argument(s) in the type <<*e*, *t*>, *t*> and the singleton containing their

attributed property.[4] In particular, a proposition of the form *Fa* is then represented by the set of extensional properties from (2):

(2) $\{ P_{<e, t>} \mid a \in P \text{ at } @\} \cup \{\{ x_e \mid x \in F \text{ at } @ \,\&\, a \in F \text{ at } @\}\}$

In virtue of the possibility of representing entities and propositions, the type $<<e, t>, t>$ enables the representation of all Montague types along the lines of Property 2. However, as is captured by our use of the variable @, the representations of Montagovian objects from (1) and (2) are still dependent on the current (possibly partial) index. As a result, it may happen that two distinct entities are, at some index, represented by exactly the same set of properties. For example, at the index at which only the propositions 'John is self-identical' and 'Mary is self-identical' are true, the entities John and Mary will be represented by exactly the same set of properties. But this, again, violates our assumption of a one-to-one relation between Montagovian objects and their single-type representations.

The type-$<<s, t>, t>$ representations of entities and propositions from (3) and (4) suffer from a similar restriction. The latter are motivated by the considerations for the type-$<<s, t>, t>$ representations of basic Montagovian objects from (8) and (13). In the latter, the constant φ denotes the represented proposition. A proposition's *aboutness* with respect to a given entity is defined in terms of the entity's existence at every index at which the proposition is true (such that *p* 'is about' *a* iff, for all indices *w*, if $w \in p$, then *a* exists in *w*).

(3) $\{p_{<s, t>} \mid w \in p \,\&\, p \text{ 'is about' } a\}$

(4) $\{p_{<s, t>} \mid (w \in p \text{ or } p = \varphi) \,\&\, \text{for some } x_e, \varphi \text{ 'is about' } x \,\&\, p \text{ 'is about' } x\}$

In view of the representability of the type $<s, t>$ (cf. (7), resp. (11)), the successful global representations from (5) and (6) also disqualify as suitable representations of Montagovian objects (as a result of their violation of Simplicity)):

(5) $\{p_{<s, t>} \mid \{p = \{w_s \mid a \text{ exists in } w\}\}$

(6) $\{p_{<s, t>} \mid \{p = \{w_s \mid w \in \varphi\}\}$

This completes our exclusion of single-type candidates on the basis of their failure to satisfy the representability requirement. We next turn to the remaining types, $<s, t>$ and $<s, <s, t>>$, from the set of single-type candidates in Table 1.

### *3.3. Identifying Representational Types*

Our discussion of a 'Fregean' single-type semantics from Section 3.1 has already suggested the possibility of representing entities *a* in the type for propositions $<s, t>$. Specifically, the identification of the single basic type with the type $<s, t>$ enables the representation of every entity *a* via the set of indices at which it exists:

(7) $\{w_s \mid a \text{ exists in } w\}$

Since we assume that there exists, for every entity *a*, exactly one world *w* which only this entity inhabits, the representation of entities from (7) ensures a correspondence between entities and their single-type representations. This fulfills the representability requirement from Property 2.

However, in virtue of the above, the correspondents of entities in the type $<s, t>$ carry only very *weak* semantic information. The correspondents of entities in the type $<s, <s, t>>$

[4] To avoid an unnecessary complication of the material, we restrict ourselves to the consideration of single-argument propositions.

compensate for this shortcoming.[5] In particular, their representational strategy follows the strategy for the representation of entities in the type <<*e*, *t*>, *t*> (cf. (1), (2)). Only, rather than representing an entity via the set of its extensional properties at @, we represent the latter via the set of true propositions at @ which carry information about it; or, equivalently, via the set of indices at which all true propositions at @ which carry information about the entity are true. An entity *a* is then represented by the set of indices from (8):

(8) $\{w_s \mid \text{for all } p_{<s,\,t>}, \text{ if } @ \in p \ \&\ p \text{ 'is about' } a, \text{ then } w \in p\}$

To ease reference, we hereafter denote the representation of *a* from (8) by '$a^{\dagger}_{@}$'.

We illustrate the above representation strategy by means of an example: Consider the representation of John in a universe consisting of three indices @, $w_1$, and $w_2$, and two distinct entities: John (abbreviated *j*) and Mary (abbreviated *m*). Assume further that, at the current index, the propositions 'John runs' (*Rj*), 'Mary runs' (*Rm*) and 'Mary doesn't whistle' ($\neg Wm$) are true, that, at the index $w_1$, the propositions 'John runs', 'John whistles', 'Mary runs' and 'Mary doesn't whistle' are true (such that *Rj*, *Wj*, *Rm*, and $\neg Wm$ obtain at $w_1$), and that, at the index $w_2$, the propositions 'John doesn't run', 'Mary runs', 'John whistles', and 'Mary whistles' are true (such that $\neg Rj$ , *Rm*, *Wj*, and *Wm* obtain at $w_2$) (cf. Fig.1) .

Then, by the characterization of aboutness from Section 3.2, the proposition 'John runs' is the only true proposition at @ which carries information about John. As a result, we represent John at @ by the subset of the set {@, $w_1$, $w_2$} at whose members John runs. We identify the latter with the set {@, $w_1$} (underbraced in Fig. 1). The latter encodes the information that John runs at @.

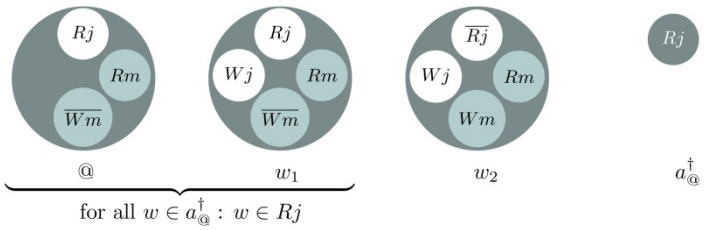



FIGURE 1. A rich $\langle s, t\rangle$-representation of John at @ (1).

Notably, the representation of entities along the lines of (8) presupposes the existence of the represented entity *a* at the current index: If *a* does not exist at @, no proposition *p* will satisfy the condition on the set from (8). Since *a* will thus be represented by the empty set of indices, its type-<*s*, *t*> representation at @ will be identified with the type-<*s*, *t*> representations of all other non-existing entities at @. But this is arguably undesirable.

To solve this problem, we hedge the condition, $@ \in p$ & *p* 'is about' *a*, on the set $a^{\dagger}_{@}$ from (8) with the conjunct '*a* exists in *w*'. The entity *a* is then represented by the set of indices from (9):

(9) $\{w_s \mid a \text{ exists in } w \ \&\ \text{for all } p_{<s,\,t>}, \text{ if } @ \in p \ \&\ p \text{ 'is about' } a, \text{ then } w \in p\}$

Figure 2 illustrates the representation of John at an index in which he does not exist. In the figure, we abbreviate 'John exists' as '*Ej*'.

[5] Since they encode semantically 'richer' information than type-<*s*, *t*> propositions, objects of the type <*s*, <*s*, *t*>> still satisfy the simplicity requirement from Property 5.

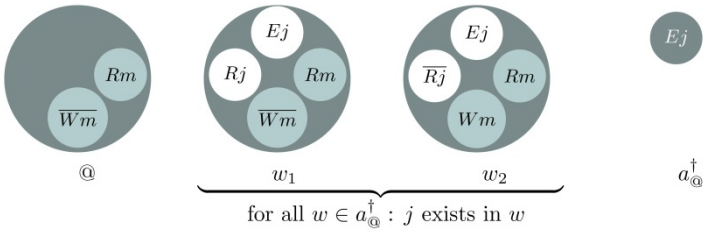



FIGURE 2. A rich $\langle s, t\rangle$-representation of John at @ (2).

This completes our description of the strong representations of entities in the propositional type <*s*, *t*>. Notably however, the former are still dependent on the current index (and do, thus, not satisfy the representability requirement from Property 2). To ensure the desired *correspondence* between entities and their rich single-type representations, we represent the entity *a* by the index-general variant, (10), of (9):

(10) {<$w_1$, $w$> | *a* exists in $w$ & for all $p_{<s,\,t>}$, if $w_1 \in p$ & $p$ 'is about' *a*, then $w \in p$}

The representation of *a* from (10) is (equivalent to) an object of the type <*s*, <*s*, *t*>>, that is associated with a function from indices to the rich type-<*s*, *t*> representation of the entity *a* at those indices. We will sometimes denote the object from (10) by '$a^{\dagger}$'.

Objects of the above form are familiar from intensional, indexical, and dynamic semantics. For example, in Landman's version of Data Semantics (Landman 1986), entities in a designated information state σ are represented via the sets of true propositions at σ which carry information about them. In Kaplan's semantics for indexical expressions (Kaplan 1989), the notion of *character* is described as a function from situational contexts (type *s*) to semantic contents (type *e,* or <*s*, *t*>). In Muskens' (1996) type-theoretic formulation of Discourse Representation Theory (cf. Kamp 1981), discourse representation structures are associated with (type-<*s*, <*s*, *t*>>) functions from indices to propositions.

We next turn to the representation of propositions in the type <*s*, <*s*, *t*>>. Following our strategy for the 'rich' representation of entities, we first describe the representation of a proposition's truth-value at the current index. The latter is obtained in the type for propositions, <*s*, *t*>.

Admittedly, the most straightforward strategy for the representation of propositions φ lies in the identification of $\varphi$ with the set of all indices at which it is true (in (11)). The latter can be lifted to the type <*s*, <*s*, *t*>> by taking the constant function (in (12)) from indices to this set.

(11) {$w_s$ | $w \in \varphi$}

(12) {<$w_1$, $w$ > | $w \in \varphi$}

However, propositional representations of the form from (12) violate our intuition that (the representations of) propositions are semantically at least as rich as (the representations of) the entities about which they carry information (hereafter, the propositions' *aboutness subjects*). For example, we expect that the representation of the proposition 'John is bald' contains the information of the representation of John.

To accommodate the semantically rich local representation of propositions, we represent the truth-value of the proposition φ at @ via the intersection (in (13)) of (the intersection of) the <*s*, *t*>-representations of φ's aboutness subject(s) at @ (cf. (9)) and the set of indices at which $\varphi$ is true (in (11)):

(13) {$w_s$ | $w \in \varphi$ & for all $p_{<s,\,t>,}$ if @ ∈ $p$ &

for some $x_e$, ($\varphi \wedge p$) 'is about' $x$, then $w \in p$}

We will sometimes denote the resulting set by $\varphi^{\dagger}_{@}$.

For greater understandability, we again illustrate our representational strategy by means of an example: Consider the rich <*s*, *t*>-representation of the proposition 'John loves Mary' (abbreviated *Lmj*) at @ in a universe consisting of three indices @, $w_1$, and $w_2$, and three (distinct) entities John (*j*), Mary (*m*), and Bill (*b*). Assume further that, at @, the propositions 'John runs' (*Rj*), 'Mary does not run' (¬*Rm*), and 'Bill runs' (*Rb*) are true, that, at the index $w_1$, the propositions 'John loves Mary', 'John runs', 'Mary doesn't run', and 'Bill doesn't run' are true (such that *Lmb*, *Rj*, ¬*Rm*, and ¬*Rb* obtain at $w_1$), and that, at the index $w_2$, the propositions 'John loves Mary', 'John runs', 'Mary runs', and 'Bill doesn't run' are true (such that *Lmj*, *Rj*, *Rm*, and ¬*Rb* obtain at $w_2$):

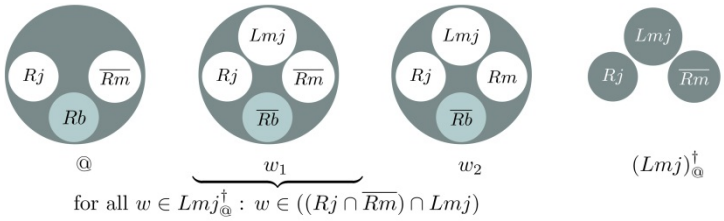



FIGURE 3. A rich $\langle s, t\rangle$-representation of the proposition 'John loves Mary' at @.

Then, the propositions 'John runs' and 'Mary doesn't run' are the only true propositions at @ which carry information about the aboutness subjects of the proposition 'John loves Mary'. As a result, we represent the truth-value of 'John loves Mary' at @ by the subset of the set {@, *w*1, *w*2} at whose members the propositions 'John runs', 'Mary doesn't run', and 'John loves Mary' are true.[6] We identify the latter with the singleton set {*w*1} (underbraced in Fig. 3).

To ensure a correspondence between propositions and their rich single-type representations (in analogy with (10)), we represent propositions φ by an index-general variant, (14), of their representation from (13):

(14) {<$w_1$, *w*> | *w* ∈ φ & for all $p_{<s,\,t>}$, if $w_1 \in p$ &

for some *x*, (φ ∧ *p*) 'is about' *x*, then *w* ∈ *p*}

We will sometimes denote objects of the above form by φ†.

This completes our elimination of single-type candidates on the basis of their failure to satisfy the representability requirement from Property 2.

Our investigation of the Booleanness and Representability requirements from Sections 3.1 and 3.2 has reduced the set of single-type candidates from Table 1 to all but two members: the types <*s*, *t*> and <*s*, <*s*, *t*>>. Since the requirements of Intensionality and Partiality depend only on the domanial structure of an algebraic type, they are unable to exclude further single-type candidates. However, we will see below that they place specific constraints on the candidates' associated objects. The latter are specified below. Since the requirement of Intensionality is trivially satisfied in virtue of the types' partiality (cf. Sect. 2.4), we only discuss the requirement of Partiality.

### *3.4. Obtaining Partial Types*

The previous subsection has described representations of entities and propositions in the type <*s*, <*s*, *t*>> as relations between pairs of indices of the form from (10) and (14). Correspondingly, at the index $w_1$ from Figure 1, the entity John and the proposition 'John whistles' are represented by the set of indices from (15):

[6] The last requirement compensates for the fact that 'John loves Mary' is undefined at @. The latter is analogous to the existence requirement on the representation of entities from (9).

(15) $\{w_s \mid$ [[John$_e$]] exists in $w$ & for all $p_{<s, t>}$, if @ ∈ $p$ &

$p$ 'is about' [[John]], then $w \in p\}$

$= \{w_s \mid$ [[John$_e$]] exists in $w$ & for all $p_{<s, t>}$, if @ ∈ $p$ &

$p$ 'is about' [[John]], then $w \in p\} \cap \{w_s \mid w \in$ [[John whistles$_{<s, t>}$]]$\}$

$= \{w_s \mid w \in$ [[John runs$_{<s, t>}$]] & $w \in$ [[John whistles$_{<s, t>}$]]$\}$

Since the property 'whistles' is true of John at $w_1$, the strong type-<$s$, $t$> representations of John and 'John whistles' at $w_1$ are the same semantic object.

In light of our previous considerations, the identity of rich local representations of entities and their associated true propositions is arguably desirable. However, our strategy for the rich representation of propositions in the type <$s$, $t$> also allows the modeling of information growth.

To see this, consider the representation of the proposition 'John is bald' at $w_1$ (in (16)). The latter is obtained through the *extension* of the set of true propositions about John at $w_1$ by the information encoded in the proposition 'John is bald'. The latter corresponds to the *elimination* of those indices from the set of indices from (15) at which the proposition 'John is bald' is either false or undefined:

(16) $\{w_s \mid$ [[John$_e$]] exists in $w$ & for all $p_{<s, t>}$, if @ ∈ $p$ &

$p$ 'is about' [[John]], then $w \in p\} \cap \{w_s \mid w \in$ [[John is bald$_{<s, t>}$]]$\}$

$= \{w_s \mid w \in$ [[John runs$_{<s, t>}$]] & $w \in$ [[John whistles]] & $w \in$ [[John is bald]]$\}$

The above suggests the representation of (type-<$e$, <$s$, $t$>>) properties of entities at a given index by type-<<$s$, $t$>, <$s$, $t$>> functions from the local representation of entities in the property's domain (here, the $w_1$-specific representation of John from (15)) to the local representation of the result of attributing the property to the relevant entity in its domain (here, the $w_1$-specific representation of the result of attributing 'is bald' to John, cf. (16)). The latter corresponds to the result of obtaining the type-<$s$, $t$> representation of John at a better-defined index $w_3$, that distinguishes itself from $w_1$ at most with respect to the obtaining of the proposition 'John is bald'.

Our considerations from the preceding paragraphs show that the possibility of modelling information growth in single-type semantics is conditional on the existence of a definedness order on indices: If all indices in the domain of the type $s$ are totally defined (such that, for all indices $w$ and propositions $p$, either $w \in p$ or $w \in \neg p$), all single-type representations of extensional properties are associated with *improper* extensions, that send single-type representations of entities at some index to themselves (if the entity has the property at the respective index) or to the empty set of indices (otherwise). But this makes it impossible to capture the informativeness of propositions in single-type semantics.

To prevent the triviality of type-<$s$, <$s$, $t$>> representations of propositions, we require the existence of underdefined indices (so-called *partial* possible worlds, or possible *situations* (Barwise and Perry, 1983; Kratzer, 1989)), that can be extended into totally defined *possible worlds*. In virtue of the above, the type <$s$, <$s$, $t$>> is only suitable as the semantic basis for natural language if its objects are associated with functions from *partial* indices to functions from *partial* indices to *partial* truth-values (or with functions from *partial* indices to functions from *total* indices to *total* truth-values, see below).

In contrast to the above, the suitability of the single basic type <$s$, $t$> is not conditional on the adoption of partial indices or truth-values. This is due to the possibility[7] of representing

[7] This holds modulo the absence of an algebraic structure on objects of this type (cf. Req. 2).

entities and propositions via (minimally informative) indices, and of representing partial indices via sets of their extending total indices. The latter is a corollary of Stone's Theorem (Stone, 1936), (cf. Davey and Priestley, 2002). The possibility of representing partial indices via sets of total indices also justifies the possibility of associating representations of entities and propositions in the type $<s, <s, t>>$ with functions from *partial* indices to functions from *total* indices to *total* truth-values. Note, however, that the resulting representations no longer satisfy the intensionality requirement from Property 3.

This completes our identification of the suitable single basic types. We close the paper by comparing Partee's preliminary basic-type choice and with our newly identified single basic types.

## 4. Single-Type Candidates and Partee's Hypothesis

The above considerations disclose three interesting facts about Partee's preliminary single-type candidate:

1. Partee's chosen type (i.e. *properties of situations*, type $<s, t>$) is a suitable single basic type that satisfies the requirements from Properties 0–5.

2. Partee places more semantic constraints on single-type objects than necessary. Granted her disregard of intensionality, Montagovian *properties of possible worlds* are equally suitable.

3. Partee neglects an alternative type (for *situation-to-proposition functions*, type $<s, <s, t>>$), whose objects are semantically 'richer' than type-$<s, t>$ objects.

The observations from items (1) to (3) support Partee's basic-type choice. However, they point out the possibility of adhering more closely to Montague's original ontology (cf. the adoption of possible worlds; (2)), and of not prematurely excluding competing candidates (3). More generally, the possibility of representing partial situations via sets of their extending possible worlds (in Sect. 3.4) emphasizes the role of semantic operations and representational strategies in ontology (as opposed to the identity of the different object types).

**Kristina Liefke**

Munich Center for Mathematical Philosophy, Ludwig-Maximilians-Universität, München &
Tilburg Center for Logic and Philosophy of Science, Universiteit Tilburg
Kristina.Liefke@lrz.uni-muenchen.de

## References

Barwise, J., and J. Perry. 1983: *Situations and Attitudes*. Cambridge (MA): The MIT Press.

van Benthem, J. 1991: 'General Dynamics', *Theoretical Linguistics*, 159–201.

Carstairs-McCarthy, A. 1999: *The Origins of Complex Language: An Inquiry into the Evolutionary Beginnings of Sentences, Syllables, and Truth*. Oxford and New York: Oxford University Press.

Cheney, D. L., and R. M. Seyfarth. 1990: *How Monkeys See the World: Inside the Mind of Another Species*. Chicago: University of Chicago Press.

Chierchia G., and R. Turner. 1988: 'Semantics and Property Theory', *Linguistics and Philosophy* 11, 261–302.

Church, A. 1940: 'A Formulation of the Simple Theory of Types', *Journal of Symbolic Logic* 5, 56–68.

Cresswell, M. J. 1985: *Structured Meanings: The Semantics of Propositional Attitudes*. Cambridge (MA): The MIT Press.

Davey, B. A., and H. A. Priestley. 2002: *Introduction to Lattices and Order* (2nd ed.). Cambridge and New York: Cambridge University Press.

Frege, G. 1891: 'Funktion und Begriff', in: *Gottlob Frege: Kleine Schriften*, Saarbrücken: Edition Classic Verlag Dr. Müller (2006), 125–42.

— 1892: 'Über Sinn und Bedeutung', in: *Gottlob Frege: Kleine Schriften*, Saarbrücken: Edition Classic Verlag Dr. Müller (2006), 143–62.

Gallin, D. 1975: *Intensional and Higher-Order Modal Logic with Applications to Montague Semantics*. Amsterdam: North Holland.

Kamp, H. 1981: 'A Theory of Truth and Semantic Representation', in J. Groenendijk, T. Janssen, and M. Stokhof (eds.): *Formal Methods in the Study of Language*, Amsterdam: Mathematical Centre, 233–65.

Kaplan, D. 1989: 'Demonstratives: An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and Other Indexicals', in J. Almog, J. Perry, and H. Wettstein (eds.): *Themes from Kaplan*, Oxford and New York: Oxford University Press, 481–563.

Kratzer, A. 1989: 'An Investigation into the Lumps of Thought', *Linguistics and Philosophy* 12, 607–53.

Landman, F. 1986: 'Pegs and Alecs, in: *Towards a Theory of Information: The Status of Partial Objects in Semantics*, Groningen-Amsterdam Studies in Semantics. Dordrecht: Foris Publications, 233–65.

Montague, R. 1970a: 'English as a Formal Language', in R. Thomason (ed.): *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London: Yale University Press, 188–221.

— 1970b: 'Universal Grammar', in R. Thomason (ed.): *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London: Yale University Press, 222–46.

— 1973: 'The Proper Treatment of Quantification in Ordinary English', in R. Thomason (ed.): *Formal Philosophy: Selected Papers of Richard Montague*, New Haven and London: Yale University Press, 247–70.

Muskens, R. 1995: *Meaning and Partiality*. CSLI Lecture Notes. Stanford: CSLI Publications.

— 1996: 'Combining Montague Semantics and Discourse Representation, *Linguistics and Philosophy* 19, 143–86.

Partee, B. 2006: 'Do We Need Two Basic Types?', in S. Beck und H.-M. Gärtner (eds.): *Snippets: Special Issue in Honor of Manfred Krifka* (Vol. 20). Berlin, 37–41.

Rantala, V. 1982: 'Quantified Modal Logic: Non-Normal Worlds and Propositional Attitudes', *Studia Logica* 41, 41–65.

Snedeker, J., J. Geren, and C. L. Shafto. 2007: 'Starting Over: International Adoption as a Natural Experiment in Language Development, *Psychological Science* 18, 79–87.

Stone, M. H. 1936: 'The Theory of Representation for Boolean Algebras', *Transactions of the American Mathematical Society* 40, 37–111.

# Relevanz anstatt Wahrheit?

Theresa Marx

In diesem Paper werde ich entgegen der Auffassung der Relevanztheoretiker dafür argumentieren, dass die Gricesche Maxime der Qualität nicht durch das allgemeine Relevanzprinzip ersetzt werden kann. Die Relevanztheorie kann keine befriedigenden Erklärungen für sprachliche Phänomene wie metaphorische oder ironische Verwendung von Äußerungen liefern, die dem strikten Wahrhaftigkeitsanspruch widersprechen. Wenngleich die Kritik am griceschen Programm partiell gerechtfertigt sein mag, stellt sie die Relevanztheorie vor weit größere explanatorische Schwierigkeiten hinsichtlich des Verhältnisses von kognitivem Aufwand und positivem kognitiven Effekt. Ebenso wenig kann man weiterhin von geglückter Kommunikation sprechen, wenn die Relevanzintention den Wahrhaftigkeitsanspruch verletzt, den wir als Adressaten an die Äußerung eines Sprechers stellen, und damit zu unerwünschten Resultaten führt. Ich plädiere daher für eine Revision der Relevanztheorie unter Berücksichtigung der wesentlichen Rolle des Wahrhaftigkeitsanspruchs in der normalsprachlichen Konversation.

## 1. Ausgangspunkt: Die Griceschen Maximen

Die Relevanztheorie, begründet 1986 von Dan Sperber und Deirdre Wilson[1], präsentiert sich als Alternative zu den Griceschen Konversationsmaximen. Paul Grice hatte mit den vier Maximen, die ich kurz erläutern werde, zu beschreiben versucht, welchen unausgesprochenen Regeln rationale und kooperative Gesprächsteilnehmer folgen, um den Erfolg der Kommunikation zu gewährleisten: Ein Sprecher soll sich darum bemühen, genau so viel wie nötig zu sagen, also nicht mehr und nicht weniger als zum Vermitteln der Information im gegebenen Kontext erforderlich ist (Maxime der Quantität). Das Gesagte soll in einem angemessenen Bezug zum Kontext der Kommunikation stehen, sich also beispielsweise nicht auf etwas beziehen, von dem der Adressat keine Ahnung haben kann, das keinen Bezug zu einer gestellten Frage hat etc. (Maxime der Relevanz). Des Weiteren muss eine gewisse Klarheit des Ausdrucks gewährleistet sein, um einen erfolgreichen Beitrag zur Kommunikation leisten zu können, das heißt, es werden überflüssige Mehrdeutigkeiten oder Ungeordnetheit des Gesagten vermieden (Maxime der Modalität). Meines Erachtens fällt unter diese Maxime auch das Bemühen, eine Sprache zu verwenden, die der Gesprächspartner verstehen kann (also nicht nur das Vermeiden unverständlicher Fremdwörter, sondern auch bspw. die Verwendung von *easy English* im Gespräch mit Nicht-Muttersprachlern).

Die drei bisher genannten Maximen lassen sich recht problemlos unter dem Relevanzprinzip zusammenfassen. Um es in der Sprache der Relevanztheoretiker auszudrücken: Ein Sprecher, der einen Akt ostensiver Kommunikation vollführt, hat die Absicht, dem Hörer ein Set von Annahmen zu vermitteln, das ihm relevant genug erscheint, um dem Hörer den für seine Verarbeitung notwendigen kognitiven Aufwand zuzumuten. Um das Optimum an Relevanz zu erreichen, muss ein ideales Verhältnis zwischen kognitivem Aufwand und positivem kontextuellen Effekt bestehen, der wiederum von der relativ leichten Zugänglichkeit der vom Sprecher im gegebenen Kontext vermittelten Annahmen abhängt. Solch ein Prinzip schließt

[1] Sperber und Wilson (1986): *Relevance: Communication and Cognition*, Blackwell, Oxford.

Redundanz ebenso aus wie Unklarheit und thematische Irrelevanz und kann daher als vollwertige Alternative zu den genannten drei Maximen gelten.

Anders verhält es sich jedoch mit der Maxime der Qualität, des Anspruchs an den Gesprächspartner, einen wahrheitsgemäßen Beitrag zu liefern, die Grice in zwei Untermaximen zerlegt: Sage nichts, das du für falsch hältst (Maxime der Wahrhaftigkeit), und sage nichts, für das du keine ausreichenden Belege hast (im Orginal *Maxim of Evidence*; mangels besserer Übersetzung spreche ich hier von der Maxime der fundierten Annahme).

Die Übermaxime der Qualität hat für Grice eine besondere Bedeutung:

> The maxim of Quality, enjoining the provision of contributions which are genuine rather than spurious (truthful rather than mendacious), does not seem to be just one among a number of recipes for producing contributions; it seems rather to spell out the difference between something's being and (strictly speaking) failing to be, any kind of contribution at all. False information is not an inferior kind of information; it just is not information (Grice 1989: 371).

Wenn es sich um die Erzeugung von Implikaturen handelt, spielt die Maxime der Qualität laut Grice eine ähnliche Rolle wie die drei anderen Maximen; er hält sie jedoch durchaus für eine Art Supermaxime, da die anderen drei nur in Kraft treten können, wenn zunächst die Maxime der Qualität erfüllt wurde.[2] Falls sie nicht erfüllt wird, kann der kommunikative Beitrag nicht als solcher gesehen werden und auch keine Information vermitteln.

Diese Hypothese erscheint mir allerdings etwas übertrieben. Wir können durchaus Beispiele konstruieren, die zeigen, dass man auch durch nicht wahrheitsgemäße Aussagen einen wichtigen Beitrag zum Gespräch leisten kann. Nehmen wir an, ich möchte Ihnen, der Sie noch nie vom Arabischen Frühling gehört haben, etwas über die Situation in Ägypten mitteilen, bin aber selbst entweder schlecht informiert oder möchte Ihnen absichtlich die Unwahrheit sagen, und äußere folgenden Satz:

(1) Nach dem Sturz Mubaraks 2011 wurde in Ägypten die Diktatur des Proletariats errichtet.

Wenn wir diesen Satz insgesamt betrachten, müssen wir feststellen, dass er dem Adressaten sogar eine ganze Menge an Informationen liefert. Wir können uns auch vorstellen, dass dieser nur aus dem ersten Teil des Satzes Konsequenzen zieht, sodass er sein Weltbild angemessen korrigiert, ohne durch den unwahren Teil nennenswert beeinträchtigt zu werden. Zwar lehne ich eine konsequenzialistische Konzeption der Wahrheit entschieden ab, wie wir noch sehen werden, dennoch zeigt dieses Beispiel, dass auch falsche Information einen gewissen kognitiven Wert für den Adressaten haben kann. Sogar, wenn dieser nur den erlogenen Teil in seine Sicht der Welt integrieren würde, enthielte dieser immer noch wahre Informationen (z.B: es existiert – weiterhin – ein Staat namens Ägypten).

Wenn es also darum geht, dass dem Adressaten Informationen geliefert werden sollen, ist die Maxime der Qualität zwar durchaus von großer Wichtigkeit; sie ist dafür aber nicht von absoluter Bedeutung, wie Grice uns glauben machen möchte.

Natürlich ist sich Grice auch der Tatsache bewusst, dass es gewisse Arten von Aussagen gibt, denen es zwar an Wahrheitsgehalt mangelt, die wir aber trotzdem intuitiv nicht als Lügen bezeichnen würden, nämlich immer dann, wenn der Sprecher überhaupt nicht vorgibt, der Maxime der Wahrhaftigkeit zu folgen. Diesen Fall finden wir in sprachlichen Phänomenen wie dem Erzählen von fiktiven Geschichten oder Witzen, sowie in sprachlichen Stilmitteln wie der Verwendung von Metaphern oder Ironie.

Laut Grice kann man den Gebrauch von Ironie durch eine Aussetzung (*suspension*) der Maxime der Wahrhaftigkeit erklären. Während der Aufführung eines Theaterstücks oder dem

[2] Grice 1989: 27.

Erzählen eines Witzes tut keiner der Sprecher so, als ob er dieser Maxime folge, und, was entscheidend für das Gelingen seines Sprechakts ist, sein Publikum ist sich dessen bewusst.

Im Falle der genannten (und ähnlicher) Stilmittel findet hingegen eine offene Verletzung (*violation*) der Wahrhaftigkeitsmaxime statt (im Gegensatz zum bewussten Lügen mit täuschender Absicht: dabei handelt es sich um eine verdeckte Verletzung der Maxime). Der Sprecher will in diesem Fall weiterhin einen wahren Gedanken ausdrücken, aber er formuliert diesen Gedanken so, dass er der Wahrhaftigkeitsmaxime nicht strikt zu folgen scheint. Betrachten wir die folgenden Beispiele für dieses Phänomen:

(2) *Das Kapital* von Marx zu lesen, ist das Lustigste, was ich je getan habe.

(3) Ich habe dich tausendmal angerufen!

Im ersten Fall verwendet der Sprecher Ironie: er möchte einen Gedanken ausdrücken, der dem Gegenteil des Gesagten entspricht, nämlich dass er sich beim Lesen furchtbar gelangweilt hat. Im zweiten Fall wird eine Hyperbole verwendet, um einen etwas schwächeren Gedanken als den tatsächlich geäußerten auszudrücken, nämlich dass der Sprecher den Adressaten so oft angerufen hat, dass ihm die Häufigkeit der Anrufe bemerkenswert erscheint.

Die Maxime der Qualität kann also offen verletzt oder vorübergehend außer Kraft gesetzt werden. In beiden Fällen wird dies aber konventionell angezeigt, sodass dem Sprecher nicht unterstellt werden kann, die Unwahrheit gesagt zu haben. Es besteht auch kein Zweifel daran, dass die Maxime im Falle der Fiktion sofort wieder in Kraft tritt, wenn die Erzählung beendet ist, beziehungsweise dass sie, im Falle der Stilmittel, weiterhin angewandt und die Wahrhaftigkeit des Gedankens durch Implikaturen kommuniziert wird.

## 2. Kritik an der Erklärungskraft der Maxime der Qualität

Als Mitbegründerin und überzeugte Vertreterin der Relevanztheorie versucht Deirdre Wilson in ihrem Paper von 1995 Erklärungen für solche sprachlichen Phänomene zu finden, ohne dass dazu die Maxime der Qualität herangezogen werden müsste. Laut Relevanztheorie versucht der Sprecher einen Gedanken auf möglichst relevante Weise auszudrücken, das heißt, indem er dem Adressaten so wenig kognitiven Aufwand wie möglich zumutet, ihm aber gleichzeitig einen möglichst großen positiven kognitiven Effekt verschafft (also beispielsweise das Übermitteln einer für den Adressaten wichtigen Information in einfachen und verständlichen Worten, deren Sinn sich im Kontext leicht erschließt).

Weshalb aber haben sich dann in den natürlichen Sprachen Kommunikationsstrategien, wie die Verwendung von Ironie, herausgebildet, die nicht diesem Prinzip folgen?

Wilson erwähnt zwei weitere Fälle, die der Maxime der Qualität zu widersprechen scheinen: Zum einen den sogenannten *loose talk,* also die nicht wörtlich zu nehmende Rede, die in Aussagen vorkommt wie

(4) Holland ist flach

und zum anderen die freie indirekte Rede, die wir verwenden, wenn wir das von einer anderen Person Gesagte so wiedergeben als handelte es sich dabei um unsere eigene Äußerung:

(5) Gestern habe ich Eros getroffen. Er arbeitet an einem unglaublich wichtigen Forschungsprojekt.

Diese Phänomene des Sprachverhaltens werden nicht zu den Stilmitteln gerechnet, stellen uns aber vor ähnliche Probleme wie der Gebrauch von Metaphern oder Ironie, da die Maxime der Qualität streng genommen verletzt wird.

Wilson schlägt zwei alternative Erklärungen für diese Phänomene vor, die vom Aussetzen oder Verletzen der Maxime der Qualität keinen Gebrauch machen. Im Falle der Metaphern, Hyperbolen und *loose talk* handele es sich um ungefähre Annäherungen (*rough approximation*) an das tatsächlich Gemeinte; selbiges wird also zwar nicht wörtlich kommuniziert, das Gesagte ähnelt aber hinreichend dem Gedanken, den der Sprecher ausdrücken will, sodass der Adressat das Gemeinte ohne Probleme erschließen kann. Das Konzept der ungefähren Annäherung kann so erklären, wie Äußerungen wie (4) wahr sein können, obwohl sie dies nicht im wörtlichen Sinne sind (wir können also beispielsweise sagen, dass sich „flach“ in (4) auf die relative Flachheit von Landschaften bezieht). Die gricesche Maxime wird daher, laut Wilson, nicht verletzt, sondern abgeschwächt.

Obwohl dieses Konzept meines Erachtens durchaus erklärerisches Potential hat, kann ich nicht erkennen, inwieweit es die Maxime der Qualität ausschließen bzw. ersetzen soll. Um den Gedanken auszudrücken, dass Holland für eine Landschaft relativ flach ist, und damit die gewünschten kontextuellen Effekte zu erreichen (z.B. die konversationale Implikatur, dass Holland ein geeigneter Ort für eine Fahrradtour wäre), ist der Satz (4) optimal geeignet. Tatsächlich könnten wir innerhalb des griceschen Theoriegefüges sagen, dass sich der Sprecher so gut wie nur möglich ausdrückt, nämlich indem er gleichzeitig der Maxime der Quantität und der Qualität folgt. Man muss sich bewusst machen, dass die natürlichen Sprachen in weiten Teilen aus dem Gebrauch von Metaphern bestehen; aber da wir es eben so gelernt haben, sind wir mit ihrem Gebrauch weiterhin wahrhaftig in einem pragmatischen Sinne. Es kann keinen vernünftigen Zweifel daran geben, dass die folgenden Sätze wahre Gedanken ausdrücken und nicht in Konflikt mit der Maxime geraten:

(6) Die Sonne versteckt sich hinter den Wolken. (Metapher)

(7) Das Glas ist halb voll. (Annäherung)

Ein rationaler Sprecher der (6) oder (7) äußert, hat im Normalfall die Absicht, etwas Wahres zu sagen und gute Gründe für seine Annahme, dies damit auch zu tun; daher kann ich nicht sehen, inwieweit die Maxime abgeschwächt wird, wie Wilson behauptet.

Betrachten wir aber zunächst noch ihre alternative Erklärung für die Verwendung von Ironie. Ihrer Auffassung nach verstehen wir solche Äußerungen wie Fälle von indirekter freier Rede: wir tun so, als habe eine andere Person den Satz geäußert, wenn wir ihn interpretieren. Wir suchen also einen anderen möglichen Sprecher, der den Satz geäußert haben könnte, und distanzieren uns damit vom Wahrheitsgehalt unserer Aussage, sodass klar wird, dass wir das Gegenteil kommunizieren möchten. In Sätzen wie (5) mag diese Erklärung durchaus ihre Rechtfertigung haben. Je nach Tonfall des Sprechers können wir schließen, dass er Eros' Projekt möglicherweise für nicht allzu bedeutsam hält und sich über ihn lustig macht. Allerdings ist (5) auch ein besonderer Fall: es werden gleichzeitig freie indirekte Rede *und* Ironie gebraucht. Auch nicht ironisch gemeinte Äußerungen, die freie indirekte Rede beinhalten, können problemlos so erklärt werden, dass die Worte eines anderen Sprechers wiedergegeben werden und der Adressat dies inferiert, indem er dem Relevanzprinzip folgt (Grice beruft sich wieder auf eine offene Verletzung der Maxime).

Das heißt nur leider noch lange nicht, dass dies auch auf Fälle zutrifft, in denen Ironie in direkter Rede gebraucht wird, wie in (2). Mir fällt spontan kein möglicher Sprecher ein, der allen Ernstes behaupten möchte, er habe noch nie etwas Lustigeres getan als *Das Kapital* zu lesen. Außerdem erfordert die Suche nach einem möglichen Sprecher in solchen – typischen, nicht mühsam konstruierten – Fällen einen weitaus höheren kognitiven Aufwand als es bei einer offenen Verletzung der Maxime der Qualität und damit der Schlussfolgerung auf eine gegenteilige Intention des Sprechers der Fall ist. Intuitiv erfassen wir, dass, wenn ein Sprecher etwa offenkundig Falsches sagt, er der gegenteiligen Auffassung ist. Auch das frühkindliche Erlernen von Ironie kann durch Grice, im Gegensatz zu Wilson, befriedigend erklärt werden. Stellen wir uns vor, das Kind zerbricht unachtsam im Spiel eine Lampe und die Mutter äußert daraufhin

(8) Na super. Mach nur so weiter.

An Tonfall und Mimik lässt sich leicht erkennen, dass die Mutter nicht wirklich begeistert ist. Welchen Sprecher aber soll sich das Kind denn nun vorstellen, der sich tatsächlich über seine Unachtsamkeit freut und es sogar auffordert, damit fortzufahren? Anstatt einen solchen Sprecher zu suchen, wird das Kind intuitiv verstehen, dass die Mutter ihre Verärgerung mit anderen Worten ausdrückt, z.B. um ihre Ermüdung angesichts des ständigen Bedarfs an Schimpftiraden zu unterstreichen.

Wilsons Ansatz zur Erklärung der Ironie ist nicht nur unangemessen kompliziert und widerspricht damit dem eigenen theoretischen Ausgangspunkt, er ist schlichtweg absurd.

Dennoch widerspricht der Gebrauch von Ironie, der einen höheren kognitiven Aufwand von Seiten des Adressaten erfordert, nicht zwangsläufig dem Relevanzprinzip. Wir müssen nur das Konzept eines positiven kognitiven Effekts erweitern: dieser muss nicht immer unbedingt ein epistemischer sein, sondern kann Eigenschaften aufweisen, die andere kommunikative Absichten des Sprechers und Adressaten befriedigen, beispielsweise ein unterhaltsamerer Redner zu sein oder sich der eigenen Intelligenz zu erfreuen. Die Motivation, Ironie zu gebrauchen, ist in vielen Fällen vergleichbar mit unserem Interesse an Kreuzworträtseln oder Fernsehsendungen wie *Wer wird Millionär?*. Die Ironie bewirkt auch ähnliche kognitive Effekte wie fiktive Geschichten oder das Erzählen von Witzen, obwohl weiterhin ein Gedanke kommuniziert wird, den der Sprecher für wahr hält.

Die Maxime der Qualität wird also weder außer Kraft gesetzt noch abgeschwächt, sondern auf überlegte und ostentative Weise verletzt, um verschiedene positive kognitive Effekte für den Adressaten zu erzielen.

Dieses Konzept des positiven kognitiven Effekts geht natürlich weit über den Relevanzanspruch hinaus, wie ihn Sperber und Wilson formuliert haben. Es soll hier außerdem angemerkt werden, dass der Gebrauch von Ironie ohne die Maxime der Qualität wohl überhaupt nicht möglich wäre. Gerade weil wir annehmen, dass uns der Sprecher etwas Wahres mitteilen will, suchen wir einen passenden Gedanken, der durch die ironische Äußerung ausgedrückt wird, und es freut uns nicht im Geringsten, wenn wir feststellen müssen, dass dieser Gedanke nicht wahr ist. Stellen wir uns vor, ein bekanntermaßen politisch Konservativer äußert auf ironische Weise folgenden Satz:

(9) Was für ein Glück, dass es in Ägypten jetzt eine Diktatur des Proletariats gibt.

In diesem Fall können wir nicht von einer offenen Verletzung der Untermaxime der Wahrhaftigkeit sprechen (obwohl diese Intention besteht), noch von ihrer verdeckten Verletzung (der Sprecher intendiert nicht, die Unwahrheit zu sagen), sondern von der Verletzung der zweiten Untermaxime, da der zu inferierende Gedanke („was für ein Unglück, dass...“) etwas Falsches beinhaltet. Somit kann (9) nicht als geglückter Beitrag zur Kommunikation gewertet werden (obwohl er, wie im ersten Abschnitt erörtert, immer noch informativ sein kann!), und das Versagen des Sprechaktes im kommunikativen Kontext zeigt sich gewöhnlich an den Reaktionen der Adressaten: mangelndes Verständnis, Korrekturversuche, oder gar Verachtung gegenüber dem falsch informierten Sprecher. Ohne die Maxime der Qualität kann weder der Erfolg noch das Scheitern der Ironie erklärt werden.

Wie wir gesehen haben, bietet das Gricesche Programm durchaus befriedigende Erklärungen für den Gebrauch von Ironie; wenn wir das Konzept des positiven kognitiven Effekts entsprechend erweitern, kann sogar eine der Relevanztheorie sehr ähnliche Erklärung funktionieren. In keinem Fall aber besteht die Notwendigkeit, auf eine komplizierte Drittsprecher-Theorie zurückzugreifen, noch gibt es ausreichende Gründe dafür, die Maxime der Qualität abzuschaffen.

## 3. Relevanz vor Wahrhaftigkeit: Das Beispiel Uhrzeit

In ihrem Paper von 2002 argumentieren Van der Henst, Carles und Sperber ebenfalls für die Abschaffung der Maxime der Qualität zugunsten des allgemeinen Relevanzprinzips. Die drei

Vertreter der Relevanztheorie versuchen anhand des Beispiels der Angaben von Uhrzeiten zu zeigen, dass rationale Sprecher natürlicherweise eher geneigt sind, sich möglichst relevant für den Adressaten anstatt strikt wahrheitsgemäß auszudrücken.

Wenn wir im Alltag nach der Uhrzeit gefragt werden, neigen wir dazu, eher gerundete Zeiten anzugeben: Wir sagen „es ist zehn vor sechs“ und nicht „es ist 17.52 Uhr“. Dabei ist natürlich nicht davon auszugehen, dass wir absichtlich lügen. Die Autoren erklären dieses Phänomen damit, dass man zwischen der wörtlichen Aussage und dem Resultat ihrer Verarbeitung unterscheiden muss, also den Konsequenzen, die eine Aussage im Geist des Adressaten auslöst. Die wörtliche Bedeutung der Aussage „es ist zehn vor sechs“ ist zwar falsch, die Konsequenzen, die der Adressat daraus zieht, sind allerdings die richtigen (beispielsweise, dass er noch Zeit für einen Kaffee hat).

Ein Konzept der Wahrhaftigkeit im strikten Sinne lässt sich unter solchen Bedingungen natürlich nicht aufrecht erhalten, noch besteht die Notwendigkeit dazu, sondern man spricht in einem solchen Fall von der Wahrheit der Konsequenzen. Es soll an dieser Stelle noch einmal betont werden, dass es sich dabei nicht um Konsequenzen im Sinne physischer Aktionen handelt, sondern um die Annahmen, die die Adressaten aus dem Gesagten schließen. Die Bedeutung einer Äußerung ist, in Einvernehmen mit dem griceschen Programm, die Intention des Sprechers, und diese Intention richtet sich auf die Konsequenzen, die seine Aussage im Geist des Adressaten provozieren wird (diese Konsequenzen sind auch identisch mit dem Informationsgehalt der Aussage).

In der besagten Kommunikationssituation, nämlich einem Unbekannten die Uhrzeit zu nennen, gibt es zwei entgegengesetzte Bestrebungen, die für den tatsächlichen Sprechakt verantwortlich sind: Der Sprecher möchte zum einen seinen eigenen Aufwand möglichst gering halten, zum anderen aber dem Adressaten eine möglichst relevante Information vermitteln. Dies zeigt sich auch experimentell: In den von den Autoren durchgeführten Experimenten gaben 97% der Träger von Analoguhren die gerundete Zeit an, wobei sie den eigenen Aufwand sowie auch den des Adressaten auf ein Minimum beschränkten; allerdings, und das ist wesentlich interessanter, gaben immer noch 57% der Digitaluhrträger eine gerundete Zeit an – das heißt, sie wichen bewusst von der strikten Wahrhaftigkeit ab, um ein Optimum an Relevanz für den Adressaten zu erzielen! Es scheint also, als ob Relevanz in der realen Kommunikation tatsächlich eine wichtigere Rolle spielt als die strikte Wahrheit.

Andere von den Autoren durchgeführte Experimente beinhalten das Fragen nach der Zeit um eine Uhr zu stellen sowie in Bezug auf einen zeitlich strikt festgelegten Termin. In beiden Fällen zeigte sich, dass die befragten Personen bemüht waren, so relevant wie möglich für den Fragesteller zu antworten: Sie hörten in weiten Teilen auf, die Uhrzeit auf- oder abzurunden und gaben stattdessen auf die Minute genaue Auskünfte.

Solche Ergebnisse scheinen die Theorie der Autoren zu stützen, dass Relevanz tatsächlich in der alltäglichen Kommunikation eine wesentlich größere Bedeutung als Wahrhaftigkeit hat. Daher wenden sie sich gegen die Maxime der Qualität und plädieren stattdessen für eine Art Wahrheit der Konsequenzen (also der Annahmen, die der Adressat schlussfolgert).

Sperber und Wilson 1995 vertreten eine ähnliche Ansicht:

> ...truth of the conclusions seems more crucial to relevance than truth of the premises.[3]
>
> If only true inputs were relevant, we would have to say that fictions were irrelevant. If truth of the output is what matters, then fictions can be relevant after all.[4]

Tatsächlich kann im speziellen Fall der Uhrzeitangaben der Exaktheitsgrad der Wahrhaftigkeit zugunsten der Relevanz vernachlässigt werden (und zeigt damit noch einmal,

[3] Sperber und Wilson 1995: 264.

[4] Ebd.: 265.

dass die Maxime der Qualität eben nicht allen anderen Maximen übergeordnet sein muss, wie Grice uns glauben machen möchte). Es handelt sich wiederum um einen Fall von sprachlichem Pragmatismus, ebenso wie in (4).

Dieser spezielle Fall liefert uns allerdings noch keine Rechtfertigung für die vollständige Abschaffung der Maxime, noch erlaubt er uns, künftig ein Konzept der Wahrheit als Konsequenz zu vertreten.

Betrachten wir die folgenden Beispiele:

(10) Arzt: „Sie haben Krebs im Endstadium."

Durch die Äußerung des Arztes zieht der Patient die Konsequenz, dass ihm nicht mehr viel Lebenszeit bleibt; entsprechend ändert er sein Verhalten (er macht sein Testament, verabschiedet sich von Freunden und Familie, etc.).

Nun stellen wir uns jedoch vor, dass sich der Arzt geirrt hat: Der Patient hat keinen Krebs, sondern tatsächlich ein unheilbares Herzleiden, dass ebenfalls in kurzer Zeit zum Tode führt. Obwohl die Annahmen, die der Patient als Konsequenz aus der Äußerung des Arztes zieht, korrekt sind (in beiden Fällen bleibt ihm nicht mehr viel Zeit), hat der Arzt offensichtlich nicht die Wahrheit gesagt. Er hat die zweite Untermaxime verletzt, indem er etwas behauptete, für das er nicht die entsprechenden Belege hat. Sollte sich nach dem Tod des Patienten herausstellen, dass die Diagnose falsch war, wird dies für den Arzt sicherlich Folgen haben. Obwohl der Arzt etwas für seinen Adressaten vollkommen Relevantes geäußert hat und dieser daraufhin die richtigen Schlüsse gezogen hat, wären wir über die Fehldiagnose empört, und unser Rechtssystem würde den Arzt abstrafen.

Ähnlich liegt der Fall bei

(11) Der Präsident der USA: „Wir haben Beweise dafür, dass das Land X die Atombombe entwickelt.

In Wirklichkeit liegen allerdings keine solchen Beweise vor; der Präsident äußert den Satz nur, um andere Ziele zu verfolgen, beispielsweise, um durch einen Krieg an die Ölreserven des Landes X zu gelangen. Doch nachdem die USA in besagten Krieg eingetreten sind, stellt sich heraus, dass im Land X tatsächlich Experimente durchgeführt werden, die auf das Entwickeln von Nuklearwaffen abzielen. Bedeutet das, dass der Präsident nicht gelogen hat? Intuitiv lehnen wir es ab, eine Äußerung zu akzeptieren, die als Lüge gemeint war, nur weil sie sich später als wahr entpuppt. Der Arzt in (19) hat nicht absichtlich gelogen (und dennoch kann er bestraft werden); der Präsident in (11) hat jedoch bewusst die Untermaxime der Aufrichtigkeit verletzt, was uns zu moralischer Entrüstung führt, mit all ihren politischen Konsequenzen, obwohl seine Äußerung höchst relevant wahr. Wir könnten sogar sagen, dass wir besser dran sind, weil wir die Lüge geglaubt haben, da sie uns ja dazu bewegt hat, die gefährliche Wahrheit zu entdecken; die Äußerung war informativ, obwohl sie als Lüge gedacht war, um noch einmal auf meine Kritik an Grice zurückzukommen. Dennoch würden wir uns nicht mit der reinen Relevanz und Informativität der Aussage zufrieden geben – denn offenbar ist uns die Wahrhaftigkeit des Sprechers wichtiger.

Man könnte mich an dieser Stelle dahingehend kritisieren, dass ich Ethik und Sprachphilosophie vermenge; allerdings sollte meiner Meinung nach ein pragmatisch-deskriptiver Ansatz möglichst alle Aspekte unseres Sprachgebrauchs erfassen, wozu auch unser moralischer Anspruch an sprachliche Äußerungen gehört.

Wir können uns problemlos weitere Fälle vorstellen, in denen aus Lügen wahre Konsequenzen gezogen werden (und natürlich auch das Gegenteil: dass aus wahren Äußerungen falsche Konsequenzen gezogen werden, z.B. aufgrund von Missverständnissen). Doch obwohl der Zustand der tatsächlichen Welt in (10) und (11) die aus den Falschaussagen gezogenen Konsequenzen rechtfertigt, besteht für uns kein Zweifel daran, dass der Arzt einen

schweren Fehler begangen hat und vielleicht für seinen Beruf nicht geeignet ist, und dass der Präsident sein Volk belügt und seines Amtes enthoben werden sollte. Das sind sozusagen moralische Konsequenzen, die direkt aus der Verletzung der Maxime der Qualität resultieren, in einem Fall aufgrund mangelnder Belege, im anderen aufgrund einer verdeckten Verletzung der Untermaxime der Wahrhaftigkeit.

Die Aufrichtigkeit des Sprechers ist uns wichtig, unabhängig von der tatsächlichen Relevanz des Gesagten, und wir erwarten grundsätzlich, dass er dieser Norm folgt. Wir können keine Lügen als kooperatives Verhalten akzeptieren, nur weil aus ihnen die richtigen Konsequenzen folgen (d.h., die Interpretation des Hörers sich als wahr erweist). Daher plädiere ich für die Aufrechterhaltung des Wahrheitsanspruches an das vom Sprecher Ausgedrückte anstatt einer Wahrheit der Konsequenzen.

Wenn wir etwas fragen, erwarten wir nicht nur eine relevante, sondern vor allem eine wahrhaftige Antwort. In vielen Fällen erwarten wir außerdem, dass auch die andere Untermaxime befolgt wird: der Sprecher soll uns etwas mitteilen, wenn er gute Gründe dafür hat, dies auch für wahr zu halten. Dieser Anspruch zeigt sich deutlich bei Fragen nach dem Wissenstand unseres Gesprächspartners. Eine Antwort auf die Frage „Was studiert Peter?“ kann für uns nur relevant sein, wenn der Gefragte aufrichtig antwortet. Außerdem gehen wir davon aus, dass Antworten wie „Peter studiert Physik“ auf Grundlage adäquater Beobachtungen geäußert werden (und nicht aufgrund schwacher Inferenzen, wie z.B. dass Peter für den Sprecher ein typisches Physikergesicht hat).

Wenn die Relevanzintention für den Sprecher an erster Stelle steht, kann dies sogar zu regelrechten Falschaussagen führen, wenn bspw. nach Meinungen oder Entscheidungen gefragt wird.

Betrachten wir mein letztes Beispiel:

(12) Rosa: „Ich wollte mir zum Mittagessen eine Sardine zubereiten, die ich im Kühlschrank habe. Meinst du, die ist noch gut?“

Theresa, *die keine Ahnung vom Verfallsdatum frischen Fisches hat*: „Wie lange ist sie denn schon da drin?“

Rosa: „Nur zwei Tage.“

Theresa: „Dann ist sie bestimmt noch gut.“

Jeden Tag finden wir uns in Konversationen wieder, in denen wir uns ungewollt in Experten verwandeln, nur weil wir weiterhin relevant und kooperativ kommunizieren wollen. In Beispiel (12) hätte Theresa das Gespräch direkt nach Rosas anfänglicher Frage beenden können, indem sie zugibt, von dem Thema keine Ahnung zu haben. Stattdessen stellt sie eine ihr relevant erscheinende Frage – und tatsächlich ist die Frage auch durchaus relevant, denn hätte Rosa „zwei Stunden“ oder „zwei Jahre“ geantwortet, hätte selbst jemand wie Theresa Rosas erste Frage wahrheitsgemäß beantworten können.

Allerdings hat Theresas Frage noch einen weiteren Effekt: Indem sie etwas Relevantes fragt, gibt Theresa zu, dass sie zumindest einen gewissen Grad an Wissen über die Materie besitzt, und sie lässt sich anschließend dazu hinreißen, ihr Expertentum durch ihr abschließendes Urteil zu bekräftigen, um nicht das Relevanzprinzip mit ihrer vorherigen Frage missachtet zu haben. Ihr Bedürfnis, etwas Relevantes zum Gespräch beizutragen, hätte sie wahrscheinlich zu einer vollkommen gegenteiligen Antwort bewogen, wenn Rosa das Wörtchen „nur“ durch „schon“ ersetzt hätte. Dann hätte Theresa, kooperativ wie sie im Dialog nun einmal ist, bestimmt ein negatives Urteil gefällt.

So aber nimmt sie in Kauf, dass Rosa an einer Fischvergiftung zugrunde geht.

Dieser Dialog, der wörtlich so stattgefunden hat, ist paradigmatisch für viele unserer alltäglichen Konversationen. Die Intention, relevant zu antworten, verführt uns dazu,

Aussagen über Gegenstände zu treffen, von denen wir kein wirkliches Wissen besitzen, sodass wir die Untermaxime der fundierten Annahme verletzen (weitere typische Beispiele wären „hier musst du dich links halten“, „Last-Minute-Flüge sind immer billiger“, „die Wirtschaftskrise ist schuld an der Arbeitslosigkeit in Spanien“,...). Solche Fälle erscheinen uns im Allgemeinen weniger problematisch als die Verletzung der Untermaxime der Wahrhaftigkeit (wir verzeihen Irrtümer, Lügen aber nur selten).

Es wäre interessant, festzustellen, welches Bedürfnis sich in der alltäglichen Kommunikation normalerweise durchsetzt, relevant zu sein oder fundiertes Wissen zu vermitteln, aber allein die Tatsache, dass wir meistens nicht blind auf die Exaktheit des Wissens unserer Gesprächspartner vertrauen, lässt mich vermuten, dass wir uns irgendwie dessen bewusst sind, dass eine Äußerung nicht unbedingt wahr sein muss, nur weil sie uns relevant erscheint. Deshalb vertrauen wir eher auf Experten als auf Laien, überprüfen die Route auf unserem Navigationssystem, vergleichen Flüge das ganze Jahr hindurch und lesen Zeitungen. Das schließt natürlich nicht alle möglichen Fehler aus (die arme Rosa könnte durchaus sterben), wir vertrauen häufig auch auf falsche Aussagen, aber es zeigt, dass es uns sehr wichtig ist, dass wir uns in wirklich wichtigen Fällen darauf verlassen können, dass die zweite Untermaxime befolgt wird (deshalb empört uns auch der unfähige Arzt).

Ich hoffe, dass meine Beispiele zeigen konnten, welche Probleme der normative Anspruch mit sich bringt, den Sperber und Wilson mit ihrer Relevanztheorie verbinden. Wir können uns nicht von einem reinen Relevanzprinzip leiten lassen, das komplett auf ein Konzept von Wahrheit verzichtet, weder wenn wir Sätze bilden, noch wenn wir etwas Gesagtes interpretieren.

Ich bezweifle nicht, dass die Relevanzintention in vielen Fällen die Intention mit sich bringt, etwas Wahres zu sagen, da dies schließlich im Normalfall die Information ist, die für den Hörer die größtmögliche Relevanz hat. Auch möchte ich nicht das größere explanatorische Potential der Relevanztheorie bei anderen Aspekten der sprachlichen Kommunikation leugnen, etwa wenn es um die Festlegung des Bezugs geht.[5] Wenn wir aber die fundamentalsten Aspekte des Wesens unserer Kommunikation erklären wollen, müssen wir feststellen, dass wir vor allem anderen erwarten, dass unsere Gesprächspartner uns wahre Informationen liefern. Erst nachdem diese Bedingung erfüllt ist, können wir beurteilen, ob die Maximen der Quantität, Modalität und, zusammenfassend, der Relevanz befolgt wurden. Die Tatsache, dass auch falsche Information relevant sein kann, darf uns nicht darauf schließen lassen, dass wir keinen Wahrhaftigkeitsanspruch haben, was sich ganz eindeutig zeigt, wenn wir über Sprecher urteilen, die die Maxime der Qualität verletzt haben. Man kann diese Maxime nicht durch das Relevanzprinzip ersetzen, ohne das gesamte Konzept der Bedeutung der Wahrheit in Frage zu stellen.

Zusammenfassend können wir sagen: Die Übermaxime der Qualität zu befolgen ist ein fundamentaler Bestandteil einer glücklichen Kommunikation. Obwohl man auch dadurch, dass man etwas Falsches sagt, wichtige Informationen liefern kann, im Gegensatz zur Griceschen Auffassung, ist die Maxime konstitutiv für unsere Interpretation des Sprechaktes im Ganzen. Die offene Verletzung der Maxime im Falle der Ironie setzt voraus, dass wir erwarten, dass sie befolgt wird – ohne sie gäbe es überhaupt keine Ironie. Relevanz allein kann uns keinen ausreichenden Ersatz für Wahrhaftigkeit des Sprechers und Korrektheit seiner Information liefern. Wir erwarten Antworten, die nicht nur das Optimum an Relevanz erfüllen, sondern auch die ehrliche und fundierte Meinung des Sprechers ausdrücken. Gerade weil wir Konsequenzen aus den Äußerungen der Anderen ziehen, haben diese eine gewisse Verantwortung, aufrichtig zu antworten. Wenn wir als Sprecher die Maxime der Qualität verletzen, kann dies unerwünschte Folgen haben. In solch einem Fall ist es wohl kaum noch angebracht, vom Erfolg der Kommunikation zu sprechen.

[5] Siehe bspw. Wilson und Matsui 1998.

Um ihren Status als vorherrschende pragmatische Theorie aufrecht zu erhalten, muss sich die Relevanztheorie mit der Bedeutung des Wahrheitsanspruchs in unserer Kommunikation auseinandersetzen.

**Theresa Marx**

Universität Heidelberg
theresa.marx@gmx.net

## Literatur

Grice, H.P. 1989: *Studies in the way of words*. Harvard: Harvard University Press.

Matsui, T. Y Wilson, D. 1998: „Recent approaches to bridging: truth, coherence and relevance“, *UCL Working Papers in Linguistics* 10: 173-200.

Van der Henst, J-B., Carles, L. y Sperber, D. 2002: „Truthfulness and relevance in telling the time“, *Mind & Language* 17: 457-466.

Sperber, D. y Wilson, D. 1995: *Relevance. Communication & Cognition*, Oxford: Blackwell Publishing, Second Edition.

Wilson, D. 1995: „Is there a maxim of truthfulness?“ *UCL Working Papers in Linguistics* 7: 197-212.

# 2. Metaphysik und Ontologie

# The Fundamental Question of Metaphysics and the Question of Fundamentality in Metaphysics

Brandon C. Look

Why is there something rather than nothing? This question is often considered the fundamental question of metaphysics. This paper will concern a related question, one that is arguably even more fundamental: Is nothingness possible? For it is only if nothingness is possible that the fundamental question of metaphysics really develops any force. There are two ways to justify the possibility of nothingness. One can hold that "nothing is simpler and easier than something" – that nothingness is, as it were, the default state of the universe that only an act of creation can overcome. Or one can argue, as Thomas Baldwin and others have done, that it is possible to "subtract" concrete objects from worlds until one has arrived at a world without concrete objects – an "empty world." It will be argued in this paper that the premises of the Subtraction Argument rest on tendentious and question-begging assumptions about ontological dependence and the grounding relation. In other words, questions of fundamentality in metaphysics reveal the fundamental question of metaphysics to be ill-formed and arguments purporting to show the possibility of nothingness invalid. Against the view of metaphysical nihilism, this paper argues for metaphysical aliquidism – the view that there must be something.

## 1. Introduction

*Why is there something rather than nothing?* This question, Leibniz tells us in the *Principles of Nature and Grace*, is the very first question that we ought to ask ourselves: "Assuming this principle [the Principle of Sufficient Reason], the first question we have the right to ask will be, why is there something rather than nothing?"[1] Indeed, Heidegger has called it "the fundamental question of metaphysics [*die Grundfrage der Metaphysik*]"; he writes, "Die dem Range nach erste, weil weiteste, tiefste und ursprünglichste Frage [ist]: *»Warum ist überhaupt Seiendes und nicht vielmehr Nichts?«*" (1983, 3). Wittgenstein, too, seems to be moved to awe: "Nicht *wie* die Welt ist, ist das Mystische, sondern *dass* sie ist" (1922, 186; *Tractatus* 6.44). As fundamental as this question may be, it is not the question to be addressed in this paper – at least not directly. Rather, this paper will concern a related question, one that is arguably even more fundamental: Is the fundamental question of metaphysics a well-formed and meaningful question? One might think, for example, that it is a kind of complex question. The famous example, of course, is "Have you stopped beating your wife?" Or, one might just recognize that *Why?*-questions are often problematic. The fundamental question of metaphysics assumes that nothingness is possible, but is it? That is, is nothingness possible? Or, given that there is obviously something, might there be (or might there have been) nothing? For it is only if nothingness is possible that the fundamental question really develops any force.

[1] "Ce principe posé, la premiere question qu'on a droit de faire, sera, *Pourquoy il y a plustôt quelque chose que rien?*" (Leibniz 1965, hereafter 'G', VI 602)

## 2. The Possibility of Nothingness

There are two ways to justify the possibility of nothingness. One can hold, as Leibniz himself did, that "nothing is simpler and easier than something" (G VI 602) – that is, that nothingness is the natural or default state and only an act of will, an act of free creation, can overcome this. Put differently, for Leibniz, given the Principle of Sufficient Reason, *being* requires a ground, or a reason, or a justification; *nothingness* does not require metaphysical grounding. Yet, this is less an argument for the possibility of nothingness than it is an article of faith. And, indeed, as I suggested above, the possibility of nothingness needs to be shown *given that there is something*. And this is what Thomas Baldwin (1996), Gonzalo Rodriguez-Pereyra (1997; 2000; 2002; 2004) and others have tried to do, to show that it is possible to "subtract" concrete objects from worlds until one has arrived at a world without concrete objects – an "empty world." This "Subtraction Argument" for "metaphysical nihilism" has occasioned a fair bit of discussion in the philosophical literature in the last decade, and it is worth considering seriously. David Armstrong (2004, 89–91) has written approvingly of the Subtraction Argument and has given an additional argument appealing to truthmakers for the possibility of an empty world. Jonathan Lowe (1998; 2002, 252–55), Alexander Paseau (2002; 2006), Ross Cameron (2006) and others, on the other hand, have countered that the Subtraction Argument fails and that metaphysical nihilism is not possible. Moreover, David Lewis (1986, 73–74) takes his modal realism – or rather his definition of a world – to block the possibility of a world without any concrete objects, and David Armstrong, in earlier work (1989, 24–25, 63–64), likewise rejects the possibility of an empty world.

The Subtraction Argument as advanced by Baldwin and Rodriquez-Pereyra is as follows. We start with the following premises:

(1) There might be world with a finite domain of 'concrete' objects.

(2) These concrete objects are, each of them, things that might not exist.

(3) The non-existence of any of these things does not necessitate the existence of any other such thing.

Starting from some possible world $w_1$, accessible from our actual world, we pick out a concrete object $x_1$, remove it completely, and go on to the next world $w_2$, which is qualitatively identical with $w_1$, except that it lacks $x_1$. By the premises above, our subtraction of a particular concrete object, $x_n$, does not entail the existence of any "new" objects in other possible worlds. Now, we are to continue this process to $w_{min}$, a world with one concrete object. And, finally, according to the Subtraction Argument, we are to imagine removing the one concrete object of $w_{min}$, whereby we arrive at $w_{nil}$ – a world with no concrete objects. Thus, if accessibility is transitive, an empty world is possible, and metaphysical nihilism is tenable.

This argument looks good on its surface. But are the premises above reproach? Most of the attention in the literature has been directed at premise (3), and I will turn to that in a moment. But before getting there let's take a moment to focus on premise (1) – a seemingly innocent premise: "There might be world with a finite domain of 'concrete' objects." Certainly, we all know what is meant by such a claim. But is it beyond reproach? It seems to me that one could easily argue against it on three related grounds. One might believe in an actual infinity in the world of concrete objects, as Leibniz did, at least with respect to the material objects of the phenomenal realm, that any particular part of matter is actually divided into or composed of an infinity of other creatures: "I believe that there is no part of matter that is not – I shall not say divisible – but actually divided, and consequently the least

article ought to be considered as if it were a world full of an infinity of different creatures."[2] I doubt there are many advocates of the Leibnizian position, but there are some.[3] I suspect there might be advocates of a related view: that there is no fundamental level to reality. Chairs are made up of atoms, which are made up of protons, neutrons and electrons, which in turn are made up of subatomic particles, and so on. An added twist, and one that seems to have more going for it scientifically, is to say that there *is* a fundamental level, but that it is different states of a quantum field. Perhaps this entails "vague objects"; perhaps it simply means that the idea of singular, discrete concrete objects is in itself rather question-begging.

Of course, advocates of the Subtraction Argument are not going to let this stand in the way of a nifty argument. And perhaps they should not. Again, it is obvious enough what is meant by concrete object. And Rodriguez-Pereyra makes this explicit in a later piece, saying that "a concrete* object is an object that is concrete, memberless, and a maximal occupant of a connected region [in space]." (Rodriguez-Pereyra 2002, 73) In other words, it is *stipulated* that the concrete objects of the Subtraction Argument are discrete, atomic, and fundamental.

As I mentioned, the majority of philosophers who have objected to the Subtraction Argument have been concerned with premise (3): "the non-existence of any of the concrete objects does not necessitate the existence of any other concrete object." Are there reasons to believe that this is true? Even Baldwin concedes that "[i]t is not .. easy to think of a direct argument for this premise." (1996, 235) On its surface, it would seem that the premise says simply that concrete objects or substances are independent from each other; that is, it would seem that this premise simply appeals to the *independence-criterion* of substancehood, which has a long tradition in philosophy. According to Aristotle, "Some things can exist apart and some cannot, and it is the former that are substances."[4] Or as Descartes puts it in the *Principles of Philosophy*, "By *substance* we can understand nothing other than a thing which exists in such a way that it depends on no other thing for its existence."[5] Of course, Descartes quickly adds to this definition of substance that, strictly speaking, only God is truly independent *but you know what I mean*. And Spinoza, of course, knew exactly what Descartes meant, going from this independence criterion of substance to monism. But that's another story.

The notion of metaphysical or ontological dependence is at issue in another sense in premise (3), which comes out in Jonathan Lowe's criticism of the premise. According to Lowe, premise (3) leads to metaphysical nihilism, a view he rejects; therefore, premise (3) must itself be rejected. It's as simple as that. The positive argument for his contrary position is also formally quite simple, though metaphysically complex:

(4) Some abstract objects, like natural numbers, exist necessarily.

(5) Abstract objects depend for their existence upon there being concrete entities.

Therefore,

(6) It is necessary that there are concrete entities.[6]

Now, this argument takes us into very deep metaphysical waters. I agree with Lowe, but I would like to find independent reasons to reject the Subtraction Argument.

---

[2] "Ainsi je crois qu'il n'y a aucune partie de la matiere qui ne soit, je ne dis pas divisible, mais actuellement divisée, et par consequent, la moindre particelle doit estre considerée comme un monde plein d'une infinité de creatures differentes." (Letter to Foucher, 1693; G I 416).

[3] Consider the discipline of continuum mechanics. "The continuum theory regards matter as indefinitely divisible." (Lai, Rubin, and Krempl 2009, 1)

[4] *Metaphysics*, XII.5: 1070b36-1071a1 from Aristotle (1984).

[5] "Per *substantiam* nihil aliud intelligere possumus, quam rem quae ita existit, ut nulla alia re indigeat ad existendum." (Descartes 1996, VIII 24/Descartes 1985, I 210, translation altered).

[6] This is Rodriguez-Pereyra's version (2000, 335) of what is found in (Lowe 1998, 252–55).

Alexander Paseau (2002) presents a more direct challenge to premise (3), arguing that on either of two interpretations of (3) the argument is ultimately invalid. Labelling the finitely many concrete objects of (1) and (2) $x_1$ to $x_n$, Paseau claims that (3) could mean that "the non-existence of any given one of the finitely many objects .. does not necessitate the existence of any other given one of these $x_i$." Or (3) could mean something stronger: "the non-existence of any one of these $x_i$ does not necessitate that there is even one of the $x_i$." (2002, 74) But Paseau asks us to consider a model, containing three entities: $x_1$, $x_2$, and $x_3$. We, therefore, have the following seven sets or worlds: $\{x_1\}$, $\{x_2\}$, $\{ x_3\}$, $\{ x_1, x_2\}$,$\{ x_1, x_3\}$, $\{ x_2, x_3\}$, $\{ x_1, x_2, x_3\}$. It should be easy to see that on this model, (1), (2), and either reading of (3) are all true, and yet there is no empty set ("null world"). The argument is, therefore, invalid.

Rodriguez-Pereyra appears to concede that, on the above interpretations of (3), the Subtraction Argument is invalid. But he claims that there is another reading of (3), which, is more perspicuous and gets at the original intention. This time it is more explicitly expressed in terms of possible worlds:

> (3*) The non-existence of any of the $x_i$ that exist in $w_1$ does not necessitate the existence of any other concrete* object, whether or not these exist in $w_1$. That is: for all worlds $w$ and for all the *concreta** $x_i$ in $w_1$, if $x_i$ exists in $w$ then if there is a world $w$* where $x_i$ does not exist, then there is a world $w$** where the only existing *concreta** are those of $w$ except $x_i$ (i.e. $w$** is such that for every concrete* object $y$, $y$ exists in $w$** if and only if $y \neq x_i$ and $y$ exists in $w$). (2002, 172)

In other words, if $x_i$ is a contingent concrete object in $w$, there is some world $w$** that contains everything in $w$ except for $x_i$. Thus, the Subtraction Argument can proceed as before: we subtract contingent concrete entities until we arrive at some world $w_{min}$, which has only one object, and there is nothing preventing us from subtracting that object, so that we have $w_{nil}$ – the null world.

I shall not concern myself with the question of whether or not the possible-worlds interpretation is equivalent to the original premise (3). Rather I wish to consider the amended argument as it stands. The scenario the nihilist imagines is clear enough. Suppose we have the following concrete objects: $x_1$, $x_2$, and $x_3$; we then have the following sets or possible worlds: $\{x_1, x_2, x_3\}$, $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1\}$, $\{ x_2\}$, $\{x_3\}$. Our largest world is $w_{max} = \{x_1, x_2, x_3\}$. But there are also singleton worlds, including $w = \{x_1\}$. Since there is a $w$* that does not have $x_1$ as a member (e.g., $\{x_2, x_3\}$), $x_1$ is contingent. Now, according to premise (3), there must also be a world, $w$** = {Ø}, for the non-existence of $x_1$ does not necessitate the existence of any other concrete objects.

Even with this improvement to premise (3), I believe that there is still an obvious problem. The Subtraction Argument may be acceptable in talking about the subtraction from $w_{max}$ to any of the worlds of two or one element. But, surely (!), the "subtraction" of concrete objects from two-object worlds to singleton worlds differs greatly from the subtraction of lone concrete objects from singleton worlds to a null-world. Indeed, it would seem that this is deeply question-begging, for the null-world is presupposed in the very argument that is intended to prove its possibility. Moreover, I think that the clause "there is a world $w$** where the only existing *concreta** are those of $w$ except $x_i$" is deceptive, for, when there are no *concreta* in $w$ other than $x_i$, the world-mates of $x_i$ are described (and treated) as both existent and non-existent. Further, we ought to be suspicious of the existential quantifier "there is a world $w$**.." absent any kind of definition of a world. Indeed, it would seem that subtraction to an empty world is only possible if one has already objected to a pretty straightforward notion of a world as a set of objects that are spatio-temporally or causally connected to each other. As I said earlier, Lewis, for example, thinks that his definition of a world blocks the very possibility of a null world at the outset, and I am inclined to agree with him.

Although premise (3) seems to capture the apparently innocent notion of the independence criterion of substance, we need to notice how truly odd the wording of this premise really is: the *non-existence* of any concrete object *does not necessitate* the existence of any other concrete object – *in another possible world*. How many dubious metaphysical ideas are at work in this premise? We have here (a) the causal powers of *absences* or *absences* grounding other states of affairs and, perhaps, (b) the causal powers or grounding of a state of affairs in one world on another world. If premise (3) is taken to be metaphorical – *just because you take away an object, that doesn't mean that another one has to pop up in its place* – then how seriously do we have to take the argument itself?

## 3. Reason and Ground

I have referred to Leibniz several times thus far, and let me do so again by returning to one of his "great principles of all our reasoning" – the Principle of Sufficient Reason (PSR). The PSR seems to be making something of a comeback among metaphysicians these days, and there is a perfectly good reason that this rationalist workhorse is doing so. It is simply an ancestor to the truth-maker principle endorsed by so many contemporary philosophers. Leibniz, of course, employs the Principle of Sufficient Reason throughout his metaphysics, most famously, in giving his cosmological argument for the existence of God. When we consider the Subtraction Argument, however, we should realize that premise (3) violates the Principle of Sufficient Reason in an odd way. The PSR, as Leibniz usually states it, is this:

> **PSR* (Nihil est sine ratione)**
>
> There is nothing for which there is not a reason why it is (and why it is so and not otherwise).

Of course, the PSR, with its double-negation, can also be reformulated thus (a reformulation that Leibniz himself also uses):

> **PSR** (Omne ens habet rationem)**
>
> For every being, there is a reason (or ground) for its being (and being so and not otherwise).

While we have seen that there is a relatively standard and innocent *independence criterion* of substancehood at work in premise (3), at the same time it would seem to violate a normal reading of the Principle of Sufficient Reason as PSR**. Every finite contingent being must be grounded in some other being. When we get to the Subtraction Argument's $w_{min}$ with only one finite, contingent, concrete object, we can ask, supposing PSR**, where or what is its ground? It should be noted that the Subtraction Argument's "progression" from a $w_{max}$ to a $w_{min}$ resembles Leibniz's move from contingent, finite causes back to an original, contingent being. For Leibniz, there's no problem in identifying its ground: given PSR**, the ground is ultramundane, God. But I do not believe any of the advocates of metaphysical nihilism would follow Leibniz here. Rather, I believe they would simply reject the Principle of Sufficient Reason (PSR**); that is, advocates of metaphysical nihilism require that PSR** not hold in the case of premise (3).

But should the Principle of Sufficient Reason be so quickly jettisoned? When we focus on the truth-maker relation, we can just as easily see that there is no metaphysical or ontological ground, no reason, no truth-maker for any fact concerning a world of no concrete objects.[7] And if metaphysical nihilists reject the PSR**, then their argument is in trouble for a different

[7] I realize there are ways of explaining absences in truth-maker lingo. But it is not entirely clear to me that there can be a truth-maker for a proposition about a world of no concrete objects. At the very least, it seems to me that some work needs to be done here.

– though related – reason. The Principle of Sufficient Reason, or the notion of metaphysical or ontological grounding, underlies the principles that lead us to say that there is some fundamental level of reality. They are the principles that seem to prevent us from saying, "It's turtles all the way down.." Therefore, denying the Principle of Sufficient Reason or the idea that there must be an ontological ground for any being undermines the idea that allowed us to admit the finite, concrete objects of premise (1). In other words, if the metaphysical nihilist accepts some kind of grounding thesis to motivate his intuitions about finite, concrete objects in premise (1), then he should likewise admit of a grounding thesis in premise (3). If he denies a grounding thesis in premise (3), then he cannot have the finite, concrete objects that he wants for premise (1).

A conciliatory (or purposefully ambiguous) conclusion would be that the Subtraction Argument proves *nothing*. But I should like to be more blunt: in my view, the Subtraction Argument *fails*. Its explicit and implicit premises rest on tendentious and question-begging assumptions about *ontological dependence* and the *grounding relation*.[8] For example, the crucial premise ("the non-existence of any of the *concreta* does not necessitate the existence of any other concrete object") assumes that there is no ontological dependence of one concrete thing on another. The argument assumes that abstract objects are not grounded in concrete objects – which might be taken as a reason to reject the argument *prima facie*. And the argument assumes that a world is not grounded in or dependent upon concrete objects. While the Leibnizian intuition that nothingness is more natural than something and that therefore there must always be a *reason* or *ground* for the existence of something, the Subtraction Argument assumes that its empty worlds, constituted presumably by *abstracta*, can exist and can exist *so* simply by fiat. Or rather, it is assumed that there is no *reason* or *ground* for the worlds' *So-sein* – which may be as metaphysically presumptuous as anything Leibniz dreamed of. Thus, questions of fundamentality in metaphysics reveal the fundamental question of metaphysics to be ill-formed and arguments purporting to show the possibility of nothingness invalid.

## 4. The Fundamental Question Forsaken

A lot is riding on the outcome of this dispute. If we can show that metaphysical nihilism is not possible (that is, that it is not possible that there be nothing), it seems that we can undercut the motivation of the fundamental question of metaphysics. For it is only when nothingness is possible that we need to address the question why there is something rather than nothing. Naturally, if the fundamental question loses its force, so too does any form of the cosmological argument for the existence of God. If, on the other hand, it is impossible that there be nothing, something must exist. Therefore, *either* we need give no explanation of the fact that something exists – it's a brute fact – *or* we say, as Spinoza and Russell did, that the world is, as it were, a *causa sui*. Of course, one could demand why things are *this* way and not some other way, but that is a different question, answerable in large part (perhaps completely) by natural science. Thus, if metaphysical nihilism is impossible and if we recognize that the Leibnizian prejudice that *nothingness is simpler and easier than being* is simply a prejudice, then the fundamental question of metaphysics can no longer be sensibly asked.

In this paper, I have *not* shown that metaphysical nihilism is *impossible*, only that one of the contemporary arguments for it is wanting. But the Subtraction Argument's weakness should also alert us to some problematic aspects of any argument purporting to show the possibility of nothingness, and my analysis should also lead to a scepticism with respect to the fundamental question of metaphysics itself. For example, it is unlikely that an argument for

[8] On these matters, see Correia (2008), Fine (2010), and Rosen (2010).

metaphysical nihilism can be produced that does not depend upon a controversial or ambiguous notion of a "world." Moreover, the fundamental question of metaphysics derives much of its force from the historically and culturally contingent thought that "God could have chosen not to create anything" – a thought that should hardly move a steely-eyed atheist. The fundamental question of metaphysics also implicitly appeals to the thesis that conceivability entails possibility – for while I can (perhaps) *think* of a world prior to creation or not having any "stuff" in it, this thought differs from a proof of its metaphysical possibility. Finally, the fundamental question also derives force from the completely unanalyzed claim that "nothing is simpler and easier than something." There is, then, a presumption towards nothingness that underlies the fundamental question of metaphysics that I reject.

It seems rather that the burden of proof should really be upon those who advocate metaphysical nihilism to show that it is possible that nothing exist. To the credit of advocates of the Subtraction Argument, they have tried to provide such an argument. But, as I hope to have shown, this attempt is a failure, and to date I have been unconvinced of any attempt to prove the possibility of nothingness. Indeed, the contrary view, that there must be something, should become the default position, a position I should like to call "metaphysical aliquidism" – the view that there must be *something*. There are two versions of *metaphysical aliquidism*. The first might be seen as a kind of monism; its thesis is simply that there is a necessarily existing concrete object – the world. Such a view is attributable, of course, to Spinoza, Russell and, more recently, Jonathan Schaffer (e.g. in Schaffer (2009; 2010)). The second version can be expressed thus: for any world, necessarily, there is a concrete object in it. This view is advocated by Lewis – though one need not be a modal realist to endorse it. Either version of *aliquidism*, I believe, is more in line with our intuitions of ontological dependence. And if we accept *metaphysical aliquidism*, the fundamental question of metaphysics ought no longer to be a question for us.[9]

**Brandon C. Look**

University of Kentucky
look@uky.edu

## References

Aristotle. 1984: The Complete Works of Aristotle. J. Barnes (ed.). 2 vols. Princeton: Princeton University Press.

Armstrong, D. M. 1989: A Combinatorial Theory of Possibility. Cambridge: Cambridge University Press.

— 2004. Truth and Truthmakers. Cambridge: Cambridge University Press.

Baldwin, T. 1996: 'There Might Be Nothing', Analysis 56, 231–38.

Cameron, R. 2006: 'Much Ado About Nothing: A Study of Metaphysical Nihilism', Erkenntnis 64, 193–222.

Correia, F. 2008: 'Ontological Dependence', Philosophy Compass 3, 1013–1032.

Descartes, R. 1985: The Philosophical Writings of Descartes. J. Cottingham, R. Stoothoff, and D. Murdoch (eds.). 2 vols. Cambridge: Cambridge University Press.

Descartes, R. 1996: Oeuvres de Descartes. C. Adam and P. Tannery (eds.). 11 vols. Paris: J. Vrin.

Fine, K. 2010: 'Some Puzzles of Ground', Notre Dame Journal of Formal Logic 51, 97–118.

[9] My thanks to the audience at the GAP8 conference for helpful questions and suggestions.

Heidegger, M. 1983: Einführung in die Metaphysik. Frankfurt a.M.: Vittorio Klostermann.

Lai, W. M., D. Rubin, and E. Krempl. 2009: Introduction to Continuum Mechanics. 4th ed. Amsterdam: Elsevier.

Leibniz, G.W. 1965: Die Philosophischen Schriften. C. I. Gerhardt (ed.) 7 vols. Hildesheim: Olms.

Lewis, D. 1986: On the Plurality of Worlds. Oxford: Basil Blackwell.

Lowe, E.J. 1998: The Possibility of Metaphysics: Substance, Identity, and Time. Oxford: Oxford University Press.

— 2002: 'Metaphysical Nihilism and the Subtraction Argument,' Analysis 62, 62–73.

Paseau, A. 2002: 'Why the Subtraction Argument Does Not Add Up', Analysis 62, 73–75.

— 2006: 'The Subtraction Argument(s)', Dialectica 60, 145–56.

Rodriguez-Pereyra, G. 1997: 'There Might Be Nothing: The Subtraction Argument Improved', Analysis 57, 159–66.

— 2000: 'Lowe's Argument Against Nihilism', Analysis 60, 335–40.

— 2002: 'Metaphysical Nihilism Defended: Reply to Lowe and Paseau', Analysis 62, 172–80.

— 2004: 'Modal Realism and Metaphysical Nihilism', Mind 113, 683–704.

Rosen, G. 2010: 'Metaphysical Dependence: Grounding and Relation', in B. Hale and A. Hoffmann (eds.): Modality: Metaphysics, Logic, and Epistemology, Oxford: Oxford University Press, 109–35.

Schaffer, J. 2009: 'On What Grounds What', in D. Chalmers, D. Manley, and R. Wasserman (eds.): Metametaphysics: New Essays on the Foundations of Ontology, Oxford: Oxford University Press, 347–83.

— 2010: 'Monism: The Priority of the Whole', Philosophical Review 119, 31–76.

Wittgenstein, L. 1922: Tractatus-Logico Philosophicus. London: Routledge & Kegan Paul.

# Why Dispositions Are Not Higher-order Properties

Joshua Mugg

In this paper I defend C.B. Martin's identity theory in which intrinsic properties of concrete objects are simultaneously qualitative and dispositional. Using three arguments from Sydney Shoemaker, I demonstrate that there are epistemic difficulties with ontologically separating dispositional and qualitative properties. I use Prior, Pargetter, and Jackson as a paradigm case of such an attempt to separate these two kinds of properties. The difficulty with Prior et al.'s higher-order account of dispositions is this: given an asymmetry relation, the qualitative properties can vary without necessarily altering the object's dispositions. Given that our interaction with an object is with its dispositions, our knowledge of objects becomes severely limited. Therefore, we ought not posit qualitative and dispositional properties as ontologically distinct.

## 1. Introduction

An ontological account of properties should not logically separate qualitative and dispositional properties. In this paper, I used three epistemic arguments from Sydney Shoemaker to demonstrate that separating these two kinds of properties leads to epistemic difficulties. I then point out some difficulties with (or at least *prima facie* oddities of) Shoemaker's purely dispositional account. I conclude by offering an alternative explanation of the relation between qualitative and dispositional properties, one advocated by C.B. Martin and John Heil. Martin and Heil argue that qualitative and dispositional properties are merely different ways of looking at the same property; all properties are both dispositional and qualitative (Martin 2008: 68). Heil puts it thus:

> If P is an intrinsic property of a concrete object, P is simultaneously dispositional and qualitative; P's dispositionality and qualitativity are not aspects or properties of P; P's dispositionality, $P_d$, is P's qualitativity, $P_q$, and each of these is P: $P_d = P_q = P$. (Heil 2003: 111)

In other words, when I say that a sphere has the property of 'roundness' I am looking at (or conceptualizing) the property qualitatively. In another respect, the sphere has the property of 'readiness to roll,' which is one and the same property (ontologically) as 'roundness.' This version of monism is called the identity theory.

According to Shoemaker, any property of an object is a causal power (a power to cause something) of that object (2003: 214-215). Furthermore, there are no non-dispositional[1] properties. On this theory, when we say that a sphere is 'round,' we mean that the sphere has a property 'to roll.' There is no qualitative property of 'roundness,' just the property 'to roll.' Shoemaker argues for this view by demonstrating that epistemic difficulties arise from positing qualitative properties that are ontologically distinct from dispositional properties. I will demonstrate that Shoemaker's arguments support the identity theory as much as they support his pure powers account.

[1] I use the term "non-dispositional" for Shoemaker's view because he not only rejects qualitative properties, but any other kind of property that is not dispositional.

Shoemaker's arguments begin with a supposition (for *reductio*) that the identity of properties consists of something logically independent of causal powers (Shoemaker 2003: 214-215). That is, these two kinds of properties really are independent of each other; they are not just two ways of conceptualizing the same property. They might interact with one another, but not necessarily so. A common view of this sort is one in which dispositions are higher-order properties with qualitative properties as a supervenience base. It will be helpful to have a specific account in mind. I offer the account of Elizabeth W. Prior, Robert Pargetter, and Frank Jackson.

In their paper "Three Theses About Dispositions" Prior, Pargetter, and Jackson argue that all dispositions have causal bases; that these causal bases are distinct from their dispositions; and that dispositions "are causally impotent with respect to their manifestations" (1982: 251). It is their second thesis, that dispositions are higher-order properties, with which I am concerned. Although they wish to "remain as neutral as possible on various wider metaphysical issues concerning...realism about properties and the distinction between categorical properties and dispositional ones," such neutrality is not an easy task. If we assume realism about qualitative and dispositional properties, then their view leads to the ontological separation of these properties. Let me explain the view of Prior et al. with the assumption that qualitative and dispositional properties are real.[2]

Prior et al. argue that two objects might have the same dispositions, yet have different molecular structures (1982: 253). That is, two objects might have identical dispositions, yet lack identical qualitative properties. Prior et al.'s thesis concerning the relation between dispositions and qualitative properties is similar to the functionalist thesis concerning the relation between physical and mental states. Two subjects might be in the same pain state while being in different physical states. In other words, if subject X is in mental state $m$ and physical state $p_1$ and subject Y is in state $m$ and physical state $p_2$, we cannot infer that $p_1$=$p_2$. In the same way, if object X possesses disposition $d$ with causal base of qualitative properties $q_a$,...$q_m$[3] and object Y possesses disposition $d$ with causal base of qualitative properties $q_n$,... $q_z$ we cannot infer that $q_a$ ,... $q_m$ = $q_n$ ,... $q_z$; if two objects have identical disposition they need not have identical qualitative properties.

Prior et al. explain that even if there is only one set of qualitative properties (call it Q) that serves as a causal basis for some disposition $d$, it need not be the case that an object's possessing Q implies that object's having $d$. Consider that some object might possess Q while possessing some further properties Q* "which swamp[s] the effect of having" that disposition (1982: 253). Because this swamping effect is possible, dispositions are not identical to their qualitative bases. In fact, Prior et al. are saying something stronger: two objects possessing identical qualitative properties need not possess identical dispositional properties. Dispositions, according to Prior et al., are multiply realizable, higher-order properties. However, this thesis leads to epistemic problems.

---

[2] Although I think Prior et al. have in mind realism of both dispositional and qualitative properties, they could deny the existence of qualitative properties which would imply (given their first thesis that every disposition has a disposition base) that dispositions never 'bottom out.' That is, that there can be no fundamental properties. Consider some disposition x. On their view this disposition must have some causal base (by their first thesis), call this property y. However, by supposition, y is not qualitative, so it must be dispositional. Again, by their first thesis, y must have a causal base, which will also be dispositional, which must have a causal base...

[3] According to Prior et al., the disposition might have a single qualitative property or a set of qualitative properties as its causal base.

## 2. Shoemaker's Arguments

### *2.1 Argument 1*

Given an account of dispositions as higher-order properties, it is possible for two or more properties to make (in every circumstance) exactly the same contribution to the object's dispositions. If one of these properties were to change, the object's dispositions would not change because there is still another non-dispositional property contributing to the dispositions. I can put it most precisely in the following way: at $T_1$ an object *O* possesses properties *A* and *B*, which make exactly the same contribution to *O*'s dispositional properties $d_1$, $d_2$, $d_3...d_i$. At $T_2$ object *O* loses property *A* such that there is no *A* to contribute to *O*'s dispositional property $d_1$, $d_2$, $d_3...d_i$, but property *B* still contributes to *O*'s dispositional properties $d_1$, $d_2$, $d_3...d_i$. Thus, even though one of *O*'s properties has changed, its dispositions remain the same. Because of the asymmetry relation in Prior et al.'s account, the higher-order dispositional properties supervene on the qualitative properties *A* and *B*, but the alteration of *A* or *B* alone does not change the higher-order dispositional properties.

This argument presents us with an epistemic problem: any interaction we have with *O* will be with its dispositional properties and, since there has been no dispositional change, we cannot know that this object has changed. Thus, we may be mistaken in judging objects to be unchanged—for the object may have lost or gained some property that makes an identical contribution to that object's dispositionality.

Perhaps an example will help. Think of a watermelon that has the disposition to be eaten, to roll, to be cut, and to make my mouth water. These higher-order properties supervene on the watermelon's qualitative properties (i.e. its being juicy, its being roundish, its being ripe, etc.). On a higher-order account of dispositionality two of these qualitative properties can make exactly the same contribution to the watermelon's dispositions. For example the juiciness of the watermelon and its water content divided by its volume might make exactly the same contribution to its causing my mouth to water. Suppose the water content divided by the watermelon volume changed, but the watermelon remained juicy. Since these make the same contribution to the watermelon's dispositions, the change will go unnoticed.

The defender of dispositions as higher-order properties may respond that, if two qualitative properties always contribute to the object's dispositions in the same way, they are in fact the same property. In the case of the watermelon, one might say that water content divided by watermelon volume is how we measure the fruit's juiciness. I think this is a sensible claim, but what would justify this inference on Prior et al.'s account? If these non-dispositional properties really are logically distinct from dispositional properties, then the contribution they make to dispositions does not determine their identity. On Prior et al.'s account that two properties make exactly the same contribution to an object's dispositions does not imply that the two properties are identical.

### *2.2 Argument 2*

Consider objects *X* and *Y* possessing identical dispositions, but not possessing identical qualitative properties. Although the objects do not possess identical qualitative properties, we would judge the objects to be identical to one another. However, we would be mistaken. That is, it would be impossible for us to recognize that the two objects are not identical. Two objects possessing identical dispositions is necessary (but not sufficient) for their being qualitatively identical (since to be identical they must also possess the same qualitative properties). Again, any interaction with objects must be with its dispositional properties. So when we observe these objects, they will appear similar in every way because all their causal powers are identical, but the objects are not identical since their qualitative properties are

different. So if these two objects appear to be the same in every way, we cannot know that they are qualitatively dissimilar. After all, the objects might have some qualitative property that does not affect their causal powers (Shoemaker 2003: 214-215).

Perhaps another example will help. The watermelon and its watermelon clone might be dispositionally indistinguishable. However, in the cloning process this cloned watermelon's genome sequence may have been slightly changed in an unnoticeable way—a way that does not affect its dispositionality. The set of lower-order properties that constitutes a causal base for each watermelon's dispositions may not be identical, but this does not imply that higher-order properties must also differ. As such it may be incorrect to say that the two watermelons are (qualitatively) identical.

### *2.3 Argument 3*

On a higher-order account of dispositions, qualitative properties can vary independently of dispositions. If the dispositions remain unchanged, it is impossible for us to know that something has retained a property over time. That is, it would be impossible for us to know that something has undergone a change with respect to its qualitative properties. Consider that if an object's dispositions depend upon its qualitative properties, then there could exist some qualitative property that is not supervened upon at all. Not all the lower level properties need affect an object's dispositions. As such, a change in that object's non-dispositional property would go unnoticed.

More precisely: at $T_1$, object *O* has numerous dispositional properties, $d_1$, $d_2$,... $d_n$, which supervene on qualitative properties *X*, *Y*, *Z*, and *O* has one qualitative property *A* upon which no dispositional properties supervene. At $T_2$, *A* has changed while *X*, *Y*, *Z* have not; therefore $d_1$, $d_2$,... $d_n$ have not changed either. Remember that any interaction with *O* will be with its dispositional properties. Thus, I cannot know that *O* has changed because I will not know that *A* has changed. In this case, I will judge that *O* has not changed, when in fact it has. In this case, as in the other two we find an epistemic difficulty—qualitative properties (or any non-dispositional properties of any kind, for that matter) end up being properties we may not know about.

### *2.4 Objection and Reply*

Now you may wish to question Shoemaker's assumption that any interaction with an object will be with its dispositional properties. This premise is essential if these arguments are to carry any force. Consider that, if I interact with an object, I do not interact with its non-dispositional properties (even if they exist) because, simply by definition, non-dispositional properties are not causal powers. As soon as I interact with a property, the property has had some kind of effect upon my senses, and thus is dispositional. So any interaction with non-dispositional properties is mediated by dispositional properties. Although this response will be unconvincing to the metaphysician who believes all properties are qualitative, this paper is not directed toward such a philosopher. I wish merely to persuade the proponent of a higher-order account of dispositions that qualitative and dispositional properties are not logically distinct from one another.

According to Shoemaker, the preceding arguments, if sound, demonstrate that a property's identity is intimately connected to that property's causal potentialities. More precisely, these arguments aim to persuade us that positing qualitative properties as ontologically separate from dispositional properties leads us to an epistemic quandary. A higher-order account of dispositionality results in skepticism about objects' non-dispositional properties and so a skepiticism about the objects in general (Shoemaker 2003: 214-215). Of course, it is possible for properties to exist without our knowledge of them. However, most philosophers prefer to

avoid such speculations when another option exists. Shoemaker offers pure dispositionality as an alternative. Without taking into consideration the idea that qualitative and dispositional properties are the same (which is admittedly a surprising identity (Heil 2003: 218)), a pure powers view of properties does look like the most viable theory.

## 3. Pure Dispositionality

Many metaphysicians have noted problems with pure dispositionality, usually assuming that this position stands or falls with a purely relational account of substance (for example, see Heil 2003: 102). Martin is among the few who do not assume that pure dispositionality stands or falls with a relational account of objects. Instead, Martin argues that a purely dispositional account of properties leads to two regresses. The first comes from properties reducing to dispositions, the second from the manifestation of dispositions reducing to dispositions. I will briefly outline his reasons for thinking that two regresses follow.

First, let us examine what a pure powers account of properties says about alleged 'qualitative properties.' On a pure powers account of properties, a property that appears qualitative (e.g. length) reduces to a capacity (e.g. being capable of being measured at six inches), but this capacity reduces to a disposition "for the formation of other capacities" (Martin 2008: 63). Now, these capacities, which are the reduction of a disposition, also reduce to the formation of other capacities. The same holds with a dispositional property, which is a capacity. This capacity reduces to a disposition "for the formation of other capacities" (Martin 2008: 63). These capacities, which are the reduction of this particular disposition, also reduce to more capacities. Problematically, any time we are confronted with a property, it reduces to a capacity, and any time we are confronted with a capacity, it reduces to another capacity. Because every capacity reduces to a reducible capacity, we have an infinite regress. So we need qualitative properties in our analysis of property identities.

Now let us examine the manifestation of dispositions. According to the pure powers view, when a disposition manifests, there is no manifestation of a qualitative property. The manifestation is merely another causal power (or a collection of causal powers). And when this new casual power is manifested, it will also merely be manifesting another causal power. Martin writes, "this image appears absurd even if one is a realist about capacities—dispositions. It is a promissory note that may be actual enough, but if it is for only *another* promissory note that is [for *another* promissory note, and so on], it is entirely *too* promissory" (Martin 2008: 63). The purely dispositional account does not allow that a real thing happens in a manifestation (I mean something beyond its gaining the power to cause something else); a manifestation is merely a promissory note. By this, Martin means that we receive a promise that, although the manifestation was not what we expected, some real event—the new dispositions—will lead to an eventual manifestation that we expect. In other words, the causal potentiality that the object acquired may lead to causing a real something to happen. However, each new disposition is only a disposition for gaining or losing dispositions (see Martin 2008: 61-63). Therefore, a world of dispositions is a world of *mere* potentialities. All that happens is the introduction or loss of potentialities.

Perhaps the above reasoning is unconvincing to one who is already a pure power theorist. My hope is that it gives some reason to deny the elimination of qualitative properties. But now it seems that we face a dilemma: on the one hand, positing qualitative properties as ontologically distinct from dispositional properties leads to epistemic worries; on the other hand, eliminating qualitative properties seems absurd. The identity theory is a way between these two unpalatable options. In asserting that properties are simultaneously dispositional and qualitative the identity theory avoids the epistemic worries arising from positing qualitative properties as distinct from dispositional properties.

## 4. Shoemaker's Arguments Support the Identity Theory

Argument 1 demonstrates that positing logically distinct, non-dispositional properties leads to skepticism concerning the relationship between qualitative and dispositional properties. The identity theory dissolves the epistemic problem. When two qualitative properties make exactly the same contribution to an object's dispositionality, it makes the most sense to say they are the same property. Thus, if $P_{q1} = P_{d1} = P_{q2}$, then $P_{q1} = P_{q2}$. Let me revisit our object with qualitative properties *A* and *B* and dispositional property *d*. Remember that *A* and *B* make exactly the same contribution to *d*. On the identity theory, when property *A* changes, so does property *B* and property *d*, and when property *B* changes, so does property *A* and property *d*. Furthermore, the argument reveals a problem with the view that a qualitative property merely *makes a contribution* to an object's disposition. When we think a qualitative property makes a contribution to a disposition, we are seeing the intimate connection between qualitative and dispositional properties because these two types of properties are the same.

The second argument demonstrates that positing an ontological distinction between dispositional and qualitative properties leads to an inability to know if two objects are identical. The identity theory assures us that we can correctly judge objects *X* and *Y* as identical because if all their dispositions are identical, all their qualitative properties must be identical as well. So the suggestion that two objects might have identical dispositional properties while differing in their non-dispositional properties demonstrates the absurdity of positing such non-dispositional properties. For in that case, we would judge objects *X* and *Y* to be identical, but our judgment would be wrong. So qualitative properties must not be ontologically distinct from dispositional properties; in other words, the identity theory holds true.

The third argument demonstrates that, if qualitative and dispositional properties vary independently of one another, then there may be qualitative changes of which we cannot (in principle) detect. Instead of doing away with qualitative properties, let us say that, when an object's qualitative properties change, its dispositional properties must change as well. Because the identity theory states that qualitative and dispositional properties are identical, when a qualitative property changes, the dispositional property changes as well. For example, if a sphere's qualitative property of 'roundness' changes to 'cubical,' then the sphere's dispositional property 'to roll' would change as well. Dispositional and qualitative properties must vary in harmony with one another because they are identical to one another.

Shoemaker demonstrates an epistemic quandary arising from the claim that qualitative and dispositional properties are distinct. While Shoemaker eliminates qualitative properties to resolve the problem, a pure powers view of properties is not the only option. Martin and Heil can affirm the arguments I have outlined in this paper. If we do not want to posit qualitative properties as distinct from dispositional properties, let us abandon the ontological distinction between the two. The distinction between the two would instead be merely conceptual—two ways of thinking of the same property.

## 5. A Brief Objection

One difficulty in criticizing a pure dispositionality view while defending the identity theory is that, according to the identity theory, every property is dispositional. I have stated that there are no non-dispositional properties under the identity theory (of course there are also no non-qualitative properties either). From this it may seem that the identity theory is subject to the same problems as a pure dispositionality view. However, I do not think this is the case.

The regress problems Martin poses to the pure dispositonalist are solved by the existence of qualitative properties. The first regress problem is solved because, when we offer a qualitative property, there is no need to reduce it to a disposition. Both views agree that some of the properties resulting from a manifestation of a disposition will be dispositions. This only leads to a regress if the manifestation is *only* a change in dispositionality. However, under the identity theory the manifestation will be qualitative as well. So although each manifestation is a promissory note for another promissory note, each promissory note contains what was promised as well. Because there are qualitative properties, we do have real manifestations as we expect.

## 6. Conclusion

Shoemaker demonstrates that positing non-dispositional properties, as Prior et al. do, leads to epistemic problems. Non-dispositional properties are properties about which we cannot know and, consequently, if we posit non-dispositional properties, we cannot really know an object. Because these arguments are epistemic in nature, they do not *demand* that we deny non-dispositional properties and, indeed, non-dispositional properties may exist. However, because we have no way of knowing if this is the case, and, furthermore, because there is no need to make such a proposal, we will be better off if we reject non-dispositional properties.

Shoemaker proposes not only that we reject non-dispositional properties but also that we reject qualitative properties. For those of us who would rather hold onto qualitative properties, I propose that we adopt the identity theory. The identity theory proposes that all properties are simultaneously dispositional and qualitative. If we propose that every property is dispositional (though it is also qualitative), then we avoid the problems presented by Shoemaker's arguments. Any change in a qualitative property will change the dispositional property that is identical to that qualitative property. If we know that two objects have identical dispositional properties, we know they have identical qualitative properties. Thus, these arguments, originally intended by Shoemaker to persuade us to accept a pure powers ontology, may persuade us instead to accept the identity theory.

**Joshua Mugg**
York University, Toronto
joshuamugg@gmail.com

## References

Heil, J. 2003. *From an Ontological Point of View*. Oxford: Claredon Press.

Martin, C. B. 1997. 'On the Need for Properties: The Road to Pythagoreanism and Back', *Synthese* 112, 193-231.

— 2008. *The Mind in Nature*. Oxford: Oxford University Press.

Prior, E., R. R. Pargetter, and F. Jackson. 1982. 'Three These About Dispositions', *American Philosophical Quarterly* 19, 251-257.

Shoemaker, Sydney. 2003. *Identity, Cause, and Mind*. Oxford: Clarendon Press.

# The Point of Action

Michael Oliva Córdoba

It is a commonplace in the ontology of action that actions are *events*. Thus we are entitled to raise questions that must be admissible if this supposition is to hold in the first place. For example, some (not all) events are *extended in time*. Hence we may ask: is every *action* extended in time? *Durationalism*—the position answering this question affirmatively—seems to be the received view. In contradistinction to that view, however, this paper suggests that the question is most likely to be answered negatively: We have reason to admit *point actions*, i.e. actions that are point events. Although the conclusion is potentially far-reaching with regard to the field of action theory, it is by no means revolutionary. It is in accordance with our *descriptive metaphysics* and suggested by three different, independent views that are widely accepted in the field of *event ontology*, the *semantics of achievement verbs*, and *speech-act theory*.

## 1. The Durational View

Is every action extended in time? Call any view given by an affirmative answer to this question (action theoretic) *durationalism*, and theorists subscribing to it *durationalists*. *Prima facie*, durationalism is quite convincing given what is generally assumed in the metaphysics and ontology of actions and events. After all, aren't all actions either bodily movements or at least very closely linked to such events? Moreover, isn't every action at least an attempt to change the world—and as such the exertion of a causal interference with the way things are? Surely, only temporal entities can interfere causally. If they do so, however, how could this interference not be (extended) in time?

It seems that most philosophers (and virtually all action theorists) are durationalists. Certainly, there isn't any debate about whether one should or should not subscribe to this view. So adherence to durationalism will most likely be implicit. (I'm pretty sure that you, dear reader, have durationalist sympathies yourself! ;-)) Nevertheless, there may be reasons, none worse than those above, why we should be more hesitant concerning durationalism. These reasons involve some likewise well-received views, which I shall introduce below. If what follows is convincing, however, we should be prepared to concede that not every action is extended in time—there are point actions. Further, we would have to admit that point actions don't play a marginal role in our social and communicative behaviour, but are quite central to it.

## 2. Two Sample Durationalists

Widespread as the view is, to pick out champions of durationalism has an air of arbitrariness. I hope that it is without injustice to say that durationalism plays a more central role in the work of, say, *Christine Korsgaard* and *Harry Frankfurt* than in the work of many other theorists. For Korsgaard, the control we take of our movements is an essential feature of our constitution of agents (Korsgaard 2008: 13). For Frankfurt, the control or guidance we have about or concerning our behaviour is the essential feature that the causal approach in action theory gets wrong (Frankfurt 1978: 157). As Frankfurt's own action theory is developed in

sharp contrast to that approach, it is literally built on his understanding of what taking control consists of.

Now, both in Korsgaard and in Frankfurt, control or guidance is essentially understood durationally—as extended in time. For Korsgaard, control is a process, and every process is extended in time. In Frankfurt it seems that guidance is only given when the ability to interfere with the course of our behaviour is manifested *while the behaviour is being carried out*—and this can only be given during a stretch of time.

So what is the problem when looked at against the background of two theories that do more than just pay lip service to durationalism? There is a simple rule we have to remind ourselves of here: If $\varphi$-ness is an essential feature of $x$, and if being $\varphi$ implies being $\psi$, then $x$ is $\psi$. Given this, if $x$ wasn't $\psi$, $\varphi$-ness couldn't be an essential feature of $x$. This applies to the case in question as well. To be sure, action theorists sympathetic to the views of Korsgaard and Frankfurt may not *think* of the durationalist implications of such views as particularly interesting—yet they *are*: If it was not the case that all actions are extended in time, control and guidance could not be essential features of action.

## 3. Three Problems

So are all actions extended in time? Three problem cases spring to mind: The cases of event ontology, achievement verbs, and speech-acts. As a warm-up, let's start with the least daunting:

### *3.1 Event Ontology*

It is generally assumed that actions are events (cf. Davidson 1980). If so, there is *prima facie* no reason to suppose that the variety of types of events isn't mirrored in the subdomain as well. One particularly interesting distinction here is that between extended events on the one hand and point events on the other. An extended event has a temporal extension (i.e., duration); a point event has a date (i.e. a point in time) but no duration. So far this does not come as news; that distinction is well known from the physicist's tool-box. However, let's not succumb to the temptation to play this distinction down. We cannot dismiss it as mere tool of the physicist. Quite to the contrary, it is firmly rooted in the fundamental way we think of the world, our *descriptive metaphysics*. We *do* make a distinction between events for which it makes sense to ask how long they took and those for which it just doesn't. Contrast, e.g., The Glorious Revolution (extended) vs. its conclusion (point), the Apollo 11 mission (extended) vs. man's first setting foot upon the Moon (point), the Universe (extended) vs. its end / the last of all moments (point). To pick out just one example: 'How long did Neil Armstrong set foot upon the moon?' is an irredeemably ill-framed question. In Wittgensteinian terms it is *unsinnig* (and not merely *sinnlos*). By contrast, 'How long did the Apollo 11 mission last?' is not. Given our descriptive metaphysics, then, we think of man's first setting foot upon the Moon as an unextended event in time—a point event. So, as we think of it, in time (like in space) we have both extensions and points. But if there are point events, and if actions are events—why should there not be point actions? Clearly, we cannot answer: 'Because actions must have temporal extension'. That would be to beg the question.

### *3.2 Semantic of Achievement Verbs*

The case is strengthened when we turn to achievement verbs. Obviously, part of the observation employed in the preceding paragraph is due to *Gilbert Ryle*. He made it a commonplace to say that it does not make sense to ask how long I have been *finding* my keys, whereas it is perfectly intelligible to ask how long I have been *searching* for them. Phenomena of this kind led him to distinguish between verbs describing *processes* or

*activities* and those describing *achievements* (Ryle 1949: 149). *Searching* is an activity, *finding* an achievement. Here the temporal contrast in question is quite clearly displayed: While it makes sense to attribute temporal extension to an activity, it does not make sense to attribute it to an achievement. So it's a matter of which category a given verb belongs to. But then, aren't there achievement verbs describing actions (e.g.: *finishing* a paper)? And if so, isn't this, again, an indication that there are actions thought of as having a position in time but no extension therein—point actions?

### *3.3 Speech-acts*

The last case in point rests on J. L. Austin's famous theory of *speech-acts*. In his 1955 William James lectures he inquired into the class of utterances, the issuing of which 'is the performing of an action—it is not normally thought of as just saying something' (Austin 1962: 6). For the sake of brevity, I shall spare you the many now classic examples. But note that actions performed via speech-acts are paradigm cases of actions: They involve a doing by a person, they are intentional under at least some description, and they can be accounted for by giving reasons rather than just citing causes. And note further that we could hardly account for the way in which we as human beings communicate or act if we were to marginalise speech-acts. They play a central role in our communicative and social practice. Now if we apply the Ryle test to them we easily find that very many (if not all) actions performed by speech-acts do not admit of intelligible answers to 'how long?' questions. 'How long have you been taking Mary to be your lawful wedded wife?' and 'How long have you been betting John sixpence it will rain tomorrow?' are as unintelligible as 'How long have you been finding your keys?' So, given our descriptive metaphysics actions performed by the very common and by no means artificial means of speech-acts do not have temporal extension. They are instantaneous, hence: point actions.

## 4. Summary: Some Worries and an Alternative Perspective

At this point the question may come up whether we are *really* forced to accept the conclusions lurking in the discussions of the problems described. Did I really do enough to justify the claim that most modern philosophers think an action must be extended in time? Did I really show convincingly that Korsgaard and Frankfurt are committed to this view? To these charges I readily plead guilty. I would like to add, however, that brevity has a price and that elaborating on these issues would have contributed nothing to the still interesting question whether there is a problem with durationalism *if* we assumed that many modern philosophers are durationalists or *if* we assume that Korsgaard and Frankfurt are committed to that view. But there may be another worry too. Isn't the Ryle argument very derivative and is it really sufficiently developed in the first place? The sceptic might object that the fact that the 'how long' question does not make sense might indicate a lot of things—it isn't clear that this in itself shows that the action is instantaneous. To this I would reply: Well and good, much would be accomplished if it was finally acknowledged in action theory that 'how long' questions concerning actions quite often do not make sense. But let's not forget that the arguments presented here are not designed to show that there *really are* point actions but that actions seem to be *thought of* as being unextended in time—and not only in some rare and exceptional cases but in a wide range of important cases indispensable for our communicative practice. This feature of the descriptive metaphysics of actions has largely gone unnoticed. It is in itself a remarkable setback for durationalism.

So, we are at a crossroads. We might revise our metaphysics and take issue with the cases of event ontology, achievement verbs, and speech-acts. Or we might accept and endorse our descriptive metaphysics of actions and admit that according to it there are point-actions. If so, however, durationalism fails. It may help to note, though, that not all is in danger if we

admit point actions. There are action theories quite capable of accommodating this consequence. Observe, e.g., the theory of Austro-American philosopher (and economist) *Ludwig von Mises*. In his theory of human action (termed *praxeology*) he explicitly reckons with the possibility that action may be temporally unextended:

> Action is not concerned with the future in general, but always with a definite and limited fraction of the future. This fraction is limited, on the one side, by the instant in which the action must take place. Where its other end lies depends on the actor's decision and choice. [...] We may call the fraction of future time for which the actor in a definite action wants to provide in some way and to some extent, the period of provision. [...] Every choice implies also a choice of a period of provision. In making up his mind how to employ the various means available for the removal of uneasiness, man also determines implicitly the period of provision. (Mises 1949: 478)

According to Mises, where the time of provision is the present moment, the action performed is a point action. So there seems to be at least one theory of action leaving room for point actions. Surely, to explain and defend such a theory (which, in the case of Mises, has actually quite a lot of interesting views worth being reconsidered in action theory) is nothing that can be done in passing. I must leave that for another occasion. But what has been established here still bears some weight: (i) There are reasons to assume point actions; (ii) the issue is of importance if we want to devise a theory of action suitable for describing our social and communicative practice; and (iii) there are ways out of a seeming impasse apparently reached when we acknowledge the reality of point actions.[1]

**Michael Oliva Córdoba**

University of Hamburg
Department of Philosophy
Von-Melle-Park 6
20146 Hamburg
Germany
michael.oliva-cordoba@uni-hamburg.de

## References

Austin, John L. 1962: *How to Do Things with Words*, ed. J.O. Urmson. Oxford: Oxford University Press.

Davidson, Donald 1969: 'The Individuation of Events', in Davidson 1989, 163–180.

— 1970: 'Events as Particulars', in Davidson 1989, 181–187.

— 1989: *Essays on Actions and Events*. Oxford: Clarendon Press.

Frankfurt, Harry G. 1978: 'The Problem of Action', in Frankfurt 2009, 69–79.

— 2009: *The Importance of What We Care About*: Philosophical Essays. Cambridge: Cambridge University Press.

Korsgaard, Christine 2008: *The Constitution of Agency*. Oxford: Oxford University Press.

Mises, Ludwig von 1949: *Human Action*. San Francisco [4]1996: Fox & Wilkes.

Ryle, Gilbert 1949: *The Concept of Mind*. London: Hutchinson.

[1] Many thanks go to Nathan Wildman and Ben Hoefer for comments and suggestions. I am particularly grateful to Rolf W. Puster for discussion and comments on various drafts.

# Bennett on Dismissivism

Laura Cecilia Porro

In this paper I introduce the topic of dismissivism in metaphysics. In particular I engage with the view of a recent important philosopher, Karen Bennett, who has addressed this issue in Bennett (2009). A discussion of Bennett's work will prove useful to delving into two important topics. First of all Bennett delves into the reasons why one may want to be a dismissivist. She describes three main reasons (antirealism, semanticism, and epistemicism). Bennett's three reasons may constitute a guideline or framework to help me diagnose what is going on in various metaphysical debates. Secondly, Bennett offers a strategy to find out when a debate should be dismissed for epistemic reasons. I will explore some general features of this strategy and investigate whether it can be generalized beyond the case studies Bennett addresses.

## 1. Three Reasons to be a Dismissivist

The purpose of this section is to present and describe Bennett's view. Bennett (2009) focuses on questions such as: what is the type of disagreement philosophers are having? Is it substantive or not? What are the ways, if any, to solve metaphysical disagreement? She discusses the dismissivist attitude in metametaphysics. She uses 'dismissivism' as a label 'for the view that there is *something* deeply wrong' with (at least some) metaphysical questions (p. 39). In particular, in her paper, she addresses three questions:

1. What are the possible reasons to be a dismissivist?
2. How do we assess whether it is appropriate to be a dismissivist in each specific case?
3. What is the appropriate method of this inquiry?

Bennett thinks philosophers like Putnam, Sidelle, and Carnap endorse dismissivist attitudes towards some metaphysical questions in that they would answer them 'who cares?'. She calls them 'neo Carnapian naysayers' (p. 38). She uses this expression to refer to all philosophers who would react to at least some metaphysical questions by saying 'who cares?'.

Bennett lists three different reasons to be a dismissivist about metaphysical disputes. Bennett does not argue for the claim that these are the only three possible reasons to be a dismissivist. The first reason to be a dismissivist Bennett describes is called 'antirealism' and is defined as follows (Bennett 2009: 39):

> There is no fact of the matter about whether or not there are *F*s. 'There are *F*s' does not have a determinate truth-value.

For instance, take the dispute whether there are Fs from the point of view of someone who thinks that F is a metaphysically vague object (leave aside doubts about the consistency of the view, for the sake of the example). According to Van Inwagen, as paraphrased by (Hawley 2001b: 5)[1]:

> [...] there are borderline cases of lives: an example may be the activity of the simples in a region we would ordinarily describe as 'occupied by a virus'. If it is indeterminate

[1] Let us leave aside doubts about the consistency of Van Inwagen's view, for the sake of the example.

> whether the activity of some things constitutes a life, then it is indeterminate whether those things compose an organism and thus, for van Inwagen, it is indeterminate whether they compose anything at all. Roughly speaking, it is a vague matter whether the virus exists, and a vague matter whether there are any viruses.

If it is vague whether there are viruses, then 'There are viruses' does not have a determinate truth-value. In this scenario, a philosopher who likes antirealism might think that there is no point discussing whether *F* exists. Rather, the dispute should be dismissed.

Bennett thinks antirealism is not a good reason to be a dismissivist, because she is not 'entirely sure what it means' (p. 40). She explicitly drops the discussion about it in her paper, because this view does not play a central role in it. Antirealism will not play an important role in what follows, so I do not delve into it any further.

The second reason to be a dismissivist is called 'semanticism' (p. 40):

> The dispute about whether there are *F*s is purely verbal. The disputants assign different meanings to either the existential quantifier, the predicate *F*, or the negation operator, and are consequently just talking past each other.

Bennett highlights that semanticism is a different view from antirealism. As she points out, two disputants can agree that there is a fact of the matter about whether there are *F*s, yet disagree about what 'There are *F*s' means, because they use words in different ways. An example of semanticism, according to Bennett, is Hirsch (2002).

The third version of dismissivism, epistemicism, has two formulations, a strong and a weak one. Bennett defends the weak one, whose definition is as follows (Bennett 2009: 42):

> Disputes about the truth value of 'There are *F*s' are not verbal disputes. But there is little justification for believing either that it is true or that it is false.

Bennett highlights that this view is compatible with as strong a form of realism as one wants. Note that the definition just given above does not mention whether disputants think there is a fact of the matter about 'There are *F*s'. This means that the view can be developed in a realist spirit as well as in an antirealist one. Epistemicism interestingly highlights the relationship between verbal disputes and dismissivism. It is easy to slide from the idea that if a dispute is verbal, then it should be dismissed, to the idea that if a dispute is non-verbal, then it should not be dismissed. Bennett emphasises that it is possible to have reasons to dismiss non-verbal disputes.

Looking at the features of antirealism, semanticism, and epistemicism I can formulate two very general reasons to be a dismissivist:

(a) disputants seem to talk about *x*, but they are actually talking about *y*;

(b) disputants seem to talk about *x*, but there actually is no point talking about *x*.

For instance, (a) is basically the definition of Bennett's semanticism. (b) on the other hand is the general scheme of Bennett's antirealism and epistemicism. The difference between antirealism and epistemicism lies in the specific reason why there is no point talking about *x*. (a) and (b) are dismissivist attitudes, although they need to be implemented with more precise reasons why they hold of some debate.

Bennett argues that two recent popular metaphysical debates should be dismissed for epistemic reasons. The first one is the composition debate, which tries to answer the question: 'When do simples compose a larger thing?'. Believers answer: always. Nihilists answer: never. The second debate is about colocation: 'Can different objects spatio-temporally coincide?' Multi-thingers answer: yes. One-thingers answer: no. Bennett argues that these two debates are genuine, i.e. they are not verbal. This means that in these debates disputants are not having a verbal misunderstanding, nor they are discussing what is the

correct way to use words in English. Rather, these are substantive questions about existence, however we cannot *know* which the best answer is.

Bennett shows this by means of an argument that starts from conditions a debate needs to meet, in order for it to be dismissed for epistemic reasons. I am going to summarize the argument now. It is appropriate to dismiss a debate for the epistemic reason listed above when the following conditions are met (note that these are sufficient but not necessary conditions):

(1) One of the disputants postulates more entities than the other[2];

(2) 'Both sides try to minimize their differences from their opponents' (Bennett 2009: 62)

(a) The disputant who postulates more entities 'insists that her extra ontology is nothing over and above what the [other disputant] accepts' (*ibid.*);

(b) The disputant who postulates less entities 'tries to recapture most of the claims that the [other disputant] accepts' (*ibid.*).

If (1) and (2) then:

(3) 'It is not obvious that the low-ontologist's view is simpler than the high-ontologist's view' (Bennett 2009: 63);

(4) 'The problems for the high-ontologist rearise for the low-ontologist' (*ibid.*).

If all this is the case, then the debate should be dismissed for epistemic reasons.

I am going to show how the debate about composition meets the above requirements, according to Bennett. According to Bennett in the debate about composition, the believers postulate more entities than the nihilists, because the former postulates more types of entities. According to the believer (the high-ontology side) composite objects and simples exist, while according to the nihilist (the low-ontology side) only simples exist. Condition (1) is thus met.

Bennett also shows that believers try to minimize their differences from the nihilists by arguing that composite objects are nothing over and above simples. The believer says: 'Necessarily, if there are simples arranged *F*-wise in region *R*, then there is an *F* in *R*.' (Bennett 2009: 48). The believer tries to show the nihilist that composite objects supervene on the simples. Thus once one accepts simples, composite objects are an ontological 'free lunch'. The believer is trying to convince the nihilist that composite objects are not an extra addition to one's ontology, but rather automatically come once the existence of simples is acknowledged.

From the opposite side, the nihilist tries to minimize the difference between the claims he makes and the claims the believer makes. Since for the nihilist there are no composite objects, claims such as 'There is a table' are either inaccurate or false (different versions of nihilism endorse one or the other). The nihilist then tries to recapture the believer's claim, with periphrases such as: 'There are simples arranged table-wise'. These examples show how the composition debate meets conditions (2a) and (2b).

From these first remarks, Bennett argues that it is not possible to establish whether the nihilist's view is simpler than the believer's or vice-versa (condition (3)). This happens because on one hand the believer combines a less parsimonious ontology with easily understandable claims (such as 'There are tables'), while on the other hand the nihilist combines a more parsimonious ontology with not-so-easily understandable claims (such as

[2] From now on I will follow Bennett in referring to the disputant who postulates more entities as the 'high-ontologist', and to the disputant who postulates less entities as the 'low-ontologist'.

'There are particles arranged table-wise'). Bennett clearly explains this in the following terms (Bennett 2009: 64):

> The high-ontologist multiplies objects while the low-ontologist multiplies properties. [The nihilist] buys her way out of ontology with the coin of ideology. So even if the low-ontologist wins the battle of ontological commitment, he does not win the war of simplicity. On at least one way of reckoning simplicity, the two come out roughly on a par.

Lastly, Bennett describes in details four challenges that arise both for the believer and the nihilist. This shows, according to Bennett, that both views have the same amount of negative features, since they run into the same issues. A detailed analysis of these four challenges is not relevant to the pursuit of the purpose of this paper, thus I am not going to discuss them.

According to Bennett, since the composition debate meets the four requirements listed above, it should be dismissed for epistemic reasons. The question whether simples compose larger things is a genuine one, and the debate over the correct answer is not due to a misunderstanding about what 'simples', 'things' or 'exist' mean. However the two views seem to display the same positive and negative features, while neither is more complex than the other. Thus we have no good reasons to prefer one over the other. In turn this is, according to Bennett, a good reason to dismiss the debate about composition, i.e. stop trying to find out which view is right.

## 2. Moving Forward

So far I have discussed Bennett's work in general, describing her taxonomy of reasons to be a dismissivist. I have also described her argument strategy to show when some debate should be dismissed for epistemic reasons. Now I turn to analyse and criticize her argument strategy in favour of epistemicism. This is interesting because if Bennett's argument is sound, it can be used as a general 'test' for all metaphysical debates. If some metaphysical debate meets conditions (1) to (4) above, then it should be dismissed for epistemic reasons. The purpose of this section is to discuss this argument in detail, to check whether it is sound and applicable in general.

Before I start it is important to highlight an aspect of Bennett's thought about this. Bennett states the argument in general terms. However she never explicitly says that this argument is meant to be a general strategy that we can use to diagnose other metaphysical debates. Her conclusion is that the composition and colocation debates should be dismissed for epistemic reasons, and not that all debates that fit the described argument strategy should be dismissed for epistemic reasons. However, since she states the argument in general terms, it is reasonable to think that Bennett hopes that her strategy could be used in other debates. Even if this is not Bennett's intention though, I still think it is a possibility worth exploring, in the context of this work. Given these remarks it should be clear that when I argue that Bennett's strategy cannot be generalised, I do not mean this as a criticism of her view, because she might not have meant it to be generalisable.

Let us start by briefly summarising Bennett's argument:

(1) If in a debate there are high and low-ontology sides;

(2) if a debate is difference-minimizing;

(3) then no view is simpler than the other;

(4) and both face similar issues;

(C) then we should dismiss the debate for epistemic reasons.

The argument holds in one direction and is not a bi-conditional, thus it could be the case that a debate should be dismissed for epistemic reasons even though it is not difference-minimizing. I am going to challenge (1) and (2), but it is important to note that this does not block the argument from (3) and (4) to (C). The strategy I will use to challenge (1) and (2) is motivated by two reasons. First of all, my purpose is to check if we can generalize this argument to all metaphysical debates (apart from composition and colocation), thus I am going to explore whether we can expect conditions (1) and (2) to be met in general by metaphysical debates. In particular, in the case of (2) I would like to find out whether disputants *should* difference-minimize, or rather whether difference-minimizing is a mistake[3].

The second reason why I want to challenge (2) comes from a suggestion Bennett makes in her paper (Bennett 2009: 72, original emphases):

> One way to resist the lessons I am drawing is to say that it is a mistake to difference-minimize. In particular, one way for the low-ontologist to resist is to embrace his view with a braver heart, and *stop trying to say everything the other side says!*

This suggestion amounts to stopping step (2b) listed above. I am going to delve into Bennett's suggestion, exploring the reasons why we could consider this type of difference-minimizing a mistake.

In what follows I will firstly express some worries about (1), then show that there is no good reason to recapture opponent's claims (challenge to 2b), and finally show that only in a very small number of debates it is possible for the high-ontology side to downplay excess ontology (challenge to 2a). My remarks will lead me to be skeptic about the possibility to apply Bennett's argument strategy to other metaphysical debates.

### *2.1 A problem With High and Low Ontology Sides*

The first condition a debate needs to meet is that one side of the debate postulates the existence of more entities than the other does. Bennett thinks that this condition is met by both the composition and colocation debates, because in both cases one side (believers/multi-thingers) postulates more *types of entities* than the other (nihilists/one-thingers). I would like to raise the question of how we count entities. In order to answer this question we need to decide first of all *what* we count. Various options open up here, which I describe drawing on an important distinction made by Schaffer (2008)[4]. We could count the entities each theory says exist, or we could count the entities each theory says are fundamental. Or we could, as Bennett does, count the types of entities whose existence is acknowledged by each theory.

Bennett does not state the rationale behind her choice. This is a problem, because had we chosen to count the number of entities, rather than the number of types, both the composition and colocation debates would fail to meet condition (1). This happens because if we count the number of entities, both parties in each debate turn out to be committed to an infinite number of entities, thus there are no high/low-ontology sides. From this it emerges that Bennett is somehow begging the question. She says that a debate should be dismissed for epistemic reasons if it meets conditions (1) to (4), and then chooses a counting method that makes the debate meet condition (1), without stating any reasons. It seems wrong to adopt a counting method or another just to make a debate fit the argument strategy. Rather, a

[3] I am aware that Bennett herself does not ask this question, and I am not criticizing her on the ground that she does not. I just want to expand Bennett's project in the direction I highlighted in the introduction, and this is the reason why I ask this question now.

[4] Schaffer does not address directly the issue of how to count entities, however he highlights that `the quantifier commitments are what a theory says exists, while the truthmaker commitments are what a theory says is fundamental' (Schaffer 2008: 18). For further discussion on this see Hawley (2001a).

counting method should be adopted for independent theoretical reasons and upon reflection on quantitative and qualitative parsimony.

If this is right, then it seems that before being able to apply Bennett's argument strategy, we need to figure out the best method for counting entities, which in turn requires to delve into a methodological enquiry about the concept of parsimony. Unfortunately, the status of the debate about parsimony in the literature is not encouraging, because philosophers do not seem to have reached any agreement about a definition of parsimony (see for example Nolan (1997), especially section 3). We thus find ourselves stuck with the following two horns of a dilemma:

– either address questions about what parsimony is, and suspend questions about dismissivism until those are answered;

– or use a strategy to find out whether a debate should be dismissed for epistemic reasons that does not need to answer questions about parsimony first.

If this is correct, this lowers the chances to apply Bennett's argument strategy to other debates.

### *2.2. A Problem With Up-playing Expressive Powers*

I now turn to discuss the attempt of the low-ontology side to minimize the difference between his claims and the high-ontology side's claims. First of all, Bennett discusses different possible strategies the nihilist has to preserve ordinary judgements about what there is and is not (Bennett 2009: 57-58). Then Bennett says (Bennett 2009: 58-59):

> All nihilists want somehow to recapture the claims that the believer takes to be true[5]. [...] As long as they do not simply proclaim statements about composites false, and stop there, revolutionary nihilists are still up-playing their expressive power. They are still difference-minimizers.

This implies:

– Recapturing believers' claims is a way to minimize the differences between nihilists and believers.

– Up-playing expressive power has the purpose of difference-minimizing.

Bennett is not explicit about the reason why nihilists try to recapture believers' claims. In general, Bennett says (Bennett 2009: 72):

> All the participants [i.e. believers and nihilists] want somehow to preserve our ordinary judgements of persistence, of sameness and difference, of what there is and isn't.

Intuitively, the reason why the nihilist wants to preserve ordinary judgements is that he does not want to say: 'there are no toasters; revise your breakfast plans' (Bennett 2009: 58). The more interesting question is why nihilists try to recapture believers' claims. This question is important because if it turns out that nihilists have no reason to do so, this should make us suspect that there are issues with difference-minimizing. However it is not easy to find out why nihilists recapture believers' claims. It is apparent from p. 57-58 that Bennett takes the task of preserving ordinary intuitions and claims and the task of recapturing believers' claims to be related. Since Bennett does not say anything about the rationale behind difference-minimizing in this case, I think I can formulate Bennett's thought in two different ways. The first version is most supported by textual evidence and is: (1-) Nihilists try to recapture believers' claims by up-playing their expressive powers. This claim refrains from stating the reason why this is the case. A second possible version tries to interpret Bennett's words, and

[5] Bennett argues that also one-thingers try to recapture multi-thingers' claims.

thus runs the risk of not reflecting properly Bennett's thought: (1) Nihilists try to recapture believers' claims by up-playing their expressive powers, because they want to preserve ordinary judgements. The fact that I cannot show that Bennett definitely endorses (1) is not a big issue, because what I am about to say should undermine both (1-) and (1).

I am going to argue against claim (1), but what I say should also show that there are issues with (1-). First of all let us clarify the claim itself. There are two ways of interpreting it:

(1) Nihilists try to recapture some of believers' claims, ...

(1*) Nihilists try to recapture all believers' claims, ...

Bennett explicitly endorses (1) and not (1*). I will discuss two sets of reasons against the validity of (1).

1 Bennett misinterprets the debate. It is not the case that nihilists try to recapture believers' claims, because the purpose 'trying to preserve ordinary judgements' cannot be achieved by means of recapturing believers' claims.

2 In general any philosopher should not try to recapture his opponent's claims.

If ♋ and ♌ are correct, they show that difference-minimizing in this way is a mistake. Note that my second argument is more general than the first one. If one disagrees with ♋ and agrees with Bennett that nihilists are in fact difference-minimizing, this is not enough to show that my second argument is wrong. Furthermore, if one thinks that (1) misrepresents Bennett's thought, and (1-) is correct, ♌ shows that even if the nihilist is in fact recapturing believers' claims, he is thus making a mistake and should stop doing so.

Let us start by analysing what is going on in the composition debate. I would like to argue, against (1), that the nihilist up-plays his expressive powers to recapture ordinary judgements, rather than believers' claims. Consider Bennett's following example from the composition debate.

(B) Believer's claim: 'There are tables and they are composite objects'.

(N) Nihilist's claim: 'There are particles arranged tablewise'.

Bennett holds that (N) is a way of recapturing (B). However it seems that both (B) and (N) are also ways of recapturing:

(O) Ordinary claim: 'There are tables'.

I would like to argue that (N)'s first aim is that of recapturing (O), rather than (B). In order to show this, consider the following example:

(B1) Believer's claim: 'The right ontology should countenance composite objects and simples'.

(N1) Nihilist's claim: 'The right ontology should countenance only simples'.

The difference between these two examples will help me clarify why I disagree with Bennett's claim. Bennett argues that (N) is a way to recapture (B), but also that the nihilist does not try to recapture (B1). Bennett and I agree on the second example, i.e. that the nihilist has no interest in recapturing (B1) and does not try to do so. Doing so would not do any good to the nihilist's view[6]. My argument is based on thinking about what the nihilist wants to achieve.

[6] Note that this is a counterexample to (1*), because it presents a case in which the nihilist is not trying to recapture a claim made by believers. However, as already highlighted, Bennett never claims that the nihilist tries to recapture all believer's claims (see Bennett 2009: 62). Bennett can thus say that the latter example does not count as a counterexample to her claim, because (B1) is one of believer's claims that the nihilist does not try to recapture. Nonetheless, I want to argue not only against the claim that

Bennett says the nihilist wants to preserve ordinary judgements, thus showing his opponents and ordinary people that his view can answer the composition question and does not lose on the ground of expressive power. It is not clear at all how exactly recapturing believers' claims helps the nihilist preserve ordinary judgements, and Bennett does not say anything to enlighten this. The tricky point, which is at the root of the disagreements between me and Bennett I suspect, is that in the case of claims about ordinary objects, the believers' claims are very close to ordinary claims. (B) is similar to (O) more than (N) is similar to (O), because the believer is closer to common sense in this respect than nihilists are. This may give the impression that the nihilist is in fact trying to recapture the believers' claims. However, if we think about (B1) and (N1), it becomes apparent that the nihilist has no reason to recapture (B1). This is because recapturing (B1) does not help the nihilist preserve ordinary judgements, or boost his position's expressive power, or answer the composition question. If this is true in the case of (N1) and (B1), why would things be different in the case of (B) and (N)? Recapturing (B) does not help the nihilist preserve ordinary judgements or show how the nihilist can answer the composition question. Thinking about (N1) and (B1) helps understanding what is happening in the debate, whereas thinking only about (B), which is so similar to the ordinary judgement's claim, can be misleading.

So far I have argued that Bennett's diagnosis of what is happening in the composition debate is wrong. However we can make a further step ahead and make a more general claim. I want to argue that in general it is a mistake for a philosophical view to recapture his opponents' claims. Thus even if one was not convinced by my argument ♋ and thought that nihilists are in fact difference-minimizing, I am going to give him further reasons to be worried about this aspect of difference-minimizing. Let us think about the purposes philosophical views have (in no particular order):

- explain what it is supposed to explain;
- have its position understood;
- preserve ordinary judgements[7];
- show its opponents are wrong or at least worse off;
- ...

Re-expressing or re-stating opponent's claims is useful in order to show why the opponent's position is wrong, however it does not serve any of the other purposes just listed. Moreover, even when one recaptures his opponent's claims in order to prove them wrong, this is most definitely not a case of difference-minimizing. The reason why one recaptures his opponent's claims here is rather to maximize the differences between himself and the other, to show why he is right and the other is wrong. A very clear example of what I am saying comes from the metaphysical debate between tropes' and universals' ontologies. Take the universalist's claim that a table is red because it instantiates the property of 'redness'. On the other hand the tropist claims that a table is red because it has a trope of 'being this shade of red'. Both claims are ways of recapturing the ordinary claim 'The table is red'. The tropist's claim (granting just for the sake of the argument that he plays the role of the low-ontology side) patently does not try to recapture the universalist's claim, rather it explains the ordinary claim in tropes' terms.

---

sometimes nihilists recapture believers' claims, but also against the weaker claim that nihilists ever try to recapture any of opponents' claims.

[7] A further question can be raised, i.e. 'should any philosophical position reconstruct ordinary judgements?'. This is an interesting question, however it is not relevant to my enquiry at present. I am here only focusing on arguing against Bennett and thus taking for granted that at least some philosophical views try to preserve ordinary judgements.

### *2.3 A Problem With Downplaying Excess Ontology*

The third challenge concerns step (2a) of difference-minimizing. In this case, it is the high-ontology side which tries to minimize the differences between his less parsimonious ontology and the low-ontology side's more parsimonious one. I am here raising a doubt, in a very different way from the previous paragraph. This time I am not going to argue against Bennett, rather I am going to agree with her diagnosis of the composition debate in this respect. I only focus on this aspect of difference-minimizing with respect to the possibility of generalizing Bennett's argument.

I have already described in the previous section how the believer tries to minimize the differences between his ontology and the nihilist's ontology. I now want to highlight that the reason why the believer can argue the way he does is because the debate between believer and nihilist satisfies the following condition:

believer's ontology = nihilist's ontology + composite objects.

The nihilist shares some ontological commitment with the believer, i.e. the commitment to the existence of simples. The believer's ontology is exactly the same as the nihilist's with the addition of composite objects[8]. This is the reason why the believer's difference-minimization strategy works. If the believer shared no ontological commitments with the nihilist, his strategy would not work.

I think Bennett is correct in holding that the believer tries to minimize his differences from the nihilist. My worry on this matter concerns the possibility to re-use Bennett's strategy in other metaphysical debates. I think that very few debates meet the condition that one disputant's ontology is exactly the same as the other disputant's plus some other commitment. Think for instance of the debate about the nature of modality, the existence of temporal parts, the existence of tropes or universals, ..., in all these cases disputants do not share any ontological commitment, and thus fail to meet condition (2a).

### *2.4 Conclusion*

With all these remarks I have provided some reason to undermine the first and second step of Bennett's argument. As highlighted above, I have said nothing against steps (3) and (4) of the argument and I will not. What I have achieved is to diminish the hope that we can use Bennett's argument strategy as a handy 'dismissivist-test' machine for all metaphysical debates, either because some steps are dubious, or because very few debates meet the requirements. What I have shown is that it is really hard to generalize from the features of one debate to other debates. It thus seems that most of the work to find out whether a debate can be dismissed has to be done 'manually', and on a case-by-case basis, without much help from Bennett's scheme.

**Laura Cecilia Porro**

University of St Andrews (United Kingdom)
lauracecilia.porro@gmail.com

[8] Some believers think that there is gunk. Such believers agree with what the nihilists claim exists, but they disagree about the nature of such things. The nihilist argues that only simples exist, while a believer in gunk thinks that even if chairs are made of parts, then also those parts are made of parts, and so on ad infinitum. Since Bennett does not delve into gunk lovers, and it will not be relevant for my work to consider the details of such option, I will not address this further.

## References

Bennett, K. (2009), Composition, Colocation, and Metaontology, in D. Chalmers,

D. Manley & R. Wasserman, eds, '*Metametaphysics*'.

Hawley, K. (2001a), Ontological Innocence, in D. Baxter & A. Cotnoir, eds, '*Composition as Identity*'.

Hawley, K. (2001b), 'Vagueness and Existence', *Proceedings of the Aristotelian Society* CII(2), 125–140.

Hirsch, E. (2002), 'Quantifier variance and realism', *Philosophical Issues* 12, 51–73.

Nolan, D. (1997), 'Quantitative Parsimony', *The British Journal for the Philosophy of Science* 48(3), 329–343.

Schaffer, J. (2008), 'Truthmaker commitments', *Philosophical Studies* 141, 7–19.

# 3. Logik und Wissenschaftstheorie

# Regularity Theories of Mechanistic Constitution in Comparison

Jens Harbecke

This paper examines the relation of two regularity theories of mechanistic constitution developed by Harbecke (2010) and Couch (2011). By comparing the central definitions it is shown that, irrespective of various similarities, the two approaches differ in important details. Taking into account these differences, Harbecke's theory in comparison to Couch's theory will be judged as more adequate for the definition of mechanistic constitutions.

## 1. Introduction

The question about the relation of the cognitive abilities of human beings to the neural mechanisms of the human brain is a central topic in the philosophy of mind. Whether the mind with its various cognitive functions and processes is identical to the brain and, hence, subject to the laws of nature, or whether it is in a special sense independent of the human body and causally interacts with it in a contingent way has consequences for the understanding of the mind. While in the philosophy of mind since the beginning of the 1980s it had been widely acknowledged that the relation between the mental and the physical is adequately described using the term 'supervenience' (cf. Kim 1984), this assumption has recently been called into question. The term 'supervenience' seemed to offer a plausible relation of determination between body and mind, without presupposing an identity of the mental and the physical. On the basis of recent developments in the sciences, however, there is a growing consensus that the mental and neuronal processes are more systematically connected than the classical notions of supervenience can express.

An important step towards this realization was a detailed study of recent findings of neuroscience from representatives of the 'mechanistic approach' to neurobiological explanation (cf. Machamer et. al 2000). It was convincingly argued that successful neuroscientific explanations of cognitive phenomena are characterised, on the one hand, through a procedural understanding of cognitive abilities such as representing, concluding, deciding, calculating etc. and, on the other hand, by a method of analysing the phenomena in terms of their local and temporal relationships to certain neural mechanisms. Proponents of the mechanistic approach have termed the systematic relation of cognitive phenomena and neuronal mechanisms 'mechanistic constitution'. The term plays an important role for the mechanists' theory of explanation as well as for their favoured ontology.

Although the mechanistic approach soon received widespread support from philosophers of science, the term 'mechanistic constitution' remained surprisingly vague in the original key contributions. Subsequently, various different definitions have been proposed, none of which has so far reached the status of a standard definition. Even the worse, some of the proposed definitions are inconsistent with another. Craver uses the term 'mechanistic constitution' to describe a relationship between components, or objects, and phenomena. Other authors such as Fazekas and Kertesz (2011, 372) equate mechanistic constitution with a mereological relationship, i.e. a relationship between individuals or objects. Again others, such as Soom

(2007, 83) and Mandik (2011) seem to identify the notion with supervenience, i.e. a relationship between classes of properties.

The most detailed definition of the term 'mechanistic constitution' has so far been developed by Harbecke (2010) and Couch (2011), who independently of each other have defended a regularity theory of mechanistic constitution.[1] The central idea of these approaches lies in the assumption that mechanistic constitution can be defined in extensional language. The authors transfer instruments successfully deployed in regularity theories of causation to the definition of mechanistic constitution.

The current paper examines the precise relationship between the two regularity theories of mechanistic constitution by analysing in detail their commonalities and differences. Furthermore, an evaluation of strengths and weaknesses of the two alternatives is offered. The investigation proceeds by the following steps: First the relevance of the problem for the conceptual basis of neurobiological explanation is discussed (Section 2). Afterwards, the core assumptions of Couch's and Harbecke's theories are presented (Section 3 and 4), before the commonalities (Section 5) and differences (Section 6) of the approaches are investigated. Section 7 contains an evaluation of the theories against the background of the preceding investigations. Section 8 summarizes the results and makes some suggestions concerning possible future research on regularity theories of mechanistic constitution.

## 2. On the Relevance of the Problem and the Method of Analysis

The general philosophical project in which regularity theories of mechanistic constitution are embedded applies a particular methodology, which aims to establish philosophical conclusions on the basis of a logical analysis of prototypical explanations in neuroscience. It is then attempted to develop a structural model of successful explanations in neuroscience, which is descriptively adequate as well as normatively binding. The contributions to this project therefore typically reconstruct pertinent results of neuroscientific research and reflect them philosophically. Harbecke reports a widely accepted theory of the localisation of the cognitive function of spatial representation in the brain of rats (cf. T. Bliss und T. Lømo 1973; Lømo, T. 2003; R. Morris 1984; R. Morris, E. Anderson, G. Lynch and M. Baudry 1986). According to this theory, spatial representation in rats is constituted through a 'neuronal map' in the hippocampus, which again is constituted by a long term-potentiation (LTP) of pyramidal cells in area CA1 of the hippocampus, where the LTP of these cells is constituted essentially through the activation of the NMDA-receptors at their synapses. The theory is supported by various experiments, which have shown a correlation of an impairment of spatial orientation - indicated through disorientation and ineffective food searching in an experimental environment - and the blockade of NMDA-receptors or simply the surgical removal of the hippocampus.

In this sense, the celebrated neuroscientific explanation of spatial representation in rats uses the following four theoretical terms, whose referents are linked by mechanistic constitution:

1. Spatial Representation
2. Generating a neuronal 'map' in the hippocampus
3. LTP of pyramidal neurons
4. Activation of NMDA – receptors

[1] Harbecke's definition uses the term 'mechanistic constitution', whilst Couch chooses the notion of 'constitutive relevance'. However, it is clear that both authors wish to define the same kind of relation.

The philosophical investigation is mainly concerned with the question about the nature of the relation of constitution. Apart from certain ontological dimensions, this project has a direct relevance for neuroscientific methodology. This is indicated by the fact that the language in which neuroscientists present their results often displays a striking disunity. To characterize the relation between the described phenomena colloquial language terms such as "is responsible for" (Bliss & Lomo 1973, 331), "gives rise to" (Morris et. al. 1986, 776), "plays a crucial role in" (Davis et. al. 1992, 32), "contributes to", "forms the basis of" (Bliss & Collingridge 1993, 38) and "is constitutively active in" (Malenka et. al. 1989, 556) are all in use. The non-unified choices in language indicate that the nature of the described relation in neuroscience remains itself somewhat unclear. While a successful analysis of this relation answers the question of the relation of cognitive processes to the neuronal mechanisms of the human brain, it also makes a contribution to the clarification of neuroscientific terminology.

## 3. Mechanistic Types of Regularities: Harbecke

Harbecke's theory is centred on the notion of a 'minimal theory', which has been applied successfully in regularity analyses of causation in order to solve the problem of spurious regularities. A minimal theory is based on a biconditional in the form "$X1 \vee X2 \vee ... \vee Xn \leftrightarrow Y$", where 'X1', 'X2', ..., 'Xn' stand for conjunctions of mechanistic properties or types, and 'Y' stands for a to-be-explained phenomenon that as well is a property. Such a biconditional is a 'minimal theory' if each of X1, X2, ..., Xn is minimally sufficient, or an 'INUS-condition' (cf. Section 4 below), of Y, and if $X1 \vee X2 \vee ... \vee Xn$ is minimally necessary for Y. The definition of mechanistic constitution offered by Harbecke explains true minimal theories as descriptively adequate for the relation in question, if the types occurring therein fulfil certain further conditions. According to this definition a mechanistic type φ constitutes a mechanistic type ψ ("Cφψ") if, and only if:

(i) φ is part of a minimally sufficient condition φ&X1 of ψ, such that...

(ii) φ&X1 is a disjunct in a disjunction $\varphi\&X1 \vee X2 \vee ... \vee Xn$ of minimally sufficient type conjunctions that is minimally necessary for ψ, such that...

(iii) if φ and X1 are co-instantiated, then their instances are a mereological part of an [an individual that instantiates] ψ, and such that...

(iv) the [individual instantiating] ψ mentioned by (iii) is a mereological part of [an individual that results from a fusion of the individuals instantiating φ and X1 mentioned by (iii)]. (Harbecke 2010, 277)

According to Harbecke, mechanistic constitution is a second-order relation between properties or types. The author explains that a mechanistic property such as LTP is to be understood in a minimal way as the set of all events which fall under the predicate "...is/instantiates a LTP". With this idea it is suggested that certain kinds of objects are logically "built into" the properties, and that mechanistic properties are understood as dynamic properties with an input state and a final state.

A mechanistic property φ is then believed to constitute a mechanistic property ψ always relative to at least one complex mechanism φ&X1 involving sometimes more mechanistic properties. These are coinstantiated in a regular but non-redundant way with the constituted property ψ ("φ is a part of a sufficient condition φ&X1 of ψ"). The instances of the mechanistic types standing in the constitution relation are always mereologically connected, i.e. the properties are instantiated at the same place at the same time (cf. conditions (iii) and (iv)). Finally, the definition allows for alternative constituents ("φ&X1 is a disjunct in a

disjunction φ&X1 ∨ X2 ∨ ... ∨ Xn of minimally sufficient type conjunctions that is minimally necessary for ψ “).

Harbecke’s definition models various features of neuroscientific explanations. In particular, it seems plausible that the prototypical theory of spatial representation presented in Section 2 postulates neural mechanisms as explanantia, whose instantiations are not themselves sufficient for the to-be-explained phenomenon, but only in connection with further conditions. Moreover, it is clear that the generation of LTP occurs at the same place and time as the spatial representation. In other words, the relation of mechanistic constitution is not causal but a simultaneous one.

Moreover, the theory explicitly allows for alternative constituents, i.e. it does not require that the to-be-explained cognitive phenomena are coextensional with certain neuronal event types. This corresponds to important empirical findings such as the fact that LTP occurs in different areas of the hippocampus on the basis of different micro mechanisms (cf. Urban & Barrionuevo 1996). At the same time, the definition offers a criterion for the reduction/non-reduction of cognitive phenomena to neural mechanisms. A mutual relation of constitution implies a coexensionality. And since most theorists consider the contextionslity of properties as sufficient for property identity, a reduction is implied.

## 4. Constitutional Dependence of Tokens: Couch

In an attempt to fathom the nature of mechanistic constitutive relevance (=mechanistic constitution; cf. footnote 1), Couch invokes the idea of an INUS condition as it was originally developed by Mackie (1974).2 According to Couch, “the components of a mechanism that realize a [cognitive] capacity should be seen as INUS components.” The author defines a “(...) a relevant part, then, as an insufficient but nonredundant part of an unnecessary but sufficient mechanism that serves as the realization of some [cognitive] capacity” (Couch 2011, 384). Couch provides the following example to illustrate the basic notion:

> Suppose we have a complex structure ABCD that serves to realize a capacity F. Suppose, further, that, of the parts present, only ABC are needed for the presence of F on the occasion, and the other part D is an extra part that serves no role in the presence of the capacity. In this case, the individual components A, B, C, are inus components for the capacity F, and D is an irrelevant part that should be ignored in giving the explanation. (...) Furthermore, it should be apparent from this that the complex structure that consists of ABC together with its organization serves as the mechanism for F. (Couch 2011, 384)

The crucial difference to causal INUS components lies in the fact that, in the case of mechanistic constitution, constituted phenomena and components are present at the same time. As Couch argues, an “(...) effect of a cause is typically believed to occur after the cause (...). In the case of a mechanism, though, the capacity is thought to be present at the same time the mechanism that realizes it is present.” (Couch 2011, 385) Couch conceives of the parts represented by ‘A’, ‘B’, ‘C’, and ‘D’ as tokens, i.e. as individual events or component activities that involve objects and properties (cf. Couch 2011, 384).

Moreover, according to Couch, for a conditional ABC → F to be constitutionally interpretable it has to refer to a necessary connection (i.e. “ ABC → F”). He emphasizes, however, that this assumption does not yet determine a specific kind of necessity (see Couch 2011, 386).

---

[2] Mackie’s definition of an INUS-condition is the following: “It is an insufficient but non-redundant part of an unnecessary but sufficient condition.” (Mackie 1974, 62).

## 5. Material Commonalities

The regularity theories developed by Harbecke and Couch display some obvious similarities. To make these explicit is the aim of this section. Probably the most striking similarity is the fact that both approaches base their definition on the notion of an INUS condition. Couch uses the term explicitly in the cited paragraph (Couch 2011, 384) while Harbecke uses the synonymous term of a minimally sufficient condition. The notion is found in item (i) of the definition of Harbecke (2010, 277): "A mechanistic type φ constitutes a mechanistic types ψ (Cφψ) if and only if: (i) φ is part of a minimally sufficient condition φ&X1 of ψ (...)".

Additionally, both authors emphasize that the relata of the constitution relation, unlike those of causation, overlap spatio-temporally. This assumption is expressed under items (iii) and (iv) of the definition of Harbecke, which demand a mereological restriction for those individuals that instantiate the relata of mechanistic constitution. As Harbecke follows Lewis (1986, ix-x) by considering individuals in a maximally neutral way as space-time regions, it is clear that a mereological relationship with both a temporal and spatial overlap is established. As mentioned above, Couch emphasizes the simultaneity of components and constituted capacities (Couch 2011, 385).

A further important similarity between the approaches lies in the fact that both authors integrate a multiple realization or 'multiple constitution' of cognitive capacities explicitly as a possibility in their regularity-based definitions of mechanistic constitution. Harbecke introduces the idea into his definition with condition (ii). Couch points out explicitly that not every neural mechanism that has proven to be minimally sufficient for a given cognitive ability should be regarded as necessary for this ability. He accepts that "(...) the mechanisms that realize a capacity are sufficient (in the sense of realization) for the presence of a capacity." But in his view, "it is merely the components of the mechanisms that are necessary in the circumstances for the capacity to occur." (Couch, 2011, 385) On the basis of these points, a further commonality results in the sense that the two authors do not consider mechanistic constitution as reductive a priori. Nevertheless, both provide an empirical criterion for identity: A cognitive capacity is identical to the mechanism realizing it when constitution is mutual.

## 6. Differences

As it became clear in the previous Section, the theories of Harbecke and Couch strongly converge in their core views. Nevertheless, some important differences can be identified, which will be explained in this section.

### *6.1 Type vs. Token*

As explained above, Couch interprets mechanistic constitution primarily as a relationship between individual mechanisms, i.e. between individual events or tokens (see Couch 2011, 384). According to Harbecke, constitution should primarily be understood as a second order relation, i.e. as a relation between mechanistic properties.

This difference is somewhat surprising given the fact that the concept of an INUS condition only makes sense with respect to types. The problem is that each actual event proves redundant for any further actual event. In other words, if 'A' and 'B' refer to actual events (i.e., if they are true), then the conditional 'A $\rightarrow$ B' is immediately true, but so is the conditional '¬A $\rightarrow$ B'. Hence, to avoid a certain kind of trivialization, mechanistic constitution must be understood as a second-order relation.

### *6.2 Symmetry and Reflexivity*

In his article, Couch does not explicitly deal with the question whether the constitution relation as defined is symmetrical, asymmetrical or anti-symmetrical. Furthermore, he does not specify whether constitution is reflexive. However, in personal correspondence he has stated that he agrees with Craver in considering constitutive relevance to be symmetric but not reflexive.

With respect to the issue of symmetry, this may actually be a problem. Unlike Craver, Couch provides an extensional definition of mechanistic constitution. If symmetry holds for such a relation, a conditional of the form ABC $\rightarrow$ F always implies a conditional F $\rightarrow$ ABC. The latter conditional, however, is incompatible with a multiple constitution of F, which Couch wished to allow explicitly (see Section 5 above and Couch 2011, 385). Consequently, the characterization of mechanistic constitution as a symmetric relation is somewhat at tension with other assumptions that are present in his article.

With respect to the question of reflexivity, it is clear that Harbecke's definition of a reflexive and anti-symmetric relation of constitution offers a criterion for the ontological distinction, or the reduction, of cognitive and neural processes. With the rejection of reflexivity, Couch, however, seems to have already introduced an ontological distinction. The point is that now no cognitive capacity can be identical to the neural mechanisms constituting it. Otherwise constitution should at least sometimes be reflexive.

This predetermining commitment may be a problem for the analysis of actual constitutional explanations in neurobiology, as it has been elaborated, for instance, by Craver (2007, 165-170). At an early stage of research in neuroscience, it typically remains an open question whether the phenomena under investigation can ultimately be reduced to mechanisms of lower levels or not. A theory of mechanistic constitution should therefore not exclude identity a priori.

### *6.3 Spurious Regularities*

A further difference between the two theories lies in the fact that Harbecke's definition, in contrast to Couch's, introduces a condition that excludes certain spurious regularities between mechanisms and constituted phenomena. These regularities are imagined analogous to "Factory Manchester Hooters" cases that pose a serious problem for certain regularity theories of causation (cf. Mackie 1974, 83-87). As it can be easily checked, the causal structure represented by Figure 1 is a model of the following minimal theory: A¬CF $\rightarrow$ B. However, according to the causal structure depicted, A should not be a cause of B. This problem for regularity theories of causation was only solved by Grasshoff & May (2001).

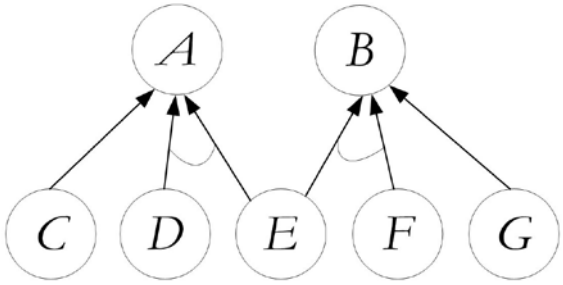



*Figure 1: "Manchester Factory Hooters" (conditions D and E are causally sufficient for A, and C is the only alternative cause of A; conditions E and F are causally sufficient for, and G is the only alternative cause of B).*

At this point it is not clear whether there are actual cases of mechanisms that stand to a cognitive phenomenon in the relation of constitution and that make true a spurious regularity in analogy to the Manchester Hooters case. However, it remains conceivable that such a spurious regularity will be found at some point. Couch can avoid this problem generally by his supposition that mechanistic constitution is symmetrical. However, since the symmetry of mechanistic constitution is problematic (see Section 6.2), a corresponding condition is required.

### *6.4 Individuals and Properties*

Couch considers mechanistic events as the relata of the constitution relation (cf. Section 4), where mechanistic events should be seen as instantiations of mechanistic properties by individuals. The author leaves open, however, how the individuals and properties figuring in these mechanistic events are related to each other. This is a relevant question, not the least because there is a strong intuition, according to which a mechanism constitutes a phenomenon *because* of the properties and individuals occurring in it. Perhaps the relationships between these entities are determined derivatively by the constitution relation; but also perhaps they are wholly new primitive relations.

Harbecke presupposes the mereological relation as primitive, whilst he defines the constitution relation in an extensional way. Actual constitution between two events or tokens is therefore defined derivatively: Two mechanistic events stand in actual constitution to another if the individuals are mereologically related and the mechanistic properties occurring in them are related by constitution.

## 7. Evaluation of the Theories

On the basis of the differences between the approaches highlighted in the previous section, this section attempts a comparative evaluation of the two regularity theories of mechanistic constitution. First, it should be emphasized that both Harbecke's and Couch's theories are richer and more detailed in several respects than the existing manipulationist theories of mechanistic constitution. The regularity theories do not encounter the conceptual difficulties that have been diagnosed by several authors for the manipulationist theories (see Harbecke 2010, 271-73; Couch 2011, Secs. 3-4; Leuridan 2011).

As explained in Section 6.1., a regularity theory must refer to properties or types of mechanistic constitution and not to objects or events if it wants to avoid a certain kind of triviality trap. Furthermore, scientific research has little interest in particular objects or singular events, but always directs its interest to the regular occurrence of compounds of event- or property types. According to Harbecke, mechanistic constitution is primarily a second-order relation that connects mechanistic types. Couch, in contrast, defines mechanistic constitution as holding between events. Accordingly, Harbecke's approach proves more adequate in this regard.

Furthermore, as it was explained in Section 6.2, there is a certain tension between Couch's adoption of a non-reflexive symmetric relation of constitution and the possibility of multiple constitution. Due to the symmetry, to-be-explained phenomena must always be coextensional with a mechanism constituting them. At the same time the non-reflexivity excludes from the outset the possibility of a reduction of to-be-explained phenomena and the mechanisms constituting them. Taken together, these assumptions are at odds with the widely held view that a coextensionality of properties is sufficient for identity of properties. Harbecke defines mechanistic constitution as anti-symmetric and reflexive. He thereby avoids this problem.

The specific problem of spurious regularities in a regularity of mechanistic constitution was presented in Section 6.3. Couch was able to circumvent this problem by introducing the

requirement of symmetry of the constitution relation. However, since the symmetry of the mechanistic structure is problematic (see Section 6.2), a pertinent additional constraint on constitutive regularities is required. Harbecke excluded the occurrence of spurious regularities with condition (ii) of his definition. This solution is analogous to certain proposals that have been put forward for regularity theories of causality.

Finally, it was shown that Couch leaves somewhat vague the specific relation that mechanisms, properties and individuals have to another. This is not necessarily a drawback of the theory, but requires clarification at some point. Harbecke had to postulate the mereological part-whole relation between individuals as primitive. However, on the basis of this step, he was able to establish a systematic relationship between to-be-explained phenomena, the mechanisms constituting them, as well as the properties and individuals figuring in these.

With these differences a certain comparative advantage of Harbecke's approach suggests itself with respect to Couch's theory, even if both theories largely agree in their basic positions. Since regularity theories were presented as successfully competing with manipulationist theories of mechanistic constitution, Harbecke's definition proves to be the currently most adequate approach for the reconstruction of mechanistic explanations in neurobiology.

## 8. Conclusion

This paper investigated the relation between the two regularity theories of mechanistic constitution by Harbecke (2010) and Couch (2011). After some introductory remarks on the position of the theories within the broader mechanistic approach, the definitions were compared for their similarities and differences. As a final conclusion, it was argued that Harbecke's theory has a comparative advantage over Couch's regardless of the extensive overlap in views between the two approaches.

At this point, a detailed analysis of the implications of these results for the manipulationist theory of constitution has not been developed. Moreover, an adequacy test of for the suggestions discussed in this paper with respect to actual neurobiological theories and explanations, e.g. regarding the possibility of spurious regularities (see Section 6.3), is still pending. These questions and issues should be considered in future research on regularity theories of mechanistic constitution.

**Jens Harbecke**

Philosophy of Science
Witten/Herdecke University
jens.harbecke@uni-wh.de

## References

Bliss, T. and T. Lømo 1973. 'Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path', Journal of Physiology 232, 331–356.

Bliss, T. and G. Collingridge 1993: 'A synaptic model of memory: long-term potentiation in the hippocampus', Nature 361(6407), 31–39.

Couch, M. 2011: 'Mechanisms and constitutive relevance', Synthese 183(3), 375-388.

Craver, C. 2007: Explaining the brain. New York: Oxford University Press.

Davis, S., S. Butcher, and R. Morris 1992: 'The nmda receptor antagonist d-2-amino-5-phosphonopentanoate (d-ap5) impairs spatial learning and ltp in vivo at intracerebral concentrations comparable to those that block ltp in vitro', Journal of Neuroscience 12(1), 21–34.

Fazekas, P. and G. Kertész 2011: 'Causation at different levels: tracking the commitments of mechanistic explanations', Biology and Philosophy 26(3), 1-19.

Graßhoff. G. and M. May 2001: 'Causal regularities, in Spohn, W., M. Ledwig and M. Esfeld (eds.) 2001: Current issues in causation. Paderborn: Mentis, 85–114.

Harbecke, J. 2010: 'Mechanistic Constitution in Neurobiological Explanations', International Studies in the Philosophy of Science 24, 267-285.

Kim, J. 1984: 'Concepts of Supervenience', Philosophy and Phenomenological Research 45, 153-176.

Leuridan, B 2011: 'Three Problems for the Mutual Manipulability Account of Constitutive Relevance in Mechanisms', British Journal for the Philosophy of Science, 63(2), 399-427.

Lewis, D. 1986: Philosophical papers, vol. 2. Oxford: Oxford University Press.

Lømo, T. 2003: 'The discovery of long-term potentiation', Philosophical Transactions of the Royal Society of London B: Biological Science 358, 617–620.

Machamer, P., L. Darden and C. Craver 2000: 'Thinking about mechanisms', Philosophy of Science 67(1), 1–25.

Mackie, J. 1974: The Cement of the Universe. Oxford: Clarendon Press.

Malenka, R., J. Kauer, D. Perkel, M. Mauk, P. Kelly, R. Nicoll, and M. Waxham 1989: 'An essential role for postsynaptic calmodulin and protein kinase activity in long-term potentiation', Nature 340(6234), 554–557.

Mandik, P. 2011: 'Supervenience and neuroscience', Synthese 180(3), 443–463.

R. Morris, E. Anderson, G. Lynch and M. Baudry 1986: 'Selective impairment of learning and blockade of LTP by an NMDA receptor antagonist AP5', Nature 319, 774–776.

R. Morris 1984: 'Developments of a water-maze procedure for studying spatial learning in the rat', Journal of Neuroscience Methods 11, 47–60.

Soom, P. 2011: From Psychology to Neuroscience. A New Reductive Account. Frankfurt: Ontos.

Urban, N. und G. Barrionuevo 1996: 'Induction of hebbian and non-hebbian mossy fiber long-term potentiation by distinct patterns of high-frequency stimulation', Journal of Neuroscience 16, 4293–4299.

# Vage natürliche Arten

Rico Hauswald

Gängige Theorien natürlicher Arten divergieren grundlegend hinsichtlich der Frage, ob bestimmte Vagheitsformen mit dem Begriff einer natürlichen Art kompatibel sind. Während Essentialisten auf diskreten natürlichen Arten bestehen, räumen Vertreter der HPC-Theorie bereitwillig die Möglichkeit vager natürlicher Arten ein. Die genaue theoretische Klärung, was Vagheit in diesem Zusammenhang eigentlich heißt, stellt allerdings bislang ein Desiderat dar. Ich will mich in meinem Beitrag dieses Desiderats annehmen und eine Bestimmung und Modellierung der Vagheit von HPC-Arten vornehmen, wobei Einsichten aus der vagheitstheoretischen Diskussion fruchtbar gemacht werden sollen. Ich argumentiere (1.) dass die HPC-Theorie als kausal-realistischer, d.h. nicht-semantischer Clusteransatz mit dem semantischen Externalismus kompatibel ist und dass die die HPC-Arten bezeichnenden Terme direkt auf diese referieren. Daraus folgt, (2.) dass es sich bei der von diesen Termen aufgewiesenen Vagheit nicht um intensionale Vagheit handeln kann. (3.) argumentiere ich, dass HPC-Arten erfolgreich mithilfe des vagheitstheoretischen Begriffs der Realisierungslücke modelliert werden können. Auf dieser Grundlage kann dann (4.) die Parallele zum sogenannten *problem of the many* verständlich gemacht werden. Ein wichtiges Ergebnis besteht in der genauen Differenzierung und Rekonstruktion unterschiedlicher Vagheitstypen, die für den HPC-Ansatz einschlägig sind. Dazu gehören synchrone und diachrone Unbestimmtheit von HPC-Arten, „imperfect homeostasis“ und die Gradualität des Natürliche-Arten-Konzepts selbst.

## 1. Einleitung

Gängige Theorien natürlicher Arten divergieren grundlegend hinsichtlich der Frage, ob es so etwas wie Vagheit, Unbestimmtheit oder unscharfe Kategoriengrenzen bei natürlichen Arten gibt. Vertreter der beiden wichtigsten Theorieansätze, d.h. des (Mikro-)Essentialismus und der *homeostatic property cluster*-(HPC-)Konzeption, nehmen diesbezüglich entgegengesetzte Standpunkte ein. Während Essentialisten auf diskreten Kategoriengrenzen bestehen, betont Richard Boyd: „It is a feature of [...] homeostatic property cluster (HPC) kinds [...] that there is always some indeterminacy or ‚vagueness‘ in their extension.” (Boyd 1999: 141) Die genaue theoretische Klärung, was „Vagheit“ in diesem Zusammenhang eigentlich heißt, stellt allerdings bislang ein Desiderat dar. Bezeichnenderweise setzt Boyd das Wort im zitierten Text wie auch an anderen Stellen in Anführungszeichen. Ich will mich in meinem Beitrag dieses Desiderats annehmen und eine genaue Bestimmung und systematische Modellierung der Vagheit von HPC-Arten vornehmen, wobei Einsichten und Unterscheidungen aus der vagheitstheoretischen Diskussion fruchtbar gemacht werden sollen. Sowohl HPC- als auch Vagheitstheorie können, wie ich meine, von diesem bisher kaum[1] geführten Dialog profitieren.

Zunächst werde ich kurz die divergierenden Standpunkte, die Vertreter der verschiedenen Theorien natürlicher Arten zum Vagheitsproblem einnehmen, darstellen. Dann führe ich einige für die Diskussion relevante vagheitstheoretische Begriffe und Unterscheidungen ein. Insbesondere greife ich den Begriff der Realisierungslücke auf und diskutiere den Gegensatz zwischen intensionaler und extensionaler Vagheit. Gegen Boyds eigene Vorbehalte werde ich

[1] LaPorte (2004) macht im Zusammenhang mit seinen Überlegungen zu natürlichen Arten zwar Anleihen bei der Vagheitstheorie, seine Stoßrichtung ist aber eine gänzlich andere als die hier angestrebte und er befasst sich auch kaum mit dem HPC-Ansatz.

dafür plädieren, den HPC-Ansatz mit dem semantischen Externalismus zu verbinden. Daraus folgt, dass Natürliche-Art-Begriffe nicht intensional vage sein können, da sie gar keine Intension haben, sondern direkt referieren. Schließlich präsentiere ich einen Vorschlag zur Modellierung von verschiedenen Vagheitsphänomenen bei HPC-Arten und differenziere zwischen (a) synchroner und (b) diachroner Unbestimmtheit, (c) „imperfect homeostasis“ und (d) der Gradualität des Natürliche-Arten-Konzepts selbst.

## 2. Vagheit bei natürlichen Arten in essentialistischen Theorien und im HPC-Ansatz

Vertreter essentialistischer Konzeptionen natürlicher Arten vertreten in der Regel die Auffassung, Vagheit sei mit dem Begriff der natürlichen Art nicht vereinbar. Bereits bei Aristoteles heißt es:

> Die Substanz scheint kein Mehr oder Weniger zuzulassen. [...] Zum Beispiel, wenn diese Substanz ein Mensch ist, so wird er nicht mehr oder weniger ein Mensch sein, weder als er selbst noch als ein anderer Mensch. Denn ein Mensch ist nicht mehr ein Mensch als ein anderer, so wie ein weißes Ding weißer ist als ein anderes und ein schönes Ding schöner als ein anderes. (Aristoteles 2006: 14f.)

Moderne Essentialisten sehen es ähnlich. Ellis (2001: 19f.) formuliert sechs Bedingungen, die eine Art seiner Meinung nach erfüllen muss, um als natürliche Art gelten zu können. Die zweite dieser Bedingungen enthält die Forderung, dass natürliche Arten kategorial distinkt sein müssen:

> [N]atural kinds must be *categorically distinct* from each other. For they must be ontologically grounded *as kinds*, and exist *as kinds* independently of our conventions. Hence, where we are dealing with natural kinds, there cannot be any gradual merging of one kind into another, so that it becomes indeterminate to which kind a thing belongs. For if there were any such merging, we should have to draw a line somewhere if we wished to make a distinction. But if *we* have to draw a line anywhere, then it becomes *our* distinction, not nature's. Natural kinds must be *ontologically* distinguishable from each other. (Ellis 2001: 19f.) (Hervorh. i. O.)

Demgegenüber betont Boyd: „It is a feature of [...] *homeostatic property cluster (HPC) kinds* [...] that there is always some indeterminacy or ‚vagueness' in their extension.“ (Boyd 1999: 141) Diese Unbestimmtheit führt dazu,

> that neither theoretical nor methodological considerations assign the object being classified determinately to the kind or to its complement, with the result that the homeostatic property-cluster definition fails to specify necessary and sufficient conditions for kind membership. (Boyd 1991: 142)

Wenn diese Unbestimmtheit tatsächlich weder durch theoretische noch durch methodologische (oder empirische) Erwägungen beseitigt werden kann, ist das ein klares Indiz für das Vorliegen eines *echten* Vagheitsphänomens. Denn echte Vagheit ist von bloßer epistemischer Unbestimmtheit zu unterscheiden. Für paradigmatische Fälle von Vagheit ist charakteristisch, dass die Unbestimmtheit prinzipiell nicht durch die Generierung zusätzlichen Wissens zu beseitigen ist. Wenn jemand ein Grenzfall des Prädikats „glatzköpfig“ darstellt, dann kann die dabei vorliegende Unbestimmtheit auch dadurch nicht beseitigt werden, dass die exakte Anzahl seiner Haare bestimmt wird.[2]

[2] Vgl. z.B. die bekannte Begriffsbestimmung von Grice: „To say that an expression is vague [is] to say that there are cases (actual and possible) in which one just does not know whether to apply the

Boyd sieht die mit HPC-Arten einhergehende Unbestimmtheit nicht als Defizit an; im Gegenteil:

> Both the property cluster form of such definitions and the associated indeterminacy are dictated by the fundamental epistemic task of employing categories which correspond to inductively and explanatorily relevant causal structures. (Boyd 1991: 142)

Insbesondere könne diese Unbestimmheit „not be remedied without rendering the definitions unnatural in the sense of being scientifically misleading." (ebd.) Gerade angesichts des „messy state of affairs" (Reydon 2009: 728) in den *special sciences*[3] tritt die HPC-Theorie ja mit dem Anspruch in Erscheinung, ein angemesseneres Modell bereitzustellen.

Klärungsbedürftig ist an Boyds Aussagen allerdings, in welchem Sinne hier eigentlich genau von „Vagheit" die Rede sein soll, welche unterschiedlichen Varianten eventuell differenziert werden müssen, wie genau diese funktionieren und auf welche Weise sie den HPC-Ansatz gegenüber anderen Natürliche-Arten-Theorien auszeichnen.

Bei der Bestimmung der hier vorliegenden Formen von Vagheit sind nun zunächst einige vagheitstheoretische Grundunterscheidungen und Begriffe zu beachten.

## 3. Extensionale und intensionale Vagheit und der Begriff der Realisierungslücke

Eine grundlegende vagheitstheoretische Unterscheidung wird in der Regel zwischen intensionaler und extensionaler Vagheit getroffen – ein Gegensatz, den Kit Fine in Anlehnung an Friedrich Waismanns Unterscheidung zwischen Vagheit im engeren Sinn und *open texture* (Porösität) eingeführt hat (Waismann 1945; Fine 1975). Intensional vage ist ein Begriff demnach im Falle einer „deficiency of intension", extensional vage im Falle einer „deficiency of extension" (Fine 1975: 266). Bei extensional vagen Begriffen treten Grenzfälle („borderline cases") auf, also Objekte, die weder eindeutig zur Extension des Wortes gehören, noch eindeutig nicht dazu gehören. Ist die Intension eines Wortes von Vagheit gekennzeichnet, resultiert dies in so etwas wie „potentieller Vagheit". Waismann illustriert es am Beispiel des Wortes „Katze": Wenn wir ein katzenartiges Wesen anträfen, das, wie wir bemerken, sehr untypische Eigenschaften aufweist, z.B. zu gigantischer Größe anwachsen kann oder wieder zu leben beginnt, nachdem es eigentlich schon tot zu sein schien, so wären wir unsicher, ob wir dieses Wesen „Katze" nennen sollten – wir hätten diesbezüglich weder klar positive noch klar negative Intuitionen.

Nun gibt es solche sonderbaren Wesen nicht wirklich. Es mag daher gut sein, dass keine tatsächlichen Grenzfälle zum Prädikat „ist eine Katze" existieren und dieses damit nicht extensional vage ist. Nichtsdestoweniger ist es, wie Waismann sagen würde, „porös", d.h. seine Anwendbarkeit ist nicht für jede ungewöhnliche, aber logisch mögliche Situation geregelt. Den Aspekt der extensionalen Präzision haben Vagheitstheoretiker mit dem Begriff der „Realisierungslücke" zu beschreiben versucht. Realisierungslücken gibt es demnach zwischen Katzen und nicht-katzenförmigen Objekten (z.B. Exemplaren anderer biologischer Spezies wie Hunden) (Pinkal 1985: 88). Ein Begriff wie „Katze" ist, wie Pinkal schreibt, „zwar extensional präzise [...], aber intensional vage" (ebd.).

Wenn man nun einen Speziesbegriff wie „Katze" als Begriff auffasst, der auf eine HPC-Art Bezug nimmt, erweisen sich allerdings zwei Dinge an dieser Beschreibung als problematisch. Zum einen können HPC-Arten sowohl in synchroner als auch in diachroner Hinsicht

---

expression or to withhold it, and one's not knowing is not due to ignorance of the facts." (Grice 1989: 177)

[3] Boyd (1999: 151) spricht von „inexact, messy, and parochial sciences".

Unbestimmtheit aufweisen, so dass die Rede von „extensionaler Präzision“ zumindest klärungsbedürftig ist und eingeschränkt werden muss. Man denke im Zusammenhang mit biologischen Spezies etwa an Hybride oder an evolutionäre Veränderungen. (Der Begriff der Realisierungslücke erweist sich gleichwohl als fruchtbar und ich werde ihn weiter unten zur formalen Modellierung von HPC-Arten heranziehen.) Der andere problematische Aspekt an der Redeweise davon, dass ein Speziesbegriff wie „Katze“ extensional präzise, aber intensional vage sei, besteht darin, dass ein solcher Begriff als Natürliche-Art-Terminus eigentlich *überhaupt* keine Intension hat, sondern namensartig direkt auf die Art referiert. Wie ich im nächsten Abschnitt zeigen möchte, sollte auch ein HPC-Theoretiker von einem solchen direkten Referieren Natürlicher-Art-Begriffe ausgehen. Waismann, der die Unterscheidung und das Katzenbeispiel ins Spiel gebracht hat, argumentierte demgegenüber noch auf der Grundlage eines Wittgensteinianischen semantischen Clustermodells, dem zufolge auch „Katze“ eine Intension *hat* – wenn auch nicht in Form einer Definition, die notwendige und hinreichende Bedingungen angibt, sondern in Form eines semantischen Clusters. In Bezug auf Natürliche-Art-Begriffe sollte dieses semantische Modell aber zurückgewiesen und eine – allerdings modifizierte, eingeschränkte – Version eines direkten Referenzmodells vertreten werden, der zufolge Natürliche-Art-Begriffe direkt auf die in der Welt bestehenden HPC-Strukturen referieren.

## 4. Der HPC-Ansatz und der semantische Externalismus

Die HPC-Theorie stellt eine Cluster-Konzeption natürlicher Arten dar. Sie postuliert keine für die Artzugehörigkeit essentiellen Merkmale, wie es der Essentialismus tut. Stattdessen ist es für die Artzugehörigkeit hinreichend, dass ein Objekt hinreichend viele der typischen Art-Eigenschaften aufweist. Die HPC-Theorie ist aber nicht mit *semantischen* Cluster-Theorien zu verwechseln, wie sie etwa in Gestalt der Wittgensteinianischen Familienähnlichkeit vorliegen. Die HP-Cluster sind reale Strukturen, die aufgrund kausaler Regelmäßigkeiten zustande kommen, und keine bloßen semantischen Artefakte. Die Grundidee ist, dass das gemeinsame Auftreten der Eigenschaften das Resultat eines selbstorganisierenden Prozesses ist, der darin besteht, dass einige der jeweils für eine Art typischen Eigenschaften das Auftreten der anderen Eigenschaften begünstigen oder dass es zugrundeliegende Mechanismen gibt, die das gemeinsame Auftreten der Eigenschaften begünstigen.[4]

Daraus folgt, dass die HPC-Theorie nicht von jenen Argumenten bedroht ist, die Putnam und Kripke gegen Cluster-Theorien natürlicher Arten vorgebracht haben, da diese Argumente *semantische* Clustermodelle treffen, nicht kausale. Semantische Clusterkonzeptionen versuchen eine Antwort darauf zu geben, was die Intension eines Wortes wie z.B. „Zitrone“ oder „Gold“ ist. Sie treten damit in Konkurrenz zu alternativen semantischen Konzeptionen, insbesondere Konzeptionen, die entweder von einer analytischen Definition im Sinne klassischer Intensionen ausgehen (also etwa: „Gold ist definiert als gelbes Metall, das diese und jene Merkmale hat“ o.ä.), oder von einer direkten Referenz des – namensartigen – Wortes auf die Art selbst. Kripke und Putnam haben überzeugend für die Überlegenheit dieser letzteren, externalistischen Auffassung gegenüber sowohl der analytisch-definitorischen als auch der Cluster-Konzeption argumentiert, zumindest mit Blick auf Natürliche-Art-Terme und Eigennamen (es mag durchaus sein, dass bei anderen Begriffstypen andere semantische Konzeptionen angemessen sind und die analytisch-

[4] Boyd formuliert es so: „I argue that there are a number of scientifically important kinds (properties, relations, etc.) whose natural definitions are very much like the property-cluster definitions postulated by ordinary-language philosophers except that the unity of the properties in the defining cluster is mainly causal rather than conceptual.“ (Boyd 1991: 141) Der Begriff einer „natürlichen *Definition*“ ist allerdings aus Gründen, die in diesem Abschnitt noch deutlicher werden, irreführend (für eine Kritik an diesem Begriff vgl. auch Millikan 1999: 99).

definitorische Theorie etwa auf *one-criterion-words* zutrifft, wie Putnam (1975a: 139) vermutet, und die Cluster-Theorie auf ein Wort wie „Spiel“, wie Wittgenstein (1984a: 277) meinte). Ein semantischer Cluster-Ansatz läuft darauf hinaus, dass die Bedeutung eines Wortes in einer Liste intensionaler Merkmale besteht – hierin ähnelt er dem klassischen Definitionsmodell –, die aber keine Menge notwendiger und zusammen hinreichender Bedingungen dafür darstellen, dass etwas von dem Wort bezeichnet wird. Ausreichend ist vielmehr, dass eine hinreichend große Zahl der prinzipiell in Frage kommenden Merkmale vorliegt. Das Argument Kripkes und Putnams gegen diese Auffassung besteht darin, dass es plausibel zu sein scheint, dass wir uns in Bezug auf alle vermeintlich typischen Eigenschaften einer Instanz einer natürlichen Art irren können, diese Eigenschaften also keine intensionale Rolle für das Wort spielen können, und zwar weder in der strengen definitorischen noch der clusterförmig „aufgelockerten“ Variante. Stattdessen muss man es sich so vorstellen, dass das Wort direkt die Art bezeichnet, so wie ein Eigenname direkt die Person oder das Objekt bezeichnet.

Wie gesagt, halte ich diese Kritik an der semantischen Cluster-Auffassung für genauso überzeugend wie den externalistischen Gegenvorschlag. Da nun die HPC-Theorie keine semantische Cluster-Konzeption darstellt, sondern eine kausale, ist sie nicht von dieser Argumentation bedroht. Sie stellt keinen Rückfall in eine Auffassung dar, die längst als unhaltbar erwiesen wurde. Mehr noch: Sie ist durchaus mit dem semantischen Externalismus kompatibel. Der Externalismus als semantische Theorie ist nicht auf einen metaphysischen (Mikro-)Essentialismus angewiesen. Die einfache Grundidee, die im Folgenden zu explizieren sein wird, besteht darin, dass der Natürliche-Art-Begriff direkt auf die Art referiert, diese Art aber eben nicht als durch essentielle Merkmale charakterisiert betrachtet wird, sondern als kausales Eigenschaftscluster.

Boyd selbst hat starke Vorbehalte hinsichtlich der Anwendung des semantischen Externalismus auf den HPC-Ansatz. Er bezieht sich auf das kausale Referenzmodell natürlicher Arten, dem zufolge Referenz eine rein naturalistische Relation zwischen außerlinguistischen Entitäten (nämlich den natürlichen Arten) einerseits und linguistischen Entitäten (den die Arten bezeichnenden Termen) andererseits ist, die vollständig kausal etabliert wird. Da dieses Bild letztlich problematisch und unhaltbar sei, komme die damit verbundene externalistische Konzeption nicht als geeignete semantische Theorie für den HPC-Ansatz in Frage (Boyd 2010: 222ff.).

Boyds Vorbehalte sind aber – so möchte ich argumentieren – nicht gut begründet, da sie auf einer übersimplifizierten Darstellung der Theorie der direkten Referenz von Termen auf natürliche Arten beruhen. Die Vorstellung einer *vollkommen* intensions- und beschreibungsfreien Etablierung der Referenzrelation ist in der Tat illusorisch und kann für externalistische Semantiken natürlicher Arten (ob nun in essentialistischen oder nicht-essentialistischen Versionen) nicht sinnvoll behauptet werden. Allerdings muss – und sollte – sich der semantische Externalist auch ohnehin nicht auf eine solche Extremposition festlegen, und es entspricht einem verbreiteten Missverständnis, dass Kripke und Putnam eine solche Extremposition propagiert hätten. Das soll im Folgenden kurz deutlich gemacht werden.

Die Auffassung, dass Namen oder Natürliche-Art-Begriffe *vollkommen* intensionsfrei, *rein* denotativ referieren könnten, ist in der Tat unvollständig; die direkte Referenz ist sozusagen nur die eine Hälfte der zu erzählenden Geschichte. Bevor eine Taufe erfolgreich sein kann und eine Referenzrelation zwischen Zeichen und Bezeichnetem hergestellt werden kann, muss zunächst erst einmal ein minimales Maß an Klarheit darüber geschaffen sein, worauf man sich bezieht und was eigentlich getauft werden soll. Es muss so etwas wie eine primäre Bezugnahme auf die zu taufende Entität geleistet werden. Diese Entität muss bereits auf die ein oder andere Art und Weise individuiert worden sein, damit ein Taufakt überhaupt erfolgreich sein kann, und für diese Individuierung spielt die Verwendung sortaler Ausdrücke

eine entscheidende Rolle.[5] Wir sind hier mit dem klassischen Problem der hinweisenden oder ostensiven Definition konfrontiert, das Wittgenstein ausführlich behandelt hat (vgl. z.B. Wittgenstein 1984a: 254). Ein deklarativer Sprechakt wie „Dies soll N heißen" allein reicht noch nicht hin, um die Beziehung zwischen Namen und Namensträger erfolgreich zu etablieren, da der indexikalische Ausdruck „dies" in dem Sinn semantisch völlig unbestimmt ist, dass damit noch keine konkrete, irgendwie näher bestimmte Entität herausgegriffen wird. Dies ist erst dann möglich, wenn der Ausdruck „dies" im Kontext einer konkreten Konversation geäußert wird, bei der unter den Teilnehmern in Bezug auf eine hinreichend große Anzahl von Sachverhalten Einigkeit im Sinne eines gemeinsamen Wissens (*common knowledge*) herrscht. Zu diesen Sachverhalten gehört z.B., dass ein hinreichend ähnlicher begrifflicher Rahmen verwendet wird, so dass die Wahrscheinlichkeit hoch ist, dass die Teilnehmer die Gegenstände in ihrer Umgebung auf gleiche oder ähnliche Weise individuieren. Erst wenn diese Einigkeit gewährleistet ist, können die Sprecher sich mit „dies" (unterstützt durch Zeigegesten oder ähnliches) auf ein- und dasselbe Objekt beziehen. Fehlt die Einigkeit, bliebe unbestimmt, ob mit „Dies soll N heißen" nun ein Mensch, oder z.B. nur der Teil eines Menschen (vielleicht auch ein zeitlicher Teil, eine Phase), oder das mereologische Aggregat der Atome, die den Menschen in diesem Moment konstituieren, oder auch einfach ein Punkt oder Abschnitt im Raumzeitkontinuum benannt werden soll. Eine solche Klärung setzt Beschreibungen voraus, die nicht unbedingt explizit vorgenommen werden müssen, sondern auch als selbstverständlich vorausgesetzt werden können. Bei der Taufe eines Menschen ist in der Regel klar, dass keine Raum- oder Zeitpunkte, sondern eben Menschen bezeichnet werden sollen (die Formulierung „Ich taufe *dich* N" ist auch insofern spezifischer, als die zu taufende Entität als *Person* herausgegriffen wird).

Gegen diese behauptete deskriptive, intensionale oder sortale Vermittlung bei einer Einführung direkt referierender Ausdrücke könnte man versuchen, (gedankenexperimentelle) Gegenbeispiele anzuführen. Was zum Beispiel, wenn sich der als N getaufte vermeintliche Mensch plötzlich als Roboter oder Zombie herausstellt? Hört dann der Name N auf, zu referieren? Das scheint kontraintuitiv zu sein; viel eher läge die Reaktion nahe zu sagen, „N ist eigentlich kein Mensch, sondern ein Roboter/Zombie", oder „N hat sich als Roboter/Zombie herausgestellt". Ich denke allerdings, dass das nur deswegen so ist, weil Menschen, Roboter und Zombies hinreichend ähnliche Entitäten sind. Was aber, wenn (um mit Wittgenstein zu sprechen) „etwas wirklich Unerhörtes geschähe"[6], wenn das Objekt, von dem angenommen wurde, dass es mit N bezeichnet wurde, sich *völlig* anders verhalten würde als ein Mensch, wenn es vielleicht verschwinden und wieder auftauchen, oder sich in mehrere „Menschen" aufspalten würde (welcher davon ist N?), die dann wieder verschmelzen, wäre es dann nicht fraglich, ob überhaupt jemals eine erfolgreiche Verbindung zwischen dem Namen N und *irgendeinem* Objekt hergestellt worden ist, ob bei der Taufe überhaupt ein Objekt vom Typ X (oder eines hinreichend ähnlichen Typs) anwesend war (letztlich könnte auch eine (kollektive) Illusion vorgelegen haben, dass ein Objekt, das man taufen könnte, anwesend ist)?

Was gerade vorrangig in Bezug auf Eigennamen (bzw. von solchen bezeichnete Einzeldinge) erläutert wurde, kann auch auf die Einführung eines Natürliche-Arten-Terms und dessen

[5] Vgl. zur sortalen Vermittlung u.a. Geach (1980: 67f.) und Lowe (2009: 29f.). Auch Evans (1982) stellt ähnliche Überlegungen an. Er möchte die direkte Referenztheorie um ein Prinzip ergänzen (er nennt es „Russell's principle"), das besagt, dass eine Bezugnahme auf bzw. ein Urteil über ein bestimmtes Objekt nicht möglich ist, solange nicht gewisse minimale Formen des Wissens über dieses Objekt vorliegen.

[6] „Wie, wenn etwas wirklich Unerhörtes geschähe? wenn ich etwa sähe, wie Häuser sich nach und nach ohne offenbare Ursache in Dampf verwandelten; wenn das Vieh auf der Wiese auf den Köpfen stünde, lachte und verständliche Worte redete; wenn Bäume sich nach und nach in Menschen und Menschen in Bäume verwandelten. Hatte ich nun recht, als ich vor allen diesen Geschehnissen sagte ‚Ich weiß, daß das ein Haus ist' etc., oder einfach ‚Das ist ein Haus' etc.?" (Wittgenstein 1984b: 222; vgl. auch Wittgenstein 1984a: 285).

Beziehung zur bezeichneten Art übertragen werden. Einen wichtigen Aspekt deutet Kripke selbst an:

> Since we have found out that tigers do indeed, as we suspected, form a single kind, then something not of this kind is not a tiger. Of course, we may be mistaken in supposing that there is such a kind. In advance, we suppose that they probably do form a kind. Past experience has shown that usually things like this, living together, looking alike, mating together, do form a kind. (Kripke 1980: 121)

Ausgehend von dieser Stelle möchte ich für die Schlussfolgerung argumentieren, dass man sich auch die Festlegung der Referenz eines Natürliche-Arten-Begriffs notwendigerweise nur intensional, sortal, deskriptiv vermittelt vorstellen kann. Kripke selbst hat in diese Richtung zwar weniger ausführlich argumentiert, Putnam dagegen schon eher (vgl. etwa Putnam 1991); prinzipiell sind sich aber beide des Problems mehr oder weniger bewusst.[7]

Wenn ein Term wie „Tiger" anhand paradigmatischer Exemplare eingeführt werden soll, muss im Voraus klar sein, inwiefern der Term auf Tiger referieren soll. So, wie die Taufe eines Menschen (implizit oder explizit) über das das Sortal „Mensch" vermittelt erfolgt, muss man sich die Taufe einer natürlichen Art über das Sortal „natürliche Art" vermittelt vorstellen. Dieser Punkt ist keineswegs so trivial, wie es vielleicht zunächst scheint; ihn nicht hinreichend zu würdigen, scheint mir viel eher eine häufige Quelle von Missverständnissen. Das Sortal „natürliche Art" legt den Taufenden gewissermaßen ungefähr darauf fest, „welche Rolle das Wort in der Sprache überhaupt spielen soll".[8] Es ist Putnam und Kripke zufolge unstrittig, dass neben Natürliche-Art-Begriffen auch nach wie vor auch andere Typen von Begriffen in unserer Sprache vorkommen, die semantisch anders funktionieren (z.B. intensional-analytisch definierte Begriffe). Die Sprecher brauchen in jedem Fall semantisches Wissen, aufgrund dessen sie verschiedene Typen von Begriffen wenigstens im Prinzip unterscheiden können.

Putnam hat ähnliche Probleme zumindest ab *The Meaning of „Meaning"* selbst recht klar gesehen (vgl. insbes. auch Putnam 1990). Hacking spricht in diesem Zusammenhang im Anschluss an Devitt (1981) vom „qua-problem": „What is ‚the same stuff'? That is the qua question." (Hacking 2007: 9) Gemeint ist die Situation, in der ein Name mit Verweis auf eine Stoffprobe oder ein Exemplar eingeführt wird, indem der Taufende sagt: „Dies und alles, was von derselben Art ist, soll X heißen". Das Problem ist, dass es nicht nur eine Weise gibt, wie etwas mit etwas anderem ähnlich sein kann und demzufolge eine einheitliche Art bildet. Der Begriff der Ähnlichkeit ist viel zu weit, um hinreichend exakte Kriterien bereitzustellen, um bei beliebigen anderen Objekten eindeutig entscheiden zu können, ob sie relativ zur Probe ähnlich sind oder nicht (vgl. dazu die klassische Darstellung bei Goodman 1970). Und selbst wenn man den Art-Begriff enger fasst, so dass nicht alles, was mit irgendetwas ähnlich ist, eine eigene Art bildet, entledigt man sich nicht des Problems, da es z.B. Arten unterschiedlicher Allgemeinheitsgrade geben kann – z.B. instanziiert eine bestimmte Probe Wasser sowohl die Art $H_2O$ als auch die Art *Flüssigkeit* (von der auch Putnam (1975b: 239) sagt, dass es ebenfalls eine natürliche Art ist).[9] Putnam reagiert auf dieses Problem, indem er „Ähnlichkeit" mit einem Index versieht: „same$_L$" was zu lesen ist als „being the same liquid as". Diese Indizierung scheint in der Tat eine recht ähnliche Strategie darzustellen, wie die sortale Relativierung, von der die Rede war. Gemeint ist nicht das, was nur irgendwie in einem unspezifischen Sinn der ursprünglichen Probe ähnlich ist, sondern das, was der Probe in *relevanten Hinsichten* ähnelt. Putnam spricht in diesem Zusammenhang von „Wichtigkeit"

[7] So geht Kripke (1980: 115) etwa kurz auf Geachs Punkte ein.

[8] „Die hinweisende Definition erklärt den Gebrauch – die Bedeutung – des Wortes, wenn es schon klar ist, welche Rolle das Wort in der Sprache überhaupt spielen soll." (Wittgenstein 1984a: 254)

[9] Mit einem solchen Problem ist jeder konfrontiert, der akzeptiert, dass natürliche Arten hierarchisch geordnet sein können – was auch der Essentialist Ellis tut (vgl. Ellis 2001: 20).

(*importance*). Was aber wichtig ist, welche Hinsichten relevant sind, hängt vom Kontext und von unseren Interessen ab. Diesen Punkt betont Putnam viel stärker als Kripke und weist darauf auch selbst explizit hin:

> Are they samples of different substances? Well, it may depend on our interests. (This is the sort of talk Kripke hates!) But the fact that there is some component of interest relativity here, and, perhaps, some drawing of arbitrary lines, does not change the fact that the degree of arbitrariness is infinitesimal compared to the arbitrariness in the ‚almost the same matter at the time of origin' criterion for identity of tables. (Putnam 1990: 68)

Ich hoffe, soweit Folgendes gezeigt zu haben: Der vom semantischen Externalismus bzw. der kausalen Referenztheorie angenommene „Taufakt" funktioniert selbstverständlich nicht rein ostensiv, sondern ist immer über Beschreibungen, sortale Individuierungen und Intentionen vermittelt. Die Notwendigkeit dieser Vermittlung ist von Kripke und Putnam bereits weitgehend selbst gesehen worden, und insofern dies nicht oder nicht ausreichend geschehen ist, lässt sich ihre semantische Theorie leicht um entsprechende Elemente erweitern. Boyds Vorbehalte erweisen sich vor diesem Hintergrund als unbegründet. Die Annahme, ein Begriff referiere semantisch direkt, ist mit der Annahme kompatibel, dass die Einführung dieses Begriffs deskriptiv und sortal vermittelt geschehen ist. Angesichts der Vorteile, die der Externalismus grundsätzlich gegenüber alternativen semantischen Theorien hat, spricht nichts dagegen – vieles aber dafür –, ihn als geeignete Semantik für die HPC-Konzeption zu übernehmen. Ein natürliche-Art-Begriff referiert demnach einfach direkt auf das HP-Cluster – und nicht wie in der mikroessentialistischen Version auf eine durch Mikroessenzen charakterisierte Art. Das bedeutet aber auch, dass der Begriff keine Intension im herkömmlichen Sinn hat (weder eine analytische Definition noch eine Cluster-artige) und demzufolge nicht intensional vage sein kann. Bei der Vagheit, die HPC-Art-Begriffe aufweisen können, kann es sich nicht um *semantische* „combinatory vagueness" handeln. Nichtsdestoweniger ist eine Form von Kombinatorik für die Vagheit bei HPC-Arten einschlägig. Das Cluster, um das es geht, ist kein Bündel von Merkmalen, die als Intension eines Begriffes semantisch zusammengefasst werden, sondern eine von unseren Prädikaten unabhängige reale Struktur in der Welt, die aufgrund kausaler Mechanismen besteht. Eine der Vagheitsformen, die in diesem Zusammenhang auftreten können, nennt Boyd „imperfect homeostasis" (Boyd 1991: 142). Sie besteht darin, dass einige der für die Art typischen Eigenschaften in einigen Exemplaren nicht realisiert sind. Wie genau diese Vagheitsform modelliert werden kann und welche anderen Vagheitsphänomene bei HPC-Arten vorkommen können, soll im folgenden Abschnitt untersucht werden.

## 5. Zur Modellierung verschiedener Vagheitsformen

Im folgenden Diagramm soll das Zustandekommen von Arten durch Eigenschaftsclustering veranschaulicht werden. Ich möchte deutlich machen, dass sich der vagheitstheoretische Begriff der Realisierungslücke generell für eine Bestimmung des Begriffs von HPC-Arten fruchtbar machen lässt. Im Rahmen dieses Modells ist es dann möglich, verschiedene Vagheitsformen zu modellieren. Ich werde im Folgenden argumentieren, dass HPC-Arten ihre Signifikanz dem Vorhandensein zum einen von hinreichend großen kausalen Eigenschaftsclustern, zum anderen dem Vorkommen von Realisierungslücken verdanken. Es wird sich zeigen, dass ein HPC-Art-Begriff durch ein gewisses Maß an extensionaler Präzision charakterisiert ist (wobei dieses Maß sich aus der Größe und Klarheit der Realisierungslücke ergibt) und zugleich verschiedene Formen von Unbestimmtheit möglich sind.

Im folgenden Diagramm sind 13 Individuen ($A_1$ bis $D_3$) in einem 19-dimensionalen Eigenschaftsraum eingetragen. Die Individuen gehören 4 Spezies an (A, B, C, D), die

wiederum zwei Genera zugeordnet sind. Spezies wie Genera stellen HPC-Arten dar. Zur besseren Vorstellbarkeit könnte man bei A und B an zwei verschiedene Säugetierarten, bei C und D an zwei verschiedene Vogelarten denken. Die HPC-Theorie lässt eine Gewichtung von Eigenschaften des Clusters zu;[10] damit der Clustergedanke deutlich wird, sehe ich vereinfachend von dieser Möglichkeit ab, d.h. alle Eigenschaften werden als gleichwertig angesehen.

| | $Genus_1$ (A/B) | | | | | | | | | $Genus_2$ (C/D) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Genus_1$typische Merkmale (A/B) | | | Spezies-A-typische Merkm. | | | Spezies-B-typische Merkm. | | | $Genus_2$typische Merkmale (C/D) | | | Spezies-C-typische Merkm. | | | Spezies-D-typische Merkmale | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 |
| $A_1$ | x | x | | x | x | | | | | | | | | | | | | | |
| $A_2$ | | x | x | | x | x | | | | | | | | | | x | | | |
| $A_3$ | x | | x | x | | x | | | | | | | | | | | | | |
| $B_1$ | x | x | | | | | x | x | | | | | | | | | | | |
| $B_2$ | | x | x | | | | | x | x | | | | | | | | | | |
| $B_3$ | x | | x | | | | x | | x | | | | | | | | | | |
| $B_4$ | x | x | x | | | | x | x | x | | | | | | | | | | |
| $C_1$ | | | | | | | | | | x | x | | x | x | | | | | |
| $C_2$ | | | | | | | | | | | x | x | | x | x | | | | |
| $C_3$ | | | | | | | | | | x | | x | x | | x | | | | |
| $D_1$ | | | | | | | | | | x | x | | | | | x | x | | x |
| $D_2$ | | | | | | | | | | | x | x | | | | | x | x | x |
| $D_3$ | | | | | | | | | | x | | x | | | | x | | x | |

Sowohl A- als auch B-Individuen ($A_1$, $A_2$, $A_3$; $B_1$ $B_2$, $B_3$, $B_4$) gehören zu $Genus_1$ und weisen daher zunächst $Genus_1$-typische Eigenschaften (1, 2, 3) auf (anschaulich: die für alle Säugetiere qua Säugetiere charakteristischen Eigenschaften). Bei wirklichen natürlichen Arten ist die Anzahl gemeinsamer, typischer Eigenschaften freilich enorm, die Cluster damit weitaus „eindrucksvoller“[11] – ich beschränke mich hier aus Darstellungsgründen auf eine überschaubare Anzahl von Eigenschaften.

Keine der $Genus_1$-typischen Eigenschaften wird von allen A- bzw. B-Individuen exemplifiziert. Damit liegt hier das vor, was Boyd (1991: 142) „*imperfect homeostasis*“ nennt. Trotzdem sind 1, 2 und 3 $Genus_1$-typische Eigenschaften, da jedes der $Genus_1$-Individuen hinreichend viele von ihnen (nämlich hier zwei der drei) besitzt.

Neben den Genus-typischen Merkmalen besitzt jedes Individuum zusätzlich Spezies-typische Merkmale, d.h. Eigenschaften, die z.B. das Individuum $A_1$ besitzt, *insofern* es zu Spezies A gehört – nämlich hier die Eigenschaften 4 und 5. Wie auch bei den Genus-typischen Merkmalen gilt wiederum, dass kein Merkmal notwendig auftreten muss; $A_1$ fehlt z.B. Eigenschaft 6, obwohl auch diese Spezies-A-typisch ist.

[10] Wobei man eventuell sogar zulassen kann, dass die relative Wichtigkeit einiger Eigenschaften gegenüber den anderen so groß ist, dass dies in diesen Fällen dem Modell des klassischen Essentialismus entsprechen würde (so auch Birds (2007: 211) Vorschlag). Der HPC-Ansatz könnte dann vielleicht als eine Art Metatheorie fungieren, in deren Rahmen unterschiedliche Typen natürlicher Arten verhandelt werden können, inklusive klassischer essentialistischer Arten, die in einigen Fällen (chemische Elemente?) ja vielleicht in der Tat vorliegen mögen.

[11] Prägnant hat Mill diesen Aspekt auf den Punkt gebracht: „[A] hundred generations have not exhausted the common properties of animals or of plants, of sulphur or of phosphorus; nor do we suppose them to be exhaustible, but proceed to new observations and experiments, in the full confidence of discovering new properties which were by no means implied in those we previously knew.“ (Mill 1973: 122)

Zu beachten ist die Möglichkeit *perfekter Exemplare*: Obwohl keine der Eigenschaften notwendig ist (also von allen entsprechenden Exemplaren exemplifiziert wird), kann es Exemplare geben, die alle typischen Eigenschaften aufweisen. So ist $B_4$ sowohl ein perfektes Exemplar von $Genus_1$ als auch von Spezies B. Der Begriff eines perfekten Exemplars muss von dem eines *klaren Exemplars* unterschieden werden. Ein klares Exemplar einer Art ist ein Exemplar, für das unstrittig ist, dass es eine Instanz der fraglichen Art (und nicht etwa einen Grenzfall) darstellt. Dazu muss es nicht alle der für die Art typischen Eigenschaften aufweisen. Auch $B_3$ ist ein klarer Fall eines Exemplars der Art B. Perfekte Exemplare sind immer auch klare Exemplare, klare Exemplare sind aber nicht immer perfekte Exemplare.

Darüber hinaus ist es möglich, dass Individuen Eigenschaften besitzen, die für andere Arten typisch sind. Beispielsweise exemplifiziert $A_2$ die für Spezies D typische Eigenschaft 16. Wenn die Art oder Anzahl abweichender Eigenschaften besonders hoch ist, oder wenn besonders wenige der typischen Eigenschaften in einem Exemplar realisiert sind (z.B. weist $D_3$ nur zwei der vier D-typischen Eigenschaften auf, nämlich 16 und 18), dieses Exemplar aber trotzdem zur fraglichen Art gehört, könnte man von *freak entities* (Hawley/Bird 2011: 13) sprechen. In Anlehnung daran wäre sicher auch ein Konzept von *freak kinds* im Verhältnis zu einem höheren Genus sinnvoll. Es würde sich dann um Arten handeln, die zwar immer noch zu einem Genus gehören, deren Exemplare aber besonders wenige (aber trotzdem noch hinreichend viele) der Genus-typischen Eigenschaften aufweisen.

Die Idee einer *Realisierungslücke* lässt sich nun so rekonstruieren: Entscheidend für das Zustandekommens eines eine Art oder Gattung konstituierenden Eigenschaftsclusters ist, dass existierende Individuen nicht einfach Eigenschaften beliebiger Teilmengen einer Eigenschaftsmenge wie E = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19} exemplifizieren, sondern dass nur einige der – im Falle von E $2^{19}$ (= 524288) – logisch möglichen Eigenschaftskombinationen tatsächlich in Individuen realisiert sind. Bestimmte Teilmengen sind realisiert (z.B. {1, 2, 4, 5} aufgrund von $A_1$), andere Teilmengen sind nicht realisiert (z.B. {1, 2, 3, 10, 11, 12}).

Aber auch dies ist noch nicht hinreichend: Die realisierten und nicht realisierten Möglichkeiten sind in der 19-dimensionalen Matrix, die sich bei der Kombination jeder Eigenschaft mit jeder anderen ergibt, nicht einfach gleich verteilt, sondern bilden Muster. Bestimmte Kombinationen und „benachbarte Kombinationen" sind realisiert,[12] bestimmte andere Kombinationen und benachbarte Kombinationen sind nicht realisiert, so dass größere Räume mit realisierten Kombinationen (*Realisierungshäufungen*) und größere Räume mit nicht (oder weniger) realisierten Kombinationen entstehen. Diese Lücken zwischen den Realisierungshäufungen können als Realisierungslücken interpretiert werden. Wohlgemerkt: Es geht dabei nicht darum, ob bestimmte Eigenschaftskombinationen häufiger oder seltener realisiert sind (ob Individuen bestimmter Typen häufiger oder seltener vorkommen), sondern es geht um die Frage, ob bestimmte Eigenschaftskombinationen überhaupt realisiert sind oder nicht. Wie groß bzw. wie „dünn besiedelt" die Lücken sein müssen, damit man von legitimen Realisierungslücken sprechen kann, dürfte sich freilich schwerlich a priori und pauschal beantworten lassen. Man wird sagen können, dass an diesem Punkt die *Gradualität des Natürliche-Arten-Konzepts*, die vom HPC-Ansatz behauptet wird, selbst offensichtlich wird.[13] Manche Arten sind klarere Fälle von natürlichen Arten, wenn und weil die Instanzen aufgrund größerer und klarerer Realisierungslücken von Nicht-Instanzen getrennt sind, andere Arten sind weniger klare Fälle natürlicher Arten, weil die Realisierungslücken kleiner

[12] Eine Kombination x von Eigenschaften ist zu einer anderen Kombination y „benachbart", wenn die meisten der Eigenschaften aus x mit denen aus y übereinstimmen.

[13] „The naturalness of a natural kind is a matter of the contribution that reference to it makes to the satisfaction of the accomodation demands of a disciplinary matrix" (Boyd 1999: 158). „Naturalness" ist hier als gradierbarer Ausdruck zu verstehen: Je natürlicher, desto größer die Eignung der Arten, zu den „accomodation demands" beizutragen.

oder weniger klar sind. Bei der Frage, ob im Rahmen einer wissenschaftlichen Disziplin bestimmte Phänomene begrifflich unterschieden oder zusammengefasst werden (taxonomisches *Splitting* und *Lumping*, vgl. McKusick 1969, Craver 2009), spielen daher auch theoretische, forschungspragmatische und andere Interessen eine Rolle.

Neben den durch die Kombinatorik von Eigenschaften zustande kommenden Vagheitsphänomenen existiert als weitere Vagheitsquelle die Möglichkeit, dass die Eigenschaften (1, 2, 3 usw.) selbst bestimmte Formen von Gradualität zulassen. Eigenschaft 1 könnte z.B. mehr oder weniger realisiert sein. Man würde dann sagen, dass das die Eigenschaft 1 bezeichnende Prädikat *sorites-unscharf* ist.

Wenn, wie oben argumentiert wurde, der semantische Externalismus auch bei HPC-Arten zutrifft, wenn also der die Art bezeichnende Terminus direkt auf die Art referiert, ähnlich wie ein Eigenname direkt auf ein Individuum referiert, dann liegt es nahe anzunehmen, dass die von natürlichen Arten aufgewiesene Vagheit eher der bei Individuen (Namen) auftretenden ähnelt und weniger der prädikativen Variante. So wird verständlich, wie die von Hawley und Bird (2011) beobachtete Ähnlichkeit der bei HPC-Arten auftretenden kombinatorischen Vagheit mit der als *problem of the many* bekannten Schwierigkeit zustande kommt. Da $D_2$ von den D-typischen Eigenschaften 16, 17, 18 und 19 nur zwei aufweist, könnte es etwa unbestimmt sein, ob es sich um ein Exemplar der Spezies D handelt. $D_2$ instanziiert klarerweise nicht die komplexe Universalie 16+17+18+19. Da es demgegenüber unbestimmt sein soll, ob $D_2$ D instanziiert, kann die Universalie D nicht mit der komplexen Universalie 16+17+18+19 identifiziert werden.

Die Parallele zum *problem of the many* ergibt sich aus folgender Überlegung. Das *problem of the many* betrifft bestimmte Einzeldinge, deren Grenzen unscharf zu sein scheinen (Unger 1980). Bei einer Wolke gibt es beispielsweise einige Wassermoleküle an oder in den Randbereichen der Wolke, bei denen unbestimmt ist, ob sie zur Wolke gehören oder nicht. Es gibt also mehr als eine jeweils exakt bestimmte Menge, bzw. ein mereologisches Aggregat von Molekülen, die man mit der Wolke identifizieren könnte. Diese Mengen oder Aggregate können als zulässige Präzisierungen des Wolkenbegriffs aufgefasst werden. Dem Supervaluationismus zufolge ist ein Satz wie „Molekül x ist Teil der Wolke“ genau dann wahr, wenn er in allen Präzisierungen des Begriffs „Wolke“ wahr ist (er ist dann „super-wahr“). Gibt es einige zulässige Präzisierungen, in denen der Satz wahr, einige, in denen er falsch ist, so gilt er als unbestimmt (ist er in allen Präzisierungen falsch, ist er definitiv falsch („super-falsch“)). Moleküle, die bei einigen Präzisierungen zur Wolke gehören, bei anderen nicht, sind keine definitiven Teile der Wolke, sondern stellen Grenzfälle dar. Ein analoges Bild ergibt sich bei der HPC-Art D. Es gibt mehrere exakt definierte Kandidaten-Universalien (16+17+18, 16+17+18+19 usw.). Einige Objekte mögen klare (oder gar perfekte) Instanzen sein (so wie B4 eine perfekte B-Instanz ist), andere Objekte sind Grenzfälle (wie $D_2$ bei D). Wiederum andere Objekte sind klarerweise keine Instanzen von D, z.B. $A_1$. Offenbar ist es, wie Hawley und Bird schlussfolgern, schlicht unbestimmt, welche der in Frage kommenden Eigenschaftsmengen der Universalie D entspricht.

Ein weiterer für das Thema Vagheit relevanter Aspekt kommt durch die Möglichkeit *historischer Veränderungen* von Arten ins Spiel. Biologische Spezies – für Boyd paradigmatische HPC-Arten schlechthin – zeichnen sich insbesondere durch ihren historischen Charakter aus. Vermutlich ist es sinnvoll, sich historisch verändernde Arten als speziellen Subtyp natürlicher Arten generell anzusehen, wobei aber zu bedenken ist, dass auch etwa chemische Elemente, die häufig als paradigmatische Fälle ewiger, unveränderlicher Arten angesehen werden, insofern historischen Charakter haben, als es sie noch nicht immer gab, sondern die verschiedenen Elemente im Laufe der kosmischen Entwicklung nach dem Urknall nach und nach entstanden sind. Anhand des Diagramms kann der Prozess der Evolution einer Art so veranschaulicht werden: Bislang typische Exemplare einer Art wie A ($A_1$, $A_3$) werden seltener (bis hin zu ihrem Verschwinden), stattdessen werden untypische

Exemplare wie $A_2$ häufiger. Damit kann sich die Menge der A-typischen Eigenschaften von {4, 5, 6} vielleicht zu {5, 6, 16} verlagern. Darüber hinaus besteht die Möglichkeit, dass die Veränderungen so drastisch sind – vielleicht sind für spätere Exemplare die Eigenschaften {6, 16, 17} typisch –, dass man gar nicht mehr von „Spezies A" sprechen möchte, sondern lieber sagen will, eine neue Art sei entstanden. Auch diese Alternative (wann ist die Veränderung „drastisch genug"?) basiert letztlich auf – freilich theoretisch orientierten und begründeten – forschungspragmatischen Entscheidungen.

## 6. Schluss

Im Gegensatz zu essentialistischen Ansätzen ergibt sich aus der HPC-Theorie keine Unvereinbarkeit des Begriffs einer natürlichen Art mit verschiedenen Vagheitsformen. Ich habe für die Notwendigkeit argumentiert, unterschiedliche Vagheitsphänomene zu differenzieren. Auf eine gewisse Form von extensionaler Präzision eines Natürliche-Art-Begriffs ist auch ein HPC-Theoretiker festgelegt. Diese habe ich mit dem Begriff der Realisierungslücke zu modellieren versucht. Vagheit kommt aber dadurch ins Spiel, dass nicht genau festgelegt ist, wie groß und deutlich die Realisierungslücken sein müssen, die eine Art von anderen abgrenzt, damit man von einer natürlichen Art sprechen kann. Dadurch erweist sich das Konzept einer natürlichen Art selbst als abstufbar und vage. Ein anderes Vagheitsphänomen besteht darin, dass bei bestimmten Exemplaren unbestimmt sein kann, ob sie eine bestimmte Art instanziieren oder nicht. Darüber hinaus besteht die Möglichkeit, dass bestimmte Exemplare einer Art zwar klare Fälle sind, obwohl sie nicht alle der für die Art typischen Eigenschaften aufweisen. Da Natürliche-Art-Begriffe direkt referieren, gibt es zudem Parallelen zum *problem of the many*. Unbestimmtheit bei der Individuierung einer HPC-Art kann es schließlich auch diachron geben, da Arten historisch veränderlich sind und keine präzise Grenze an der Stelle existiert, ab der die Veränderung so drastisch ist, dass es sich nicht mehr um dieselbe Art handelt.

**Rico Hauswald**

Humboldt-Universität zu Berlin
rico.hauswald@hu-berlin.de

## Literatur

Aristoteles 2006: *Kategorien*. Übers. und erl. von K. Oehler. Berlin: Akad.-Verl.

Bird, A. 2007: *Nature's metaphysics. Laws and properties*. Oxford: Oxford University Press.

Boyd, R. 1991: „Realism, Anti-Foundationalism and the Enthusiasm for Natural Kinds", *Philosophical Studies* 61, 127-48.

— 1999: „Homeostasis, Species, and Higher Taxa", in R. Wilson (Hrg.): *Species*, Cambridge (MA): MIT Press, 141-86.

— 2010: „Realism, Natural Kinds, and Philosophical Methods", in H. Beebee und N. Sabbarton-Leary (Hrg.): *The semantics and metaphysics of natural kinds*. New York: Routledge, 212-234.

Craver, C. 2009: „Mechanisms and natural kinds", *Philosophical Psychology* 22, 575-94.

Devitt, M. 1981: *Designation*. New York: Columbia University Press.

Ellis, B. D. 2001: *Scientific essentialism*. Cambridge: Cambridge University Press.

Evans, G. 1982: *The varieties of reference*. Oxford: Oxford University Press.

Fine, K. 1975: „Vagueness, Truth and Logic", *Synthese* 30, 265-300.

Geach, P. T. 1980: *Reference and Generality*. 3. Auflage. Ithaca, N.Y.: Cornell University Press.

Goodman, N. 1970: „Seven Strictures on Similarity“, in L. Foster und J. W. Swanson (Hrg.): *Experience and Theory*, Cambridge (MA): University of Mass. Press, 19–29.

Grice, H. P. 1989: *Studies in the Way of Words*. Cambridge: Harvard University Press.

Hacking, I. 2007: „Putnam's Theory of Natural Kinds and Their Names is not the same as Kripke's“, *Principia* 11, 1–24.

Hawley, K., Bird, A. 2011: „What are Natural Kinds?“ http://www.st-andrews.ac.uk/~kjh5/OnlinePapers/WhatAreNaturalKinds.pdf

Kripke, S. 1980: *Naming and Necessity*. Cambridge (MA): Harvard University Press.

LaPorte, J. 2004: *Natural Kinds and Conceptual Change*. Cambridge: Cambridge University Press.

Lowe, E. J. 2009*: More kinds of being. A further study of individuation, identity, and the logic of sortal terms*. 2. Aufl. Chichester: Wiley-Blackwell.

McKusick, V. 1969: „On lumpers and splitters, or the nosology of genetic disease“, *Perspectives in biology and medicine* 12, 298-312.

Mill, J. S. 1973: *A system of logic, ratiocinative and inductive. Being a connected view of the principles of evidence and the methods of scientific investigation*. Toronto: University of Toronto Press.

Millikan, R. G. 1999: „Response to Boyd's Commentary“, *Philosophical Studies* 95, 99–102.

Pinkal, M. 1985: *Logik und Lexikon. Die Semantik des Unbestimmten*. Berlin: de Gruyter.

Putnam, H. (1975): „Is semantics possible?“, in *Philosophical Papers, Vol. 2: Mind, Language and Reality*, Cambridge: Cambridge University Press, 139–52.

— 1975b: „The Meaning of 'Meaning'“, in *Philosophical Papers, Vol 2: Mind, Language and Reality*, Cambridge: Cambridge University Press, 215–71.

— 1990: „Is Water Necessarily H2O?“, in H. Putnam und J. Conant (Hrg.): *Realism with a Human Face*, Cambridge (MA): Harvard University Press, 54–79.

— 1991: „Explanation and Reference“, in R. Boyd, P. Gasper und J. D. Trout (Hrg.): *Readings in the Philosophy of Science*. Cambridge (MA): Bradford, 171–86.

Reydon, T. 2009: „How to Fix Kind Membership: A Problem for HPC Theory and a Solution“, *Philosophy of Science* 76, 724-736.

Unger, P. 1980: „The Problem of the Many“, *Midwest Studies in Philosophy* 5, 411-67.

Waismann, F. 1945: „Verifiability“, *Proceedings of the Aristotelian Society* 19, 119-50.

Wittgenstein, L. 1984a: „Philosophische Untersuchungen“, in *Werkausgabe, Bd. 1.*, Frankfurt/M.: Suhrkamp, 225–618.

— 1984b: „Über Gewißheit“, in *Werkausgabe, Bd.* 8., Frankfurt/M.: Suhrkamp, 113–443.

# Epistemische und nicht-epistemische Werte in der angewandten Forschung

Gertrude Hirsch Hadorn

Theorien können verschiedenen Arten von Zwecken dienen. In diesem Beitrag wird aus der Position eines pluralistischen Pragmatismus dafür argumentiert, dass sich die Beurteilung von Theorien in der Grundlagenforschung und der angewandten Forschung darin unterscheiden sollen, welche epistemischen und nicht-epistemischen Werte in beiden Fällen dem jeweiligen Zweck der Theorie angemessen sind. Epistemische und nicht-epistemische Werte haben verschiedene Funktionen. Epistemische Werte artikulieren meist vage ein Ideal wissenschaftlicher Theorie. Sie dienen als Standards zur Beurteilung von Theorien. Dabei ist zwischen Standards für die Adäquatheit von Evidenz und solchen für strukturelle Adäquatheit von Theorien zu unterscheiden. Epistemische Standards haben somit eine direkte Funktion, nicht-epistemische wie moralische und prudentielle Werte hingegen eine indirekte. Sie dienen dazu, epistemische Standards zu spezifizieren und zu gewichten. Diese verschiedenen Funktionen sind zu berücksichtigen wenn zu beurteilen ist, ob die Verwendung von Theorien für lebensweltliche Probleme zulässig ist. Es wird gezeigt, dass in den Richtlinien zur Beurteilung der Unsicherheit von Modellen in den Sachstandsberichten des Weltklimarates einerseits Adäquatheit von Evidenz und strukturelle Adäquatheit von Modellen unklar vermischt sind, andererseits auch die indirekte Funktion nicht-epistemischer Werte kaum bedacht wird.

## 1. Einleitung

Einer traditionellen Auffassung zufolge, wie sie beispielsweise von Giere (2003) vertreten wird, ist die praktische Beurteilung von Theorien hinsichtlich ihrer Eignung als Handlungsgrundlage für Zwecke der Lebenswelt anhand von nicht-epistemischen Werten von ihrer epistemischen Beurteilung als gerechtfertigte Überzeugung anhand von epistemischen Werten zu unterscheiden. Für die praktische Beurteilung wird die epistemische Beurteilung vorausgesetzt und unterstellt, dass die epistemische Beurteilung von Theorien in der Grundlagenforschung die Rationalität von wissenschaftlichen Theorien überhaupt sichert (Hansson 2007, Carrier 2004). Die Termini „Grundlagenforschung“ und „angewandte Forschung“ verdanken dieser Unterstellung viel von ihrer Plausibilität.

In der Kritik an der traditionellen Position wird erstens in Frage gestellt, ob epistemische Werte der Grundlagenforschung für die Beurteilung von Theorien in der angewandten Forschung überhaupt geeignet sind. Grund dafür ist, dass sich die Verwendung akzeptierter Theorien der Grundlagenforschung für die Erforschung lebensweltlicher Probleme verschiedentlich als ein Fehlschlag erwiesen hat, beispielsweise im Falle von medizinischen Problemen (z.B. Cartwright und Munro 2011, Carrier und Finzer 2011), oder von Umweltproblemen (z.B. Shrader-Frechette 1989, Kriebel et al. 2001). Zweitens wird die Unterscheidung epistemischer Werte von nicht-epistemischen grundsätzlich in Frage gestellt. Ein Argument dafür lautet, dass epistemische Werte auf sozialen Konventionen beruhen, was einen Begriff von „epistemisch“ im Sinne der analytischen Epistemologie in Frage stellt. Dieser Kritik zufolge sollte die relevante Diskussion darüber geführt werden, welche Werte auf welche Weise gut oder schlecht für die Wissenschaft sind. Ferner sollte beachtet werden, dass in allen Phasen eines Forschungsprozesses Entscheidungen getroffen werden, nicht nur

dann, wenn es um die Akzeptierbarkeit von Theorien geht (z.B. Longino 1990, Douglas 2000, Machamer und Osbeck 2004).

Ich teile die erste Kritik an der Universalität epistemischer Werte der Grundlagenforschung und argumentiere im Folgenden aus der Position eines pluralistischen Pragmatismus, wie er z.B. von Churchman (1956), McLaughlin (1970) oder Foley (1988) vertreten wird. Was die zweite Kritik betrifft, so stimme ich zwar zu, dass eine wichtig Frage darin besteht, welche Werte auf welche Weise in welchen Entscheidungen eine Rolle spielen sollen. Doch weise ich die zweite Kritik mit dem Argument zurück, dass die Grundlage für eine adäquate Antwort auf diese Frage gerade darin besteht, in geeigneter Weise zwischen epistemischen und nicht-epistemischen Werten zu unterscheiden. Dies gilt insbesondere für die Beurteilung von Theorien - womit im Folgenden auch Modelle gemeint sind -, auf die ich mich in diesem Beitrag konzentriere.

Im zweiten Abschnitt diskutiere ich die Kritik von Shrader-Frechette (1989, 1997), von Douglas (2000) sowie von Cartwright (2006) an der traditionellen Auffassung epistemischer und nicht-epistemischer Werte in der Beurteilung von Theorien. Dieser Kritik entnehme ich drei Vorschläge, die ich im dritten Abschnitt im Zusammenhang mit einer Klärung des Begriffs „epistemischer Wert“ ausarbeite: (i) einen funktionalen Unterschied zwischen epistemischen und nicht-epistemischen Werten, (ii) eine Differenzierung innerhalb der epistemischen Beurteilung von Theorien, und zwar einerseits hinsichtlich der Adäquatheit der Evidenz und andererseits hinsichtlich struktureller Adäquatheit, sowie (iii) ein pragmatisches Verständnis von „Adäquatheit“, welches beinhaltet, dass eine Theorie für den jeweiligen Zweck adäquat sein soll, was einen Pluralismus epistemischer und nicht-epistemischer Werte zur Folge hat. Daraus ziehe ich sodann Konsequenzen für eine Gegenposition zur traditionellen Auffassung, und zwar im Form einer universalen Konzeption der Funktionen von epistemischen und nicht-epistemischen Werten in der Beurteilung von Theorien, die je nach Zweck einer Theorie zumindest teilweise unterschiedliche epistemische und nicht-epistemische Werte erfordern, so auch im Falle von Grundlagenforschung und angewandter Forschung. Die Nützlichkeit dieser Überlegungen zeige ich im vierten Abschnitt anhand der Kontroverse über die Richtlinien zur Beurteilung der Unsicherheit von Modellen bzw. Modellergebnissen in den Sachstandsberichten des Weltklimarates auf. Ein zentrales Problem dieser Richtlinien ist, dass Fragen der Adäquatheit der Evidenz mit Fragen der strukturellen Adäquatheit von Modellen auf unklare Weise vermischt sind. Die Klärung der verschiedenen epistemischen Standards, die in diesem Zusammenhang relevant sind, ist gerade angesichts der nicht-epistemischen Folgen des Fehlerrisikos wichtig.

## 2. Kritik an der traditionellen Auffassung

Angestoßen durch ihre Analyse von Expertengutachten zur Sicherheit eines Standortes für stark radioaktive Abfälle und eines Standortes für schwach radioaktive Abfälle in den USA, die insbesondere im zweiten Fall zu einer gravierenden Fehleinschätzung kamen, wirft Shrader-Frechette (1989, 1997) die Frage auf, anhand welcher Kriterien entschieden werden soll, ob Theorien, die in der Grundlagenforschung akzeptiert sind, auch in der angewandten Forschung verwendet werden dürfen – ob beispielsweise das hydrogeologische Gesetz von Darcy verwendet werden darf um abzuschätzen mit welcher Geschwindigkeit radioaktive Stoffe mit dem Grundwasser in einer bestimmten geologischen Formation wandern, die als Endlager dienen soll. Ausgangspunkt ihrer Überlegungen ist, dass Modelle der Grundlagenforschung aus systematischen Gründen fehlerhaft sein können. Ein erster Grund ist die für Grundlagenforschung konstitutive Idealisierung in Form von Vereinfachungen bei der Konstruktion der Modelle von empirischen Situationen, bei der mathematischen Formulierung der Modelle und bei der Konstruktion der Experimente bzw. Simulationen. Ein

weiterer Grund sind prinzipielle Probleme der empirischen Überprüfbarkeit von Modellen, wenn beispielsweise relevante Prozesse von historischen Umständen abhängen, wenn Modellvorhersagen nicht überprüfbar sind wie im Falle grosser Zeiträume, wenn unter Wissenschaftlern kontrovers ist, welche Parameter relevant sind u.a.

Zunächst schlägt Shrader-Frechette (1989) vor, die epistemischen Werte der Grundlagenforschung um zwei zusätzliche Kriterien zu erweitern, wenn es darum geht zu beurteilen, ob die Verwendung einer in der Grundlagenforschung akzeptierten Theorie in der angewandten Forschung zulässig ist. Erstens erachtet sie es für nötig, dass die Idealisierung durch die Berücksichtigung vernachlässigter Faktoren korrigiert wird. Ihr zweites Kriterium betrifft sodann den Grad an empirischer Genauigkeit, der für eine Anwendung zu erfüllen ist. Der im dritten Abschnitt diskutierten Unterscheidung innerhalb epistemischer Werte zufolge betrifft das erste Kriterium die strukturelle Adäquatheit von Theorien, das zweite ihre Adäquatheit in Bezug auf Evidenz. In einem späteren Aufsatz (Shrader-Frechette 1997) greift sie statt dessen die traditionelle Unterscheidung zwischen einer epistemischen Rationalität von Überzeugungen und einer ethischen Rationalität von Handlungen auf und fordert „to employ ethical rationality as well as scientific rationality“ (Shrader-Frechette 1997: S149, siehe auch S157, S158). Dieses additive Verhältnis wird jedoch durch ihre weiteren Ausführungen in Frage gestellt. Ist eine Theorie nicht gut genug und keine bessere Alternative vorhanden, dann stimmt sie im Falle der Grundlagenforschung der Auffassung zu, diese Theorie aus epistemischen Gründen zu verwenden. Im Falle der angewandten Forschung vertritt sie hingegen die Position, dass ethische, d.h. moralische und prudentielle Gründe eine Verwendung verbieten, um negative Folgen für Betroffene zu vermeiden. Bei der Entscheidung, welches Fehlerrisiko zu minimieren ist, vertritt sie die Position, dass epistemische Gründe dafür sprechen, in der Grundlagenforschung das Risiko von Fehlern erster Art, d.h. von falsch positiven Resultaten, zu minimieren, während ethische Gründe dafür sprechen, in der angewandten Forschung das Risiko von Fehlern zweiter Art, d.h. von falsch negativen Ergebnissen, zu minimieren. Sowohl bei der Entscheidung, ob eine schlechte Theorie verwendet werden soll, als auch bei der Entscheidung, welches Fehlerrisiko zu minimieren ist, kann aber jeweils nur eine der beiden Alternativen gewählt werden. Das spricht gegen ein additives Verhältnis von epistemischer und ethischer Rationalität. Shrader-Frechette selbst deutet dies an anderer Stelle auch an, denn sie kritisiert „to extend criteria from one domaine (pure science) to another (applied science) for which they may not be well suited“ (Shrader-Frechette 1997: S157). Ein zweites Problem besteht darin, dass die Unterscheidung zwischen epistemischer und ethischer Rationalität selbst unklar ist. Denn bei der Frage, ob Fehler erster oder zweiter Art minimiert werden sollen, geht es nicht einfach darum, ob anhand eines epistemischen oder eines ethischen Kriteriums entschieden werden soll. Vielmehr ist jede der beiden Alternativen mit epistemischen und ethischen Kriterien verbunden, was ich im dritten Abschnitt im Zusammenhang mit einer Klärung des Begriffes „epistemischer Wert“ zeigen werde.

Für ihre Kritik an der traditionellen Auffassung von der Rolle epistemischer und nicht-epistemischer Werte bezieht sich Douglas (2000) wie schon Shrader-Frechette auf das induktive Risiko bei der Annahme empirischer Hypothesen bzw. Theorien. Ziel ihrer Argumentation ist zu zeigen, dass nicht-epistemische Werte nicht nur bei der wissenschaftsexternen Verwendung von Forschungsresultaten eine Rolle spielen und dies auch sollen, sondern auch bei zentralen Entscheidungen im Forschungsprozess selbst, und zwar nicht nur in Bezug auf die Akzeptanz von Modellen, sondern auch bei Entscheidungen über den methodischen Ansatz, die Interpretation der Daten u.a. Dabei schließt sie die Grundlagenforschung ein, weil deren Resultate als wissenschaftlich gesichert gelten und unter Umständen ohne weitere Prüfung möglicher negativer Folgen für lebensweltliche Probleme verwendet werden. Douglas greift auf die Analyse des induktiven Risikos von Hempel (1965) zurück: Da entweder das Risiko falsch positiver oder dasjenige falsch negativer Resultate durch eine entsprechende Festlegung des Signifikanzniveaus minimiert

wird, sind für eine gerechtfertigte Entscheidung darüber, ob eine Hypothese akzeptiert oder zurückgewiesen werden soll, Kriterien erforderlich, welche angeben, welcher der beiden möglichen Fehler – falsch positives oder falsch negatives Resultat – schwerer wiegt. Wenn Theorien für lebensweltliche Probleme verwendet werden, dann gilt es mögliche negative nicht-epistemische Folgen in Form von Schädigungen von Betroffenen zu vermeiden. Somit sprechen in diesem Fall ethische Gründe dafür, das Risiko von falsch negativen Resultaten zu minimieren. Douglas bezeichnet diese Funktion nicht-epistemischer Werte, in diesem Falle ethischer Werte, bei der Entscheidung über das Fehlerrisiko als eine „indirekte Rolle“ (Douglas 2000: 564).

Anhand von Studien zur Toxizität von Dioxin argumentiert Douglas sodann dafür, dass nicht nur bei der Akzeptanz von Theorien, sondern auch bei weiteren Entscheidungen ein Fehlerrisiko besteht: ob z.B. bei der Dosis-Wirkungsbeziehung ein Grenzwert angesetzt wird - und ggf. wo - oder ob linear extrapoliert wird, welcher methodische Ansatz für die Datenerhebung gewählt wird, wie die Daten interpretiert werden u.a. Bei Entscheidungen, welche Fehlerrisiken (nicht) in Kauf genommen werden sollen, sind aufgrund der nicht-epistemischen Folgen solcher Fehlerrisiken generell nicht-epistemische Werte zu berücksichtigen, so ihre These. Nicht-epistemische Werte spielen daher generell eine indirekte Rolle im Forschungsprozess. Elliot (2011) zeigt allerdings, dass nicht nur die Unterscheidung zwischen direkten und indirekten Rollen selbst, sondern auch Zweck und Anwendung der Unterscheidung bei Douglas klärungsbedürftig sind. Im dritten Abschnitt werde ich die Idee, direkte und indirekte Funktionen zu unterscheiden, aufgreifen und in einer bestimmten Weise verwenden. Nicht einig gehe ich allerdings mit Douglas in Bezug auf ihren Vorschlag, bei der Beurteilung von Theorien zwischen Grundlagenforschung und angewandter Forschung nicht zu unterscheiden. Zwar argumentiere ich im dritten Abschnitt dafür, dass nicht-epistemische Werte auch in der Grundlagenforschung eine indirekte Funktion haben. Doch spricht die Heterogenität der Zwecke, denen Theorien dienen sollen, und zwar sowohl innerhalb der Grundlagenforschung als auch der angewandten Forschung, für einen pluralistischen Pragmatismus in Bezug auf die jeweiligen epistemischen und nicht-epistemischen Werte, wie er z.B. von Foley (1988) vertreten wird.

Mit der Formel „evidence for use“ bezieht Cartwright (2006) Position gegen randomisierte kontrollierte Studien (RCTs), die als Goldstandard gesicherter empirischer Evidenz für die Wirksamkeit von medizinischen und anderen Maßnahmen gelten (Cartwright und Munro 2011, Cartwright und Hardie 2012). Für Cartwright stehen dabei nicht Kriterien zur Rechtfertigung eines bestimmten Signifikanzniveaus im Vordergrund, sondern Gründe dafür, warum eine Prognose für Situation B aufgrund eines Modelles, das sich in Situation A bewährt hat, falsch sein kann. Diese Gründe sieht sie darin, dass sowohl die kausale Rolle von Maßnahmen unklar ist als auch welche unterstützenden Faktoren in einem bestimmten Kontext relevant sind. „Evidence for use“ beinhaltet für sie die Fragestellung, welcher wissenschaftliche Zugang, z.B. bezüglich Kausalität, für welches System oder welchen Gebrauch unter welchen Randbedingungen adäquat ist. Sie kritisiert, dass randomisierte kontrollierte Studien aufgrund ihrer Vereinfachungen diese Fragestellung unterlaufen. Sie arbeitet daher an alternativen Verfahren zur Ermittlung der relevanten kausalen Komplexität in einer bestimmten Verwendungssituation. Für das von Cartwright als „evidence for use“ aufgeworfene Problem führe ich in Abschnitt 3.1 den Ausdruck „strukturelle Adäquatheit“ ein. So gesehen geht es Cartwright um die strukturelle Adäquatheit von Modellen für Massnahmen zur Lösung lebensweltlicher Probleme.

Der Kritik an der traditionellen Auffassung der Beurteilung von Theorien der angewandten Forschung entnehme ich drei Vorschläge, die ich für weiterführend erachte und im dritten Abschnitt ausarbeite: (i) einen funktionalen Unterschied zwischen epistemischen und nicht-epistemischen Werten im Anschluss an Hempel (1965) und Douglas (2000), (ii) eine Differenzierung innerhalb der epistemischen Beurteilung zwischen adäquater Evidenz und

struktureller Adäquatheit von Theorien im Anschluss an Überlegungen von Shrader-Frechette (1989) und Cartwright (2006) sowie (iii) ein pragmatisches Verständnis von „Adäquatheit", welches beinhaltet, dass Theorien für den jeweiligen Zweck adäquat sein sollen, was einen Pluralismus epistemischer und nicht-epistemischer Werte zur Folge hat, wie dies z.B. von Foley (1988) vertreten wird. Diese Vorschläge laufen der Annahme einer einseitigen epistemischen Abhängigkeit der angewandten Forschung von der Grundlagenforschung entgegen. Daher ist der Terminus „angewandte Forschung" irreführend. Während im Englischen mit dem Ausdruck „use" eine terminologische Alternative besteht - beispielsweise „evidence for use" (Cartwright 2006) oder „use-inspired basic research" (Stokes 1997) -, steht ein geeigneter Vorschlag im Deutschen noch aus, so dass ich im Folgenden den Ausdruck „angewandte Forschung" verwende.

## 3. Funktionen epistemischer und nicht-epistemischer Werte

Epistemische und nicht-epistemische Werte haben verschiedene Funktionen in der Beurteilung von Theorien: Epistemische Werte dienen als Standards zur Beurteilung von Theorien. Sie werden auf Theorien angewandt und haben somit eine direkte Funktion in der Beurteilung von Theorien. Nicht-epistemische Werte dienen hingegen dazu, epistemische Standards zu spezifizieren, z.B. die Festlegung des Signifikanzniveaus bei der Operationalisierung des epistemischen Standards „empirische Genauigkeit". Da nicht-epistemische Werte auf epistemische Standards angewandt werden, ist ihre Funktion bei der Beurteilung von Theorien indirekt. Epistemische Standards wie Universalität, Einfachheit, Erklärungskraft u.a. sind oftmals abstrakt und vage formuliert. Für ihre Anwendbarkeit sind Spezifikationen und Gewichtungen nötig, die auf unterschiedliche Weise vorgenommen werden können. Während die Praxis in der Grundlagenforschung diesbezüglich oftmals unkontrovers scheint und diskussionslos erfolgt (Kuhn 1977), gibt es dazu in der angewandten Forschung sehr kontroverse Debatten, nicht zuletzt aufgrund der traditionellen Auffassung der Beurteilung von Theorien.

### *3.1 Epistemische Werte*

Die direkte Funktion epistemischer Werte besteht darin, dass sie als Standards oder Kriterien in der Beurteilung von Theorien dienen. Mit den epistemischen Standards ist festgelegt, welche Eigenschaften erforderlich sind, um als eine gute wissenschaftliche Theorie anerkannt zu werden, oder aufgrund der Ausprägungen dieser Eigenschaften besser als eine andere Theorie abzuschneiden. Einem Vorschlag von Hempel (1965, 1983) zufolge artikulieren epistemische Standards ein Ideal wissenschaftlicher Theorie. Zu den epistemischen Kriterien zählt Hempel nicht nur empirische Genauigkeit und Konsistenz, welche die Evidenz für eine Theorie betreffen. Um als eine gute Theorie anerkannt bzw. besser als eine andere Theorie eingeschätzt zu werden, müssen Überzeugungen noch weitere Eigenschaften aufweisen, zu denen üblicherweise Universalität, Einfachheit, Fruchtbarkeit, Erklärungskraft u.a. zählt. Van Fraassen (1980) nennt diese weiteren Eigenschaften „pragmatic virtues". Im Rahmen einer pragmatischen Konzeption entsteht damit jedoch eine Unklarheit, da auch die Kriterien der Evidenz einer dem jeweiligen Zweck entsprechenden Präzisierung bedürfen, z.B. welches Fehlerrisiko minimiert werden soll. Daher übernehme ich Hempels breite Verwendung des Terminus „epistemischer Standard", unterscheide aber innerhalb der epistemischen Standards nochmals zwischen Kriterien für adäquate Evidenz und solchen für strukturelle Adäquatheit von Theorien. Hempel verwendet auch den Ausdruck „desiderata" (Hempel 1983). Dieser Ausdruck zeigt zwar den Bezug zum Ideal wissenschaftlicher Theorie als Grund für den normativen Status an, ist aber in Bezug darauf, was genau gewünscht wird, unklar: die Standards selbst oder dass Theorien bezüglich dieser Standards besser abschneiden.

Hempel (1965, 1983, 2000) wie auch Kuhn (1977, 1983) neigen zu einem universellen Ideal wissenschaftlicher Theorie, d.h. zu universellen epistemischen Kriterien. Dagegen sprechen nun meines Erachtens einige gute Gründe, die mit der Heterogenität der Wissenschaften, und zwar in verschiedener Hinsicht, zu tun haben: Erstens können Theorien, die in der Grundlagenforschung akzeptiert sind, für lebensweltliche Probleme fehlerhaft sein. Zweitens ist es innerhalb der Naturwissenschaften zu einem Wandel und einer Pluralisierung des Wissenschaftsverständnisses gekommen. Drittens gibt es eine grundlegende Heterogenität der Wissenschaftsbegriffe, über alle Disziplinen der Natur-, Sozial-, Geistes- und Technikwissenschaften hinweg betrachtet. Die verschiedenen Arten von Fragestellungen in den Wissenschaften sind mit unterschiedlichen epistemologischen, ontologischen u.a. Konzeptionen verbunden. Das spricht gegen ein universelles Ideal wissenschaftlicher Theorie. So ist beispielsweise das Kriterium der prädiktiven Genauigkeit nur sinnvoll für Theorien, welche sich auf beobachtbare empirische Regularitäten beziehen lassen. Verschiedene Arten von Fragestellungen lassen sich als verschiedene Zwecke verstehen, denen Theorien dienen sollen. Die Adäquatheit von Theorien hinsichtlich Evidenz und Struktur bemisst sich somit an ihrem jeweiligen Zweck, d.h. an den für diesen Zweck in geeigneter Weise spezifizierten epistemischen Kriterien für Evidenz und Struktur. Dies spricht für die Position eines pluralistischen Pragmatismus wissenschaftlicher Rationalität, die bereits in den Anfängen der Debatte über Wissenschaft und Werte z.B. von Churchman (1956) und McLaughlin (1970) vertreten worden ist, sowie später wieder z.B. von Foley (1988).

Aus der Position eines pluralistischen Pragmatismus lässt sich das Verhältnis von Grundlagenforschung und angewandter Forschung dahingehend bestimmen, dass die beiden Forschungsformen verschiedene Zwecke haben und demzufolge auch verschiedene Ideale bzw. epistemische Standards zur Beurteilung von Theorien verwenden sollen. Der Zweck, dem Theorien dienen - beispielsweise ob die Forschung darauf zielt, fundamentale Gesetze der Natur zu verstehen oder aber ein lebensweltliches Problem als Grundlage zur Ausarbeitung von Maßnahmen - soll daher nicht nur für die Festlegung der Kriterien für das Signifikanzniveau, sondern auch derjenigen für die strukturelle Adäquatheit einer Theorie relevant sein.

### *3.2 Nicht-epistemische Werte*

Moralische und prudentielle Werte zähle ich unter die nicht-epistemischen Werte. Sie dienen dazu, epistemische Standards zu spezifizieren und zu gewichten. Sie haben damit eine indirekte Funktion in der Beurteilung von Theorien. Moralische und prudentielle Werte haben abgesehen von ihrer indirekten Funktion in Bezug auf epistemische Standards natürlich auch direkte Funktionen in der Wissenschaft: wenn es um wissenschaftsethische Fragen des Forschungshandelns wie die Zulässigkeit oder die Effizienz von Experimenten geht, oder bei der Prioritätensetzung und Verteilung von Forschungsgeldern zu Forschungsthemen beispielsweise. Ich konzentriere mich hier jedoch auf die Funktion nicht-epistemischer Werte bei der Beurteilung von Theorien. Hier ist ihre Funktion indirekt, d.h. auf epistemische Standards bezogen.

Nicht nur in der angewandten Forschung, sondern auch in der Grundlagenforschung gilt es, vergleichsweise abstrakte und vage epistemische Standards zu spezifizieren. Nicht-epistemische Werte spielen diesbezüglich auch in der Grundlagenforschung eine Rolle. So spricht beispielsweise für die Regel, in der Grundlagenforschung das Risiko falsch positiver Ergebnisse zu minimieren, kein epistemisches Kriterium, sondern eine prudentielle Überlegung zur Effizienz in der Forschung, die beinhaltet, die Arbeit nicht in falsche Theorien zu stecken und deshalb das Risiko falsch positiver Resultate in der Grundlagenforschung zu minimieren. Es handelt sich hier um die Spezifikation des epistemischen Standards für Evidenz, die in der Grundlagenforschung aufgrund einer prudentiellen Überlegung und damit eines nicht-epistemischen Wertes erfolgt. In der angewandten Forschung spricht hingegen

eine moralische Überlegung dafür, das Risiko falsch negativer Ergebnisse zu minimieren. Dass moralische und prudentielle Gesichtspunkte für die Festlegung des Signifikanzniveaus empirischer Genauigkeit ausschlaggebend sind, also eine indirekte Funktion haben, verleiht diesen Werten somit nicht den Status eines epistemischen Standards. Auch funktionieren nicht-epistemische Werte nicht unbedingt wie ein Standard im technischen Sinne, sondern beinhalten Überlegungen, welche beispielsweise in die Festlegung von Regeln oder Richtgrössen münden.

Nicht-epistemische Werte sind auch für epistemische Kriterien struktureller Adäquatheit von Bedeutung. So sprechen Effizienzüberlegungen für einfache Theorien im Sinne von rechnerisch einfach handhabbaren Theorien, während ästhetische oder kognitive Gründe die Einfachheit der Form von Theorien betreffen. In welcher Bedeutung und ob überhaupt Einfachheit ein sinnvolles epistemisches Kriterium ist, hängt der Position eines pluralistischen Pragmatismus zufolge vom Zweck der Theorien ab. Indem nicht-epistemische Werte eine Funktion bei der Spezifizierung und Gewichtung von epistemischen Standards adäquater Evidenz und struktureller Adäquatheit haben, sind nicht-epistemische Werte für die Rechtfertigung von epistemischen Standards relevant, zusätzlich zu den ontologischen, epistemologischen u.a. Voraussetzungen des jeweiligen Forschungsgebietes, welche die Forschungsstrategie bestimmen und damit entsprechende epistemische Standards (Lacey 2004). Beruht die Konzeption eines Forschungsproblems beispielsweise auf einem Systembegriff wie im Falle der Klimaforschung, dann ist die Komplexität der zur Diskussion stehenden Modelle, d.h. Anzahl und Heterogenität der in Betracht gezogenen Variablen und Beziehungen, ein wichtiger epistemischer Standard. Der Grad der Auflösung in räumlicher und zeitlicher Hinsicht ist ein weiterer wichtiger epistemischer Standard für strukturelle Adäquatheit, wenn es auf der Basis eines Systemansatzes um die Verwendung von Modellen für lebensweltliche Probleme geht. Für die Rechtfertigung von epistemischen Standards zur Beurteilung von Theorien gilt es somit, Überlegungen theoretischer und praktischer Philosophie einzubeziehen und auf einander zu beziehen (McLaughlin 1970, Foley 1988).

### *3.3 Konsequenzen für eine Konzeption der Beurteilung von Theorien*

Der traditionellen Auffassung zufolge ist die Beurteilung von Theorien in der Grundlagenforschung und der angewandten Forschung konzeptionell verschieden. In der angewandten Forschung soll eine zweite, zusätzliche Beurteilung anhand von nicht-epistemischen Werten erfolgen, während die epistemischen Werte der Grundlagenforschung auch für angewandte Forschung gelten sollen. Ich argumentiere hingegen mit der unterschiedlichen Funktion epistemischer und nicht-epistemische Werte dafür, die Beurteilung von Theorien in der Grundlagenforschung und der angewandten Forschung nicht konzeptionell zu unterscheiden, da in beiden Fällen epistemische Standards eine direkte Funktion haben und nicht-epistemische Werte eine indirekte, d.h. bezogen auf die Spezifikation und Gewichtung epistemischer Standards. Hingegen sind die jeweiligen epistemischen Standards und nicht-epistemischen Werte für diese Forschungsformen mindestens teilweise verschieden.

Die Grundgedanken einer solchen allgemeinen Konzeption lassen sich in fünf Punkten zusammenfassen. (i) Der Kern ist ein Begriff von epistemischen Werten, welche ein Ideal wissenschaftlicher Theorie vage artikulieren und als Standards für Beurteilung von Theorien dienen. (ii) Was als gute Theorie anerkannt bzw. besser als eine andere Theorie ist, soll sich an Standards sowohl für Evidenz als auch für die Struktur von Theorien bemessen. Theorien sollen also hinsichtlich Evidenz und Struktur adäquat für die Art ihres Zweckes sein. (iii) Epistemische Standards bedürfen der Spezifikation und Gewichtung für die Anwendung. Dies ist die Funktion von nicht-epistemischen Werten wie moralischen und prudentiellen Überlegungen einerseits, aber auch von ontologischen, epistemologischen u.a. Voraussetzungen des jeweiligen wissenschaftlichen Zugangs andererseits. Unterschiedlichen

Arten von Zwecken sind mindestens teilweise unterschiedliche Spezifikationen und Gewichtungen epistemischer Standards angemessen. (iv) Der Spielraum dieses Pluralismus ist einerseits dadurch reguliert, dass die Anwendung epistemischer Standards auf Theorien sinnvoll sein muss, d.h. dass Theorien bestimmte Indikatoren für diejenigen Eigenschaften, die als epistemische Standards dienen, aufweisen müssen und bezüglich ihrer Indikatorwerte verbessert werden können. Diese Anwendbarkeitsbedingung kann die Korrektur von epistemischen Standards bzw. ihrer Spezifikation und damit der nicht-epistemischen Werte sowie der ontologischen und epistemologischen Überlegungen erfordern. Andererseits stehen wissenschaftliche Ideale und praktische Ideale bzw. Grundorientierungen in wechselseitigen Abhängigkeiten, was ein Potential zu wechselseitiger Korrektur beinhaltet. Diese Abhängigkeiten sind nicht auf die Eignung wissenschaftlicher Theorien für technische Maßnahmen beschränkt, sondern betreffen beispielsweise auch das Selbstverständnis von Menschen in einer Gesellschaft wie im Falle von Darwins Evolutionstheorie. (v) Die hier skizzierte Konzeption ist demzufolge nicht nur geeignet, die wissenschaftliche Praxis in Bezug auf die Beurteilung von Theorien zu beschreiben, sondern auch, diese unter dem Gesichtspunkt zu beurteilen, inwiefern den relevanten Unterscheidungen und Beziehungen epistemischer und nicht-epistemischer Werte Rechnung getragen wird. Damit kann sie zur Klärung von Debatten über epistemische Standards beitragen.

Diese Leitgedanken eines pluralistischen Pragmatismus wissenschaftlicher Rationalität bedürfen natürlich noch weiterer Ausarbeitung, was nicht im Rahmen dieses Beitrages erfolgen kann. Ich möchte an dieser Stelle statt dessen nochmals hervorheben, wie sich diese Position zur traditionellen Auffassung sowie zu der im zweiten Abschnitt diskutierten Kritik verhält. In Bezug auf die Funktion nicht-epistemischer Werte unterscheidet sich die hier vertretene Position eines pluralistischen Pragmatismus von der traditionellen Auffassung, der zufolge nicht-epistemische Werte ausschließlich und zusätzlich bei der angewandten Forschung ins Spiel kommen sollen, dahingehend, dass sie die indirekte Funktion nicht-epistemischer Werte in der Grundlagenforschung aufzeigt. Die Position des pluralistischen Pragmatismus unterscheidet sich aber auch von der Auffassung, nicht-epistemische Werte hätten in der Grundlagenforschung deshalb eine Rolle zu spielen, weil Grundlagenforschung letztlich doch für Anwendungen benützt wird und somit generell bei Entscheidungen, welche die Forschung betreffen, mögliche Folgen für die Gesellschaft aufgrund nicht-epistemischer Kriterien berücksichtigt werden sollten (Longino 2002, 2004, Douglas 2000, Kitcher 2001). Ich stelle vielmehr die Adäquatheit universeller Kriterien – epistemischer wie nicht-epistemischer – mit Blick auf die Heterogenität der Wissenschaften und ihrer Zwecke in Frage. Diese Heterogenität hat nicht nur Konsequenzen für Kriterien adäquater Evidenz, sondern auch für Kriterien struktureller Adäquatheit von Theorien. Daher kritisiere ich auch, die Fehlerhaftigkeit von Theorien auf das Problem des induktiven Risikos zu reduzieren, wie dies z.B. Douglas (2000) nahelegt. Wird die zentrale Frage der Beurteilung von Theorien angewandter Forschung nur in der Evidenz gesehen, verstellt sich der Blick für das, was Kriebel et al. (2001: 874) „Fehler dritter Art“ genannt haben: ein gut bestätigtes Modell, das für das fragliche Problem/Zweck strukturell inadäquat ist. Dieses Problem wird inzwischen beispielsweise von Parker (2010) oder Winsberg (2010) betont. Im vierten Abschnitt werde ich zeigen, dass die Vermischung dieser beiden Funktionen epistemischer Standards wesentlich zur Unklarheit der Richtlinien des Weltklimarates für die Beurteilung der Unsicherheiten von Modellen bzw. Modellresultaten beiträgt.

## 4. Fallbeispiel: Richtlinien des Weltklimarates zur Beurteilung der Unsicherheit von Modellen

Der Weltklimarat (Intergovernmental Panel on Climate Change IPCC) ist 1988 geschaffen und von der Vollversammlung der Vereinten Nationen anerkannt worden. Das

wissenschaftliche Gremium des IPCC hat den Auftrag, politischen Entscheidungsträgern auf der Grundlage der weltweit erschienenen Publikationen wiederholt über Informationen zum Klimawandel, die für Maßnahmen relevant sind, zu berichten. Der jüngste, vierte Sachstandsbericht stammt von 2007 (Intergovernmental Panel on Climate Change 2007). Die Berichterstattung umfasst die Beschreibung und Analyse bestehender Trends sowie Voraussagen künftiger Ereignisse und Trends. Drei Arbeitsgruppen teilen sich diese Aufgabe. Arbeitsgruppe I befasst sich mit den physikalischen Aspekten des Klimasystems und des Klimawandels, d.h. mit der Entwicklung der Temperatur und den Faktoren, welche diese beeinflussen. Arbeitsgruppe II berichtet über die Sensitivität und Vulnerabilität sozioökonomischer sowie natürlicher Systeme durch Klimawandel, die negativen und positiven Auswirkungen von Klimawandel sowie Möglichkeiten der Anpassung an diese Auswirkungen. Arbeitsgruppe III widmet sich Kosten und Nutzen von Maßnahmen, die den Klimawandel bremsen sollen, unter Bezug auf Emissionsszenarien und unter Berücksichtigung von technischen und ökonomischen Instrumenten sowie regulatorischen Maßnahmen. Der IPCC stützt sich ausschließlich auf bereits publizierte wissenschaftliche Arbeiten. Seine Aufgabe besteht darin, einerseits den inhaltlichen Stand der Informationen zusammenzufassen, der in allen drei Arbeitsgruppen zu einem großen Teil auf Modellen und Modellrechnungen beruht, und andererseits die Unsicherheiten dieser Modelle und Modellergebnisse sowie auch der Informationen anderer Art zu beurteilen, d.h. die Möglichkeit oder Wahrscheinlichkeit, dass diese wissenschaftlichen Informationen fehlerhaft sind. Beides soll der IPCC sodann in geeigneter Weise Entscheidungsträgern kommunizieren. Im Folgenden konzentriere ich mich auf die Beurteilung der Unsicherheiten durch den IPCC.

Seit dem dritten Sachstandsbericht arbeitet der IPCC verstärkt an Richtlinien für ein einheitliches Vorgehen bei der Beurteilung der Unsicherheiten wissenschaftlicher Informationen (Manning et al. 2004, Intergovernmental Panel on Climate Change 2005). Diese Richtlinien sind für den anstehenden fünften Sachstandsbericht weiter überarbeitet worden (Mastrandrea et al. 2010). Auch die Kommission, welche die Arbeit des IPCC vor einigen Jahren überprüft hat, widmet dieser Frage in ihrem Bericht ein eigenes Kapitel (InterAcademy Council 2010). Es scheint jedoch, dass die Kontroverse über die Richtlinien zum Umgang mit Unsicherheiten eher zunimmt. So widmet die führende Zeitschrift „Climatic Change“ im Jahr 2011 diesem Thema eine Sondernummer, siehe darin z.B. Moss (2011) und Jones (2011). Auch die Zeitschrift „Nature“ publiziert Beiträge dazu, weil dieses Thema mit der Gründung einer analogen Organisation für Fragen der Biodiversität und der Ökosystemdienstleistungen im Jahr 2012 zusätzlich Auftrieb erhalten hat (z.B. Turnhout et al. 2012, Westcott et al. 2012). Ich beziehe mich im Folgenden auf die zuletzt vom IPCC veröffentlichte „Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties“ (Mastrandrea et al. 2010), verfasst von einer Kerngruppe von 13 Hauptautoren des fünften Sachstandsberichtes, welche die drei Arbeitsgruppen vertreten.

In der Absicht, Entscheidungsträgern glaubwürdige Informationen zur Verfügung zu stellen, soll ihnen die Möglichkeit oder Wahrscheinlichkeit kommuniziert werden, dass die im Sachstandsbericht gemachten Aussagen nicht auf den tatsächlichen Klimawandel, seine Auswirkungen oder die Wirkungen der getroffenen Maßnahmen zutreffen. Dafür gibt es eine Vielfalt möglicher Quellen im Forschungsprozess. Die Behandlung der Unsicherheiten besteht somit in der Beurteilung vorliegender wissenschaftlicher Informationen anhand epistemischer Standards. Der IPCC schlägt einen Ansatz vor, der in allen drei Arbeitsgruppen gleichermaßen angewendet werden soll:

> These notes define a common approach and calibrated language that can be used broadly for developing expert judgments and for evaluating and communicating the degree of certainty in findings of the assessment process. (Mastrandrea et al. 2010: 1).

In diesem Zitat sind drei zentrale Problemkreise dieses Ansatzes angedeutet: (i) Den unterschiedlichen Zwecken wissenschaftlicher Modelle in der Grundlagenforschung und der angewandten Forschung wird in ihrer Bedeutung für die Adäquatheit von Modellen unzureichend Rechnung getragen, da der Ansatz generell verwendet werden soll. (ii) Überlegungen zur Adäquatheit von Evidenz und zu struktureller Adäquatheit sind im Begriff „degree of certainty" auf unklare Weise vermischt. (iii) Die indirekte Funktion nicht-epistemischer Werte sowie epistemologischer und ontologischer Voraussetzungen der verschiedenen Wissenschaften für eine adäquate Spezifikation epistemischer Standards ist zu wenig erkannt, was mit „calibrated language" angedeutet ist. Denn es wird sogleich angemerkt, dass spezifischere Richtlinien einer Arbeitsgruppe mit dem generellen Ansatz konsistent sein sollen.

Zur Beurteilung der Unsicherheiten führen die Richtlinien ein quantitatives und ein qualitatives Maß ein. Ersteres wird wie folgt erläutert:

> Quantified measures of uncertainty in a finding expressed probabilistically (based on statistical analysis of observations or model results, or expert judgment) (Mastrandrea et al. 2010: 1).

Es handelt sich hier um eine gestufte Likelyhood-Skala auf der Basis objektiver oder subjektiver Wahrscheinlichkeiten. Beispiele dafür sind, dass die Häufigkeit von Starkregen über den meisten Gebieten der Erde zumeist erst nach 1960 zugenommen hat (66-100% prob.) oder dass diese Zunahme anthropogen ist (>50-100% prob.). Die Likelyhood-Skala operationalisiert also empirische Genauigkeit, einen epistemischen Standard für Evidenz.

Das qualitative Maß wird wie folgt erläutert:

> Confidence in the validity of a finding, based on the type, amount, quality, and consistency of evidence (e.g., mechanistic understanding, theory, data, models, expert judgment) and the degree of agreement. Confidence is expressed qualitatively. (Mastrandrea et al. 2010: 1).

Auf der Basis verschiedener Kombinationen von Ausprägungen auf den beiden Dimension „evidence" und „agreement" wird eine fünfstufige Skala festgelegt. Ausdrücklich wird festgehalten, dass es sich bei „confidence" nicht um einen Begriff der Statistik handelt und „confidence" nicht probabilistisch interpretiert werden soll (Mastrandrea 2010: 3). Vielmehr soll „confidence" über die Validität von Ergebnissen nachvollziehbar informieren, und zwar anhand von Art, Ausmaß, Qualität und Konsistenz der Evidenz (Mastrandrea 2010: 2). Doch ist nicht klar, was damit gemeint ist. Falls damit die externe Validität gemeint ist, dann betrifft dies eigentlich die strukturelle Adäquatheit der wissenschaftlichen Modelle und Simulationen für das tatsächliche Klima und die beobachtbaren Prozesse. So wird auch festgehalten:

> Consider all plausible sources of uncertainty. Experts tend to underestimate structural uncertainty arising from incomplete understanding of or competing conceptual frameworks for relevant systems and processes. (Mastrandrea et al. 2010: 2).

In diesem Fall sollten die quantitative und die qualitative Skala Verschiedenes messen. Die Empfehlungen für die Verwendung der beiden Skalen deuten hingegen in eine andere Richtung. Einerseits wird festgehalten, dass die quantitative Skala zusätzlich oder alternativ zur qualitativen Skala angewendet werden kann, sofern die Voraussetzungen für die quantitative Skala gegeben sind (Mastrandrea et al. 2010: 1), andererseits wird kritisiert, dass die qualitative Skala auch verwendet wird, wenn die Anwendung der quantitativen zulässig ist (Risbey und Kandlikar 2007, InterAcademy Council 2010). Das deutet eher darauf hin, dass es sich um zwei Skalen für denselben Standard handelt. In diesem Falle fehlt jedoch ein Standard für externe Validität, d.h. für die strukturelle Adäquatheit von Modellen. Diese lässt sich nicht einfach auf der Basis empirischer Evidenz abschätzen, weil empirische Genauigkeit

auch auf bloßen Korrelationen beruhen kann, und somit keine gesicherten Voraussagen über tatsächliche Ereignisse oder Trends sowie auch Erklärungen zulassen, wie im zweiten Abschnitt im Anschluss an Cartwright als Kritik an der traditionellen Auffassung der epistemischen Beurteilung von Theorien für lebensweltliche Problem festgehalten ist. Dass Kriterien der strukturellen Adäquatheit wissenschaftlicher Modelle für die tatsächliche Entwicklung des Klimas, seiner Folgen und der Wirksamkeit möglicher Maßnahmen in den Richtlinien keine systematische Beurteilung erfährt, kommt auch darin zum Ausdruck, dass für die qualitative Einschätzung der Evidenz als Komponente der Konfidenz-Skala in der oben zitierten Erläuterung nur eine summarische Einschätzung von Typ, Umfang und Konsistenz der Evidenz bezüglich exemplarisch aufgelisteter Faktoren gewünscht wird.

Es scheint, dass sich die Richtlinien auf Evidenz beschränken, wobei der Begriff aufgrund des unklaren Verhältnisses der beiden Skalen mehrdeutig ist. Demzufolge ist die indirekte Funktion nicht-epistemischer Werte auch nur in Bezug auf das Fehlerrisiko angesprochen, und zudem zu knapp. Zwar werden Autoren angehalten zu beachten, welches Fehlerrisiko in den Originalarbeiten minimiert wurde, doch wird nicht ausgeführt, welche Konsequenzen aus dem jeweils gewählten Signifikanzniveau für die Einschätzung der Unsicherheit der Resultate im Sachstandsbericht zu ziehen sind.

In welchen Punkten der IPCC über seine gegenwärtigen Empfehlungen in der Beurteilung von Unsicherheiten hinausgehen muss, um seinen Bestrebungen nach glaubwürdiger Information über den Klimawandel, seine Auswirkungen und die Wirksamkeit von Maßnahmen tatsächlich nachzukommen, lässt sich anhand der Ausführungen im dritten Abschnitt aufzeigen. Erstens betrifft dies den soeben diskutierten Punkt, in Bezug auf die Adäquatheit von Modellen bzw. Modellergebnissen explizit zwischen Adäquatheit der Evidenz und struktureller Adäquatheit zu unterscheiden. Zweitens betrifft dies den Punkt, dass sich die Adäquatheit von Modellen bzw. Modellergebnissen am jeweiligen Zweck bemisst, was eine Überprüfung der Validität von Modellen bzw. Modellergebnissen für den intendierten Handlungsbereich erfordert. Drittens ist zu beachten, dass nicht nur nicht-epistemische Werte, sondern auch unterschiedliche ontologische und epistemologische Voraussetzungen der verschiedenen Wissenschaften, die insbesondere in der Arbeitsgruppe II vertreten sind, eine angemessene Spezifizierung der epistemischen Standards für das jeweilige Gebiet erfordern. Den Autoren der Sachstandsberichte und auch der Richtlinien sind diese Fragen keineswegs unbekannt - ganz im Gegenteil. Doch steht bis zu einer systematischen und transparenten Berücksichtigung dieser Fragen in den Richtlinien zum Umgang mit Unsicherheiten noch Arbeit an.

## 5. Schlussbemerkung

In diesem Beitrag kritisiere ich die traditionelle Position, welche die Funktion von moralischen und prudentiellen Überlegungen als eine additive Rolle bei der Beurteilung von Theorien angewandter Forschung versteht und sich dabei auf die Beurteilung des Fehlerrisikos bei der Annahme von Theorien beschränkt. Ich argumentiere stattdessen aus der Position eines pluralistischen Pragmatismus für eine indirekte Funktion nicht-epistemischer Werte in der Beurteilung von Theorien generell, welche darin besteht, epistemische Werte zu spezifizieren und zu gewichten. Da dies mit Blick auf den Zweck einer Theorie erfolgen soll, unterscheiden sich Grundlagenforschung und angewandte Forschung bezüglich der konkreten epistemischen Standards und der nicht-epistemischen Werte.

Die Argumentation stützt sich auf einen Begriff von epistemischen Werten, welche ein Ideal einer wissenschaftlichen Theorie mehr oder weniger vage artikulieren. Epistemische Werte dienen als Standards für die Beurteilung von Theorien sowohl in Bezug auf die Adäquatheit ihrer Evidenz als auch ihrer strukturellen Adäquatheit. Beide Arten von epistemischen

Standards gilt es dem Zweck der Theorie entsprechend zu spezifizieren und zu gewichten. Dies ist die Funktion von nicht-epistemischen Werten einerseits sowie von ontologischen und epistemologischen Voraussetzungen des jeweiligen Forschungsansatzes andererseits. Moralische und prudentielle Werte haben demzufolge eine indirekte Funktion in der Beurteilung von Theorien, da sie auf epistemische Standards angewendet werden um diese zu spezifizieren und zu gewichten. Ich zeige, dass sie diese indirekte Funktion auch in der Grundlagenforschung haben. Da in beiden Forschungsformen epistemische Standards eine direkte Funktion und nicht-epistemische Werte eine indirekte Funktion haben, unterscheide ich nicht auf der konzeptionellen Ebene zwischen der Beurteilung von Theorien in der Grundlagenforschung und der angewandten Forschung. Der relevante Unterschied zwischen den beiden Forschungsformen liegt vielmehr darin, dass die jeweiligen epistemischen Standards und nicht-epistemischen Werte mindestens teilweise verschieden sind, und zwar aufgrund unterschiedlicher Zwecke.

Eine Konsequenz dieser Konzeption besteht darin, im Interesse der Glaubwürdigkeit von Ergebnissen der angewandten Forschung der Frage der Adäquatheit epistemischer Standards für ihren Zweck vermehrt Aufmerksamkeit zu schenken und dabei die indirekte Funktion sowohl von nicht-epistemischen Werten also auch von ontologischen und epistemologischen Voraussetzungen des Forschungsansatzes bei der Spezifikation und Gewichtung epistemischer Standards zu beachten. Diese interdisziplinäre Aufgabe erfordert noch weitere Arbeit.

Auch die Konzeption sowie die damit verbundene Position eines pluralistischen Pragmatismus der Beurteilung wissenschaftlicher Theorien ist in diesem Beitrag nur als eine Skizze vorgelegt worden. Sie bedürfen ebenso weiterer Ausarbeitung. Wissenschaftliche Rationalität beinhaltet dieser Position zufolge, wissenschaftliche Theorien mit Bezug auf ein bestimmtes Ideal zu verbessern. Verschiedene Arten von Problemen oder Zielen erfordern verschiedene Ideale wissenschaftlicher Theorie. Churchman hat das wie folgt auf den Punkt gebracht:

> Here the situation is quite similar to that which occurs in production and distribution. [...] There is no such thing as a ‚good' rope: the best rope for anchoring a boat may be very poor rope indeed for hanging clothes – or men. (Churchman 1956: 248)

**Gertrude Hirsch Hadorn**

Institut für Umweltentscheidungen, ETH Zürich, CH 8092 Zürich
hirsch@env.ethz.ch

## Literatur

Carrier, M. 2004: „Knowledge and Control. On the Bearing of Epistemic Values in Applied Science", in P. Machamer und G. Wolters (Hrg.): *Science, Values, and Objectivity*, Pittsburg/Konstanz: University of Pittsburgh Press/Universitätsverlag Konstanz, 275–93.

– und P. Finzer 2011: „Theory and Therapy. On the Conceptual Structure of Models in Medical Research", in M. Carrier und A. Nordmann (Hrg.): *Science in the Context of Application. Methodological Change, Conceptual Transformation, Cultural Reorientation,* Dordrecht: Springer, 85–99.

Cartwright, N. 2006: „Well-ordered Science: Evidence for Use", *Philosophy of Science* 73, 981–90.

– und E. Munro 2010: „The Limitations of Randomized Controlled Trials in Predicting Effectiveness", *Journal of Evaluation in Clinical Practice* 16, 260–66.

– und J. Hardie 2012: *Evidence-based Policy. A Practical Guide to Doing it Better*. Oxford: Oxford University Press.

Churchman, C. W. 1956: "Science and Decision Making", *Philosophy of Science* 23, 247–49.

Douglas, H. E. 2000: „Inductive Risk and Values in Science", *Philosophy of Science* 67, 559–79.

Elliot, K. C. 2011: „Direct and Indirect Roles for Values in Science", *Philosophy of Science* 78, 303–24.

Foley, R. 1988: „Some Different Conceptions of Rationality", in E. McMullin (Hrg.): *Construction and Constraint. The Shaping of Scientific Rationality*, Notre Dame: University of Notre Dame Press, 123–52.

Giere, R. N. 2003: „A New Program for Philosophy of Science?", *Philosophy of Science* 70, 15–21.

Hansson, S. O. 2007: „Values in Pure and Applied Science", *Foundations of Science* 12, 257–68.

Hempel, C. G. 1965: „Science and Human Values", in *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*, New York/London: The Free Press/Collier-McMillan, 81–96.

– 1983: "Valuation and Objectivity in Science", in R. S. Cohen und L. Laudan (Hrg.): *Physics, Philosophy and Psychoanalysis. Essays in Honor of Adolf Grünbaum*, Reidel: Dordrecht, 73–100.

– 2000: „On the Cognitive Status and the Rationale of Scientific Methodology", in Jeffrey R. (Hrg.): *Selected Philosophical Essays*, Cambridge: Cambridge University Press, 199–228.

InterAcademy Council 2010: *Climate Change Assessments. Review of the Processes and Procedures of the IPCC*. Amsterdam: InterAcademy Council.

Intergovernmental Panel on Climate Change 2005: *Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties*. Geneva: Intergovernmental Panel on Climate Change.

– 2007: *Climate Change 2007: Synthesis Report*. Cambridge: Cambridge University Press.

Jones, R. N. 2011: „The Latest Iteration of IPCC Uncertainty Guidance — An Author Perspective", *Climatic Change* 108, 733–43.

Kitcher, P. 2001: *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Kriebel, D., J. Tickner, P. Epstein, J. Lemons, R. Levins, E. L. Loechler, M. Quinn, R. Rudel, T. Schettler und M. Stoto 2001: "The Precautionary Principle in Environmental Science", *Environmental Health Perspectives* 109, 871–76.

Kuhn, T. 1977: „Objectivity, Value Judgment, and Theory Choice", in *The Essential Tension*, Chicago: University of Chicago Press, 320–39.

– 1983: „Rationality and Theory Choice", *Journal of Philosophy* 80, 563–70.

Lacey, H. 2004: „Is There a Significant Distinction between Cognitive and Social Values?", in P. Machamer und G. Wolters (Hrg.): *Science, Values, and Objectivity*, Pittsburg/Konstanz: University of Pittsburgh Press/Universitätsverlag Konstanz, 24–51.

Longino, H. 1990: *Science as Social Knowledge*. Princeton: Princeton University Press.

– 2002: *The Fate of Knowledge*. Princeton: Princeton University Press.

– 2004: „How Values Can Be Good for Science", in P. Machamer und G. Wolters (Hrg.): *Science, Values, and Objectivity*, Pittsburg/Konstanz: University of Pittsburgh Press/Universitätsverlag Konstanz, 127–142.

Machamer P. und Osbeck, L. 2004: „The Social in the Epistemic“, in P. Machamer und G. Wolters (Hrg.): *Science, Values, and Objectivity*, Pittsburg/Konstanz: University of Pittsburgh Press/Universitätsverlag Konstanz, 78–89.

Manning, M.R., M. Petit, D. Easterlin, J. Murphy, A. K. Patwardhan, H.-H. Rogner, R. Swanrt, und G. W. Yohe (Hrg.) 2004: *IPCC Workshop on Describing Scientific Uncertainties in Climate Change to Support Analysis of Risk and of Options: Workshop Report*. Geneva: Intergovernmental Panel on Climate Change.

Mastrandrea, M. D., C. B. Field, T. F. Stocker, O. Edenhofer, K. L. Ebi, D. J. Frame, H. Held, E. Kriegler, K. J. Mach, P. R. Matschoss, G.-K. Plattner, G. W. Yohe und F. W. Zwiers 2010: *Guidance Note for Lead Authors of the IPCC Fifth Assessment Report on Consistent Treatment of Uncertainties*. Geneva: Intergovernmental Panel on Climate Change.

McLaughlin, A. 1970: „Science, Reason and Value“, *Theory and Decision* 1, 121–37.

Moss, R. 2011: „Reducing Doubt About Uncertainty: Guidance for IPCC's Third Assessment“, *Climatic Change* 108, 641–58.

Parker, W. 2010: „Whose Probabilities? Predicting Climate Change with Ensemble Models“, *Philosophy of Science* 77, 985–97.

Risbey, J. S. und M. Kandlikar 2007: „Expressions of Likelihood and Confidence in the IPCC Uncertainty Assessment Process“, *Climatic Change* 85, 19–31.

Shrader-Frechette, K. S. 1989: „Idealized Laws, Antirealism, and Applied Science: A Case in Hydrogeology“, *Synthese* 81, 329–52.

– 1997: „Hydrogeology and Framing Questions Having Policy Consequences“, *Philosophy of Science* 64, S149–60.

Stokes, D. E. 1997: *Pasteur's Quadrant: Basic Science and Technological Innovation*. Washington (DC): Brookings Institution Press.

Turnhout, E., M. Hulme, J. Vogel und B. Wynne 2012: „Listen to the Voices of Experience“, *Nature* 488, 454–55.

van Fraassen, B. C. 1980: *The Scientific Image*. Oxford: Clarendon Press.

Westcott, D. A., F. J. Kroon und A. W. Sheppard 2012: „Policy: Biodiversity Needs a Scientifc Approach“, *Nature* 490, 37.

Winsberg, E. 2010: “Models of Climate: Values and Uncertainties”, in *Science in the Age of Computer Simulation*, Chicago/London: University of Chicago Press, 93–119.

# Causation, Dispositions, and Mathematical Physics

Johannes Röhl

It has been denied that the concepts of cause and effect have any role in theories of systems with continuous time evolution like fundamental physics. Instead of causally connected events there are system states and their time development is described mathematically. This conflicts with causal descriptions in special sciences and with philosophical accounts that hold causation to be a relation between events. Such a central dissent could lead to a disunity of science and world. I argue that a world view with dispositions is able to accommodate both the viewpoint of mathematical physics and the one of "event causation". I claim that "cause-free" descriptions rest on presuppositions that employ causal notions: (1) Systemic closure, which means that all relevant causal features have been taken into account. (2) The mathematical description (the system's Hamiltonian) is based on the properties of the components that are relevant for their dynamic influences, their causal powers. In particular, classical forces can be understood as dispositions or as dependent on dispositions. And in quantum mechanics non-classical dispositions, propensities, capture important aspects of the ontology. Dispositions as relevant causal factors allow a unified ontology for both continuous processes and links of discrete events.

## 1. Introduction: Cause and Effect vs. Continuous Time Evolution

Philosophers have often denied that the concepts of cause and effect play any role in fundamental physics or other mathematical theories of systems with continuous time evolution (prominently Russell 1913 and more recently Earman 1986). According to this position mathematical theories of dynamical systems like the ones in both classical and modern physics and elsewhere do not deal with causally connected events. Some arguments for this position refer to the problems of the traditional view of causation. But more importantly it is stressed that in mathematical theories of physics there are only system states and a continuous time evolution of these states, described by differential equations or other mathematical operations, no discrete events or anything else that would correspond to the traditional notions of cause, effect and the causal relation between these entities. Usually, this view is taken to be eliminativist with respect to causation. On the other hand, sciences like biology and geology appeal to causes and effects in a sense very close to the traditional conceptions critized by Russell.

My goals in this paper are the following: I want to show that the conflict of these two views I call *event causation* and *continous time evolution* is deep and cannot be eliminated easily if we want to be realists about both fundamental mathematical sciences and "special sciences" dealing with higher levels of reality. Then I want to argue for a reconciliation of both views. To this end, I first try to show that the viewpoint of mathematical physics should not be taken as eliminativist, but rather as dependent on some quite general causal (but not "event-causal") presuppositions. Secondly, I argue that an ontology in which causal powers or dispositions are taken as the central elements of causation (rather then events or a relation between events) can accommodate both views and in this fashion some kind of reconciliation is possible. While striving for reconciliation it cannot be avoided to be mildly revisionist in some ways, so neither of the conflicting views will escape unscathed and there will not be a reduction of one to the other. In the remainder of this first section I will briefly sketch the

conflicting positions. In the second section I will show how a dispositional account of event causation could look like; in the third section I will discuss the causal presuppositions of the continous time evolution viewpoint and in the fourth section a dispositional grounding of this view will be sketched.

### *1.1 Event Causation*

I will be content with a somewhat rough characterization of what I take to be the core of the conception of causation in everyday speech, special sciences like medicine, biology, social sciences or the law that are not highly mathematized, and most of the philosophical literature on causation. This will be called "event causation" in the remainder of the paper. Examples abound:

(1) Tom's fall from the ladder caused the fracture of his collarbone.

(2) The rainstorm caused the flooding of the cellar.

Generally, C causes E; C and E are distinguishable events located somewhat precisely in space and time, and causation is some relation between events with certain features; it is asymmetric, irreflexive and transitive.

### *1.2 Continuous Time Evolution*

The traditional causes and effects (in the sense of event causation) are neither necessary for nor compatible with the continuity of the trajectories of a mechanical system in classical physics (or any other sufficiently mathematized theory): „In the motions of mutually gravitating bodies, there is nothing that can be called a cause and nothing that can be called an effect; there is merely a formula" (Russell 1913). Rather there is a state space that contains the states of the respective system, and their time evolution is described by mathematical operators acting on these states or a set of equations of motion that take parameters of initial states as initial values. The framework of state space formulation is a powerful and flexible tool. And the same case can be made in an ontologically less dubious fashion when looking at systems in ordinary space. If we describe a planet in its orbit around the sun we can completely describe this situation by giving the planet's continous trajectory which is functionally dependent on some parameters like the mass of the central body. There seems no need for a chain of distinguishable events like „sun being at point x at $t_0$ causes motion of earth at $t_1$" etc. The ingredients of the continous time evolution approach are:

(1) States {S(t)} of a physical system

(2) These states are elements of a „state space" (configuration space, phase space, Hilbert space).

(3) Time evolution: S(t') = U(t',t) S(t) is governed by a Hamiltionian H(q,p,t) or Lagrangian function or equations of motion for e.g. position states q(t) with initial states q(0) ... as initial values.

So the replacement of the presumably irredeemably muddy traditional concepts of cause and effect by powerful mathematics seems a very suggestive approach at first glance, but this claim apparently conflicts with descriptions that hold causation to be a relation between distinguishable, discrete events and the framework of event causation is used both in many less mathematicized special sciences and in most philosophical accounts of causation, including highly formalized approaches utilizing graph theory etc. If we take both approaches realistically as descriptions of the structure of the world a threat of a disunity of science and world arises, which would follow from a divergence in such a central concept as causation.

### *1.3 Options for Avoiding the Conflict*

There seem to be several options to avoid or mitigate the conflict. I will briefly look at the two most obvious ones: Reductionism and Pluralism. One could subscribe to hard physicalist reductionism and conclude that causal descriptions in the macroscopic realm and the special sciences should not be taken ontologically serious, but only as some kind of approximate description. Event causation is not really real, neither are the entities of higher levels. Both are epiphenomena, supervening on the basic physical entities and everything could in principle be reduced to fundamental levels without event causation. This is a viable position, although the attempts of reductionism in the last decades do not bode all that well and major philosophical accounts of reduction or supervenience are actually spelled out in the event causation approach. Another point is that there is no exact alignment between ontic levels and the possibility of a description in terms of continuous time evolution. Microscopic bacteria are not described by a mathematical theory, but mesoscopic steam engines and macroscopic solar systems are.

The other option is some kind of ontological pluralism along the lines proposed by Nancy Cartwright: The world is “dappled”; its regions and levels are strongly independent with respective “regional” ontologies and laws (Cartwright 1999). Accordingly, a similar pluralism could hold with respect to the causal relation and we could have event causation in some regions and continuous time evolution in others. But even with such a notion of strongly independent levels of reality it would seem really odd to have wildly dissimilar causal relations on different (vertical) levels (or no causation worth speaking of at all on some levels, but not on others). This seems especially problematic if causation is taken to be the “glue” between somewhat self-consistent, (horizontally) independent patches of reality. Especially for a “dappled world” like Cartwright's not to fall apart, we need a non-domain-specific account of causation as the relation that stitches the patches together. If one takes domains only vertically as levels dependent on the respective lower levels one could use relations of “constitution” between them that might be indepedent of causation. But the thorny problems of supervenience and the possibility of downward/upward causation need not concern me here. In any case it seems clear that accepting fundamentally different notions of causation in different domains or levels of reality will be a problem for a unified scientific world view.

So I will try a different approach beyond reductionism or pluralism. To this end, it is important to note that it is far from obvious what the ontological implications of the standard mathematical description should be and there is the chance of reconciliation of the conflicting views. Which relation do models like a high-dimensional phase space bear to reality which takes place in ordinary spacetime? Do all terms of the formalism correspond to entities of the world and to which kinds of entities? My aim is to show that a picture of the world in which causal powers or dispositions play a central part and are the basis of causation and laws of nature (defended by e.g. Ellis/Lierse 1994, Molnar 2003, Bird 2007) is able to accommodate both the view of mathematical physics and the one of manifestly causal descriptions. For this purpose I will offer one general argument examining the presuppositions of the continous time evolution approach and mention two examples from classical and quantum physics for the central role of causal dispositions in these fields. But first it will be shown how the event causation approach can be spelled out in terms of dispositions and their manifestations.

## 2. Dispositions and Event-causation

### *2.1 Dispositions, Manifestations, Triggers*

Without going into details I will first state the main points of my conception of dispositions. Compared with accounts in the literature it is probably closest to the one defended by Brian Ellis in (Ellis/Lierse 1994) and (Ellis 2001), but I take it to be also compatible with the slightly

different conceptions of dispositions suggested by Mumford (Mumford 1998), Molnar (Molnar 2003), Bird (Bird 2007) and others. Dispositions are properties, that is they are real features (tropes or universals) of their bearers (the things that have the dispositions). Their most important feature is that they are essentially linked to events or processes which are their manifestations: fragile → breaking, inflammable → inflammation, and so on. The nature of this link is controversial as a disposition may be present without ever being manifested, so it is not a normal ontological relation (because one relatum may not be existent) but I will take it as given for now. The manifestation event of a disposition (usually) takes place upon a "trigger" of the disposition and additional conditions have to be accounted for. (There are also untriggered dispositions like the tendency for the decay of a radioactive nucleus.) The bearer of a disposition is (usually) involved in the manifestation process. As mentioned, I will ignore the fact that not all dispositionalists use an ontology of events or processes (like Ellis 2001), but take as manifestations of dispositions instead (dispositional or non-dispositional) properties (Mumford 1998, Bird 2007). These approaches seem to be compatible to the one sketched here as long as one can give some account of the relation of the manifested properties to the manifestation process. Clearly, not all dispositions have to have processes as their manifestations. If we allow second order powers the manifestation of a disposition will be another disposition, e.g. the manifestation of the disposition to be capable to learn Finnish is to be able to speak Finnish which in turn has the speaking of this language as manifestation.

### *2.2 A Dispositional Account of Event-causation*

The event-causation view seems to be the standard view in analytic philosophy (Davidson 1967). Regardless of the way the causation relation is to be understood (constant conjunction, counterfactual dependence, transmission of a mark or of a conserved quantity), the causal relata can usually be conceived of as events. For simplicity I will ignore alternatives that use "facts" or propositions instead of events proper as causal relata. As long as "facts" are taken to be immanent entities, that is as located in space and time, event causation and fact causation seem to be sufficiently close to each other and both sufficiently different from continous time evolution that the finer distinctions may be ignored for the present purpose. Another distinction I will disregard is the one between events and processes where the former have no relevant temporal substructure whereas the latter are temporally extended. I will not distinguish between events and processes and use the expressions synonymously. Event causation is not difficult to connect with a metaphysics of dispositions like the one sketched above.

Briefly, the analysis of event causation in terms of dispositions and manifestations looks like this: Dispositions are dispositions for events or processes which are their manifestations, that is they are properties essentially linked to these manifestations. This manifestation of a disposition is an event or process that takes place upon a "trigger" (the main manifestation condition) of the disposition, and possibly additional conditions and several dispositions acting together have to be accounted for. In an event-causal description we would say:

The striking of the match (cause event) caused the match's burning (effect event).

or expressed as an ontological relation:

**causes** (c,e)

This can be translated into a dispositional account of causation as follows: We classify the "cause event" (the striking) as a trigger and the "effect event" (the burning) as the manifestation of the disposition of inflammability inhering in the match. An additional condition for the manifestation is the presence of oxygene. Thus the causal relation is analysed into a somewhat more complex one involving the disposition-manifestation link, the

triggering event and possibly additional conditions. Instead of **causes** (c,e) we have a relation (disposition, manifestation, trigger, additional conditions) relating a dispositional property, two events and whatever category conditions may belong to. This might be analysed further in terms of several two-place relations like **has_manifestation**(d,m) and **has_trigger**(d, t) (cf. Röhl/Jansen 2011 for some more formal considerations). Therefore this is not reductive in a usual sense, but more explicit about the causal factors being at work. One advantage of the dispositional approach is that the features (the dispositional properties) of the things involved in the events responsible for the causal connection are made explicit instead of relating cause and effect immediately.

## 3. Presuppositions of "Non-causal" Mathematical Descriptions

Now let us have a look at the presumably "cause-free" viewpoint. My central argument turns to the presuppositions of the apparently non-causal mathematical description. I think the very possibility of the "cause-free" description seems to rest on (at least) two presuppositions that employ causal notions:

(1) **Closure**: This means that all relevant causal features of the environment of interest for the development of the situation have been identified and taken into account and the system can (at least in principle) be causally isolated from other potential causal influences. (Or that these influences can be captured by a few control parameters to be used as inputs in the formalism.) This is often somewhat hidden in the formalism, because in a mechanical system contraints like the restriction of the movement of a ball to the surface of a bowl that reduce the degrees of freedom are used to find sets of coordinates that express only the remaining degrees of freedom, so the constraints do not show up explicitly.

It has to be noted, though, that the condition of closure could be spelled out in terms of purely functional mathematical relations between the system and its environment (or rather in terms of the *lack of* certain functional relations which show the isolation of the system). I do not claim the the event-causal view is presupposed, but it is certainly plausible that some causal structures correspond to the functional dependencies. And as it is well known functional dependencies are not the same as causal ones.

(2) **Internal dynamics**: We are often told that the "physics" of a dynamic system is contained in the Hamiltonian of a system. The Hamiltonian is the function that determines the equations of motion of a system or, in Quantum Mechanics, the time evolution operator U(t, t').

There are several options how to understand this: One could opt for an instrumentalist reading of the formalism taking it just as a technical machinery to derive predictions about the (spacetime) trajectories. This is possible, but uninteresting given the realist assumptions that lead to the conflict between event causation and continous time evolution in the first place. Or one could employ some kind of holism according to which the determination of the system by its Hamiltonian cannot be resolved any further: There are states of the system and the regular time evolution between them, but the H(q,p,t) that determines this evolution is ontologically not analysable. But in this case we would probably like to know what ontic structures correspond to a Hamiltonian function "out there". Looking at the practice of physical theorizing this seems not very plausible, as Hamiltonians are usually not conjured up out of nowhere or simply determined by data. Rather, when building Hamiltonians for complex systems we think about the fundamental properties of the component systems relevant for the interactions, that is their dynamic influences on each other. But "interaction", "coupling constant" are all causal concepts. Therefore, and that is the option that seems more fruitful to me, the ontic basis of a mathematical description can and should be analysed in causal language.

One could point out that “interaction”, “coupling constant” etc. are a mere way of speaking which does not imply causal connections in the philosophically “loaded” sense at all and that all physical theorizing and calculations are perfectly compatible with an eliminativist position (like Russell's). But again, I do not claim that event causation is presupposed, only very general causal notions that can be connected with an ontology of dispositions. To deny this seems to beg the question against any causal-ontological interpretation of the formalism.

### *3.1 A Possible Objection: Fundamental Symmetries*

Sometimes in fundamental physics a Lagrangian can be suggested by mere symmetry considerations. It has to respect certain symmetries (like Lorentz covariance) and we pick the most simple form that does everything we need it to do. So again, a formal description would be sufficient and no recourse to explicitly causal conceptions necessary. Furthermore, one could object that in fundamental physics the dynamical properties (like charges) are connected with internal (global and local gauge) symmetries by Noether's theorem.

Against the first point it can be said that symmetry and simplicity are in a way heuristic principles, not material ones. Generally, one can argue that regardless of the representation of e.g. charge as conserved quantity of an internal phase transformation, the essence of charge can only be captured by its being a specific causal power. The connection between symmetries and conserved quantities is fundamental, but it would be misleading to say that charge is “nothing but” a parameter of a gauge symmetry. A similar point has been made recently by Michael Esfeld, namely that our experimental access to reality is dependent on the dynamical coupling of microsystems to phenomena (Esfeld 2008). But for the purpose of the coupling of an electron to experimental phenomena it doesn't help very much that its charge can be connected to a gauge symmetry. So clearly, symmetries cannot replace causal powers. I will now try to show how the continous time evolution approach can be connected to the dispositional account of causation.

## 4. Dispositions and Continous Time Evolution

### *4.1 States and Processes*

First it needs to be shown how the category of “states” employed by the continous time evolution approach fits into the ontological model sketched above. This could and should be elucidated further, but for lack of space I take a “deflationary” view of both states and processes for the moment: To ascribe a state S(t) to a particular thing X means just that X instantiates a certain determinate value of the (determinable) property S at time t. And that a thing X participates in a process during the time interval (t, t') means that it continously exhibits subsequent values of S at all the intermediate time points, e.g. when it moves through a spatial region during that time interval it takes continuous values of position. So a process can be represented by $[S]_{t}^{t'}$, the sequence of S-values during (t,t').(This should not be taken to imply that states and processes can be ontologically *reduced* to things and properties, only that their relations can be described in such a fashion for our purpose.)

A further source of confusion is that what is called a *state* in physics might be described as a *process* in ontology. In classical mechanics a body may be described as being in a *state* of uniform rectilinear motion, despite being in motion and changing its position, because such a state is the “default state” of inertial motion without the action of external forces upon the body. The relevant parameter here is the velocity v(t) which is constant (both in magnitude and direction) and in this fashion the state of motion can be described by one parameter value only, rather than a sequence of values as sketched above. However, the interesting motions involving forces will always have changing values of velocity (either in magnitude or

direction), because Newton's second law connects forces acting upon a body with changes of its state of motion (characterized by velocity).

### *4.2 A Dispositional Account of Continous Time Evolution*

One could now be tempted to suggest that state S(t) causes the temporally subsequent state S(t+dt) in analogy with event-causation, because in the deterministic case S(t) determines S(t+dt). Just take S(t) as cause "event" and S(t+dt) as the effect "event". But there are at least two serious problems with such an approach. It would face Russell's objections that there is no next point in a continuous trajectory, no state following immediately in time. The continuous processes of classical physics have no obvious structure of causally linked separate events. This follows almost trivially from the continuity of a classical trajectory. And it would not agree with the dispositional account, because dispositions would have no role at all. A moving football's being at point x at t does not cause its being at x+dx at t+dt. Instead we have to look for the dispositions of the ball relevant for its movement. The ball's movement is caused by a combination of its inertial disposition to keep its initial movement (the velocity received by kicking it) and its acceleration towards the earth's center caused by the gravitational masses of the ball and the earth. All this takes place in accordance with the law of gravitation as this is based on the respective dispositions (cf. Bird 2007 for an account of laws based on dispositions).

So with the dispositional model we can give the following causal-ontological acount for such a movement or any similar continous physical process $S(t \rightarrow t')$: S is the joint manifestation process of the dispositions $D_1$, $D_2$,... of the system's constituents. The respective distances of the constituents as well as contraints are treated as additional manifestation conditions. In this fashion, the continous state space evolution can in principle be analysed in terms of causal properties, because these determine the functions or operators that determine the time evolution. Real causal dispositions are what drives a system, not some mathematical or abstract entity.

A similar approach has been pursued by Andreas Hüttemann (Hüttemann 2013).[1] Hüttemann's disposition-based process theory combines the approach of process theories of causation as proposed by (Salmon 1998) or (Dowe 2001), although with considerable differences to these authors. Hüttemann takes "quasi-inertial" processes like the inertial movement in classical mechanics as the default manifestation processes of dispositions and describes as causes in a more traditional sense only impeding factors ("antidotes") that perturb the default process, thus leading to a perturbed incomplete manifestation of the disposition. According to this approach, dispositions are not causes, but "contributing" factors. Without going into detail I want to stress that in my account dispositions are the central causal factors, but obviously no causes in the event causation sense as they are no events, but properties. Of course, I take dispositions to be more fundamental than cause events which presumably agrees with Hüttemann's characterization of traditional causes as "antidotes" (because antidotes are conceptually secondary to the disposition manifestations they are antidotes *for*).

### *4.3 Causation Involving Classical Forces*

As shown above in the case of a football, also Russell's example, the movement of a planet in its orbit, can be described as the time evolution of the system sun - planet as a process that is

[1] I became acquainted with Hüttemann's approach only after my presentation of this paper at the GAP 8 in September 2012. Professor Hüttemann kindly provided me with a pre-print of his 2013 paper, but a thorough discussion of the similarities and differences of his very interesting conception to my independently developed thoughts is beyond the scope of the present paper, so I will restrict myself to a few remarks.

the joint manifestation of the graviational mass dispositions of both bodies. Alternatively, one can introduce forces as intermediate entities in the causation complex. A realism with respect to classical forces has been convincingly argued for by Jessica Wilson (Wilson 2007). Causation by forces can be integrated into a dispositional framework. Forces are treated as a special type of dispositions which have accelerations as their manifestations. So here I depart from the view that dispositions are directly connected to processes; forces are intermediaries between the causal powers of things and the accelerations which characterize the processes these things undergo.

I suggest to understand the relation between the causal power of an entity and the executed forces as a relation of disposition and manifestation. The gravitation disposition $D_1$ of a heavy body $K_1$ is to act by force on a second gravitating body according to Newton's law.

$F(K_1,K_2)$ acting on a second body (patient) $K_2$ is a manifestation of the disposition $D_1$ of body $K_1$

$F(K_1,K_2)$ is a manifestation of a disposition of $K_1$, but the force itself has dispositional character as well. It is a disposition of the second body $K_2$ for a change of its motion (acceleration). The manifestation of the acceleration is a change of the body's motion, that is a process or a change in a process parameter.

Schematically: $D_1(K_1) \rightarrow M_1(K_2) = D_2(K_2) \rightarrow M_2$

As we could give the same description from the other bodies' perspective a (central) force is a joint manifestation of mutually dependent dispositions of two bodies. And the manifestation (acceleration) of a force-disposition of the body $K_2$ is dependent on all other forces acting on this body. In this way we can also describe forces in equilibrium where no motion takes place, although forces are active.

### *4.4 Quantum Propensities*

Finally, I consider briefly the case of propensities in quantum mechanics (QM), an approach originally suggested by Popper (Popper 1959). In QM we find a similar divergence of two apparently conflicting descriptions, one in terms of a systems' continuous time evolution and one in terms of events of measurement and their probabilities, and this seems to lie at the heart of the notorious interpretational difficulties. On the one hand the "Schrödinger" time evolution of quantum states according to Schrödinger's equation is continuous and deterministic. On the other hand the "Von Neumann" state change is discontinuous, probabilistic and seems to involve the collapse of the developed state into an eigenstate upon the triggering of a measurement-like event. This is of course a highly contested terrain, but one can interpret the quantum superposition state as exhibiting stochastic dispositions, "propensities" for the possible values of measurement results. The trigger for this projection onto an eigenstate and the manifestation of a definite value would be the causal influence of the measuring device. In the decoherence approach (cf. Giulini et al. 1996) the apparent collapse is due to the interaction with the environment. This could be taken as the deeper analysis of a propensity model of the collapse, so in this case propensities would probably not correspond to fundamental, irreducible features. Or one could subscribe to the Ghirardi/Rimini/Weber model (Ghirardi/Rimini/Weber 1986) with a modified dynamics that leads to untriggered, spontaneous collapses of the superposition state which are real, not merely apparent. In this case quantum propensities are fundamental dispositions without the necessity of a trigger. While this is an ongoing debate, it seems that propensities can capture important aspects of the ontology of quantum mechanics, and are considered as a serious option for an ontology of quantum field theory (cf. Kuhlmann 2010).

## 5. Conclusion

The conflict between event causation and continous time evolution is often ignored, but if we take causation ontologically serious we should take the conflict seriously and investigate options how to resolve it. Dispositionalists should not tie their accounts to event causation if they want their conceptions to be applicable to fundamental physics. I argued that both the conception of causation in terms of discrete events classified as cause and effect and the viewpoint of mathematical physics that seems to avoid causal concepts in favor of the mathematically described continuous evolution of states of physical systems can be accommodated by a model of causation in terms of manifestation processes of dispositional properties exhibited by the material things that interact with each other. Both approaches have to be revised slightly for this purpose. The causal link between events is not taken to be fundamental, but itself based on a more complex relation of dispositions, their manifestation conditions and their manifestations. The state space approach was shown to implicitly use causal concepts, both to isolate a closed system and to identify the relevant internal factors for the mathematical descriptions. These factors are dispositional properties of the system's components both in fundamental physics and elsewhere. An ontology with dispositions as relevant causal factors allows a unified conception of causation for both continuous processes and links of discrete events.

**Johannes Röhl**
Institut für Philosophie
Universität Rostock
18051 Rostock
johannes.roehl@uni-rostock.de

## References

Bird, A. 2007: *Nature's Metaphysics. Laws and Properties.* Oxford: Oxford University Press.

Cartwright, N. 1999: *The dappled world.* Cambridge: Cambridge University Press.

Davidson, D. 1967: 'Causal relations', *Journal of Philosophy* 64, 691–703.

Dowe, P. 2001: *Physical Causation.* Cambridge: Cambridge University Press.

Earman, J. 1986: *A primer on determinism*. Dordrecht: Reidel.

Esfeld, M. 2008: *Naturphilosophie als Metaphysik der Natur.* Frankfurt/Main: Suhrkamp.

Ellis, B./Lierse, C. 1994: 'Dispositional Essentialism', *Australasian Journal of Philosophy* 72, 27-45.

Ellis, B. 2001: *Scientific Essentialism*. Cambridge: Cambridge University Press.

Giulini, D. et al. 1996: *Decoherence and the appearance of a Classical World in Quantum Theory*. Berlin: Springer.

Ghirardi, G.C., Rimini, A., Weber, T. 1986: 'Unified dynamics for microscopic and macroscopic systems', *Physical Review D* 34: 470.

Hüttemann, A. 2013: 'A disposition-based process theory of causation' (to appear) in S. Mumford and M. Tugby (eds.): *Metaphysics of Science*, Oxford: Oxford University Press, 109-139.

Kuhlmann, M. 2010: *The Ultimate Constituents of the Material World. In Search of an Ontology for Fundamental Physics.* Ontos Verlag: Heusenstamm.

Molnar, G. 2003: *Powers. A study in metaphysics.* Oxford: Oxford University Press.

Mumford, S. 1998: *Dispositions.* Oxford: Oxford University Press.

Popper, K.R. 1959: 'The propensity interpretation of probability', *British Journal for the Philosophy of Science*, Vol.10, No.37, 25-42.

Röhl, J. and Jansen, L. 2011: 'Representing Dispositions', *Journal of Biomedical Semantics* 2011, 2(Suppl 4):S4 <http://www.jbiomedsem.com/content/2/S4/S4>

Russell, B. 1913: 'On the notion of cause', in *Logical and Philosophical Papers 1909-1913 (Collected Papers of Bertrand Russell, Vol. 6)*. London: Routledge 1992, 190-210.

Salmon, W. 1998: *Causality and Explanation*. New York: Oxford University Press.

Wilson, J. 2007: 'Newtonian Forces', *British Journal for the Philosophy of Science*, 58, 173-205.

# Between Relativism and Absolutism? – The Failure of Kuhn's Moderate Relativism[1]

Markus Seidel

In this paper I argue that a moderate form of epistemic relativism that is inspired by the work of Thomas Kuhn fails. First of all, it is shown that there is evidence to the effect that Kuhn already in his *The Structure of Scientific Revolutions* proposes moderate relativism. Second, it is argued that moderate relativism is confronted with a severe dilemma that follows from Kuhn's own argument for his relativistic conclusion. By focusing on the work of moderate relativists like Bernd Schofer and Gerald Doppelt this dilemma as well as the ultimate failure of Kuhn's moderate relativism are exhibited.

## 1. Introductory Remarks

The question of the potential relativistic implications of Thomas Kuhn's philosophy of science has been one of the key questions in the aftermath of the publication of *The Structure of Scientific Revolutions* (SSR) and continues to be the focus of much debate.[2] Some authors have defended the Kuhnian account by maintaining that it does not imply an *extreme* but a *moderate* form of relativism.[3] The basic idea of this defence can be traced back to Kuhn's later work: Kuhn argued that though there are no paradigm-transcendent standards of evaluation in theory-choice, there are transparadigmatic *values* that are shared, but weighed differently by competing scientists adhering to different paradigms.[4] Therefore, Kuhn believes, scientific change in scientific revolutions is not *wholly* irrational since science *as a whole* is a highly rational enterprise.[5] Furthermore, in single situations of theory-choice good reasons play a decisive role, but they do not determine theory-choice.[6] Consequently, so it is argued, Kuhn does not deny the rationality of scientific change and proposes instead so-called "Kuhn-underdetermination",[7] - and in doing so, just advances a *moderate* form of relativism. Thus, the basic idea of moderate relativism is that the evaluation of theories is co-determined by the facts of the world, social factors *and* good reasons.[8]

In this paper I will argue that a) Kuhn's position in SSR can in fact be interpreted to constitute a form of moderate relativism, and b) that an application of Kuhn's own argument in SSR leads to a dilemma for proponents of moderate relativism like himself, Gerald Doppelt and Bernd Schofer.[9] My conclusion is that there is no plausible moderate position between relativism and absolutism.[10]

---

[1] I would like to thank Julia Göhner for her helpful comments on earlier versions of this paper.

[2] See e.g. Bird 2011, Sankey 2012a, Wray 2011: 164-168.

[3] See e.g. Doppelt 1982, Schofer 1999.

[4] See especially Kuhn 1977 and also Kuhn 1970c: 184f.

[5] See Kuhn 1970a: 143f.

[6] See Kuhn 1970b: 261.

[7] See Carrier 2008, Schofer 1999: 23, Wray 2011: 161f.

[8] See especially Doppelt 1983: 111, Doppelt 1986: 240f, Schofer 1999: 23.

[9] Subsuming these authors under the heading "moderate relativism", I do not intend to deny the differences between them. Most importantly, Kuhn himself never described his position by using this

The major problem of many discussions about relativism is that participants remain unclear as to what exactly relativism is supposed to be. To my mind, the same can be said with respect to the discussion about relativism in Kuhn's work, because Kuhn's use of the label "relativism" is ambivalent. Sometimes Kuhn seems to believe that the question of relativism is intimately connected to the realism/anti-realism debate in philosophy of science and the related question of whether truth should be understood in terms of correspondence.[11] Sometimes, however, Kuhn discusses relativism as a position concerning situations of rational theory-choice.[12]

In what follows, I will only be concerned with the latter form of relativism and not the former: whatever the fate of realism and the correspondence-theory of truth, the question of an authentic form of *epistemic* relativism remains a focus of special interest both in recent discussions in epistemology and philosophy of science[13] as well as with respect to a proper understanding of Kuhn's position.[14] Henceforth, I will speak of epistemic relativism as the position that the epistemic evaluation – i.e. an evaluation using terms like "reasonable", "rational", "justified" and the like – of a proposition is possible only relative to a variable and local set of epistemic standards or norms.[15]

## 2. Kuhn's Moderate Relativism in *SSR*

Let us start with the well-known criticism of Thomas Kuhn provided by Imre Lakatos. Lakatos' much quoted dictum that "in Kuhn's view scientific revolution is irrational, a matter of mob-psychology" (Lakatos 1970: 178, italics omitted) is sustained by his diagnosis that for Kuhn "[there] are no rational standards for theory comparison. Each paradigm contains its own standards." (Lakatos 1970: 178). Whether or not Lakatos' attack succeeds, some of Kuhn's statements surely speak in favour of Lakatos' diagnosis. Thus, Kuhn claims that paradigms "are the source of the methods, problem-field, and standards of solution accepted by any mature scientific community at any given time" (Kuhn 1970c: 103). Therefore, according to Kuhn, "when paradigms change, there are usually significant shifts in the criteria determining the legitimacy both of problems and of proposed solutions." (Kuhn 1970c: 109). Finally, to give the quote that most critics have taken to testify that Kuhn is an epistemic relativist,

> [as] in political revolutions, so in paradigm choice – there is no standard higher than the assent of the relevant community. To discover how scientific revolutions are effected, we shall therefore have to examine not only the impact of nature and of logic,

---

label. I would like to thank Paul Hoyningen-Huene for pointing this out in his comments to my talk at GAP.8.

[10] Just for the record: my own position is an epistemic absolutist one. Though I do not think that there is a plausible, moderate form of relativism, I believe that epistemic absolutism can very well integrate the basic insights and intuitions of the epistemic relativist. My account – argued for in my PhD-thesis (see Seidel forthcoming) – is inspired by Alvin Goldman's recent proposal in the debate about epistemic relativism (see Goldman 2010).

[11] See e.g. Kuhn 1970c: 205f, Kuhn 2000: 243f.

[12] See e.g. Kuhn 1970b: 259.

[13] See e.g. Goldman 2010, Kusch 2010, Pritchard 2011, Sankey 2011a, Sankey 2012b.

[14] See e.g. Sankey 2012a.

[15] Thus, e.g. the proponents of so-called "Edinburgh relativism" or the "Strong Programme" count as epistemic relativists in this sense when they claim that the relativist "accepts that none of the justifications of his preferences can be formulated in absolute or context-independent terms" (Barnes/Bloor 1982: 27), that "[f]or the relativist there is no sense attached to the idea that some standards are really rational as distinct from merely locally accepted as such" (ibid.) and that "there are no context-free or super-cultural norms of rationality" (ibid.).

> but also the techniques of persuasive argumentation effective within the quite special groups that constitute the community of scientists. (Kuhn 1970c: 94)[16]

Now, even if we accept Kuhn's later statement that the talk of 'persuasive argumentation' here is not meant to suggest that in paradigm or theory choice there aren't any good reasons to adopt one theory or the other – Kuhn alludes to his trans-paradigmatically applied 'Big Five' here –,[17] we nevertheless want to know why Kuhn thinks that 'there is no standard higher than the assent of the relevant community'.

In a series of recent papers, Howard Sankey convincingly argues that Kuhn's argument in these passages resembles the classical Pyrrhonian sceptic's argument from the criterion:[18] in order to justify a belief we need to appeal to a criterion or standard of justification. How, however, is this standard itself justified? The Pyrrhonian argues that all available options with respect to this question fail to provide an epistemic justification of the criterion: a) appeal to another criterion inevitably leads to an infinite regress, b) appeal to the original criterion results in circular justification, and c) adoption of the criterion on a dogmatic basis leaves the criterion unjustified. Whereas the Pyrrhonian sceptic's conclusion of the argument from the criterion is suspension of belief, the epistemic relativist uses the argument to show that epistemic justification can be justification only relative to a set of epistemic standards operative in a specific context.[19]

As I have argued elsewhere, I do not think that the epistemic relativist can use the sceptical strategy as suggested by Sankey: either the epistemic relativist has to bite the bullet of the argument from the criterion that we can have *no* – relative or absolute – epistemic justification, or her use of the argument does nothing to establish epistemic relativism at all.[20] Nevertheless, I completely agree with Sankey's interpretative result that *in fact* Kuhn uses an argument that closely resembles the argument from the criterion in arguing for his thesis that "there is no standard higher than the assent of the relevant community" (Kuhn 1970c, 94). Here is the passage that precedes this quote:

> Like the choice between competing political institutions, that between competing paradigms proves to be a choice between incompatible modes of community life. Because it has that character, the choice is not and cannot be determined merely by the evaluative procedures characteristic of normal science, for these depend in part upon a particular paradigm, and that paradigm is at issue. When paradigms enter, as they must, into a debate about paradigm choice, their role is necessarily circular. Each group uses its own paradigm to argue in that paradigm's defense. [...] [T]he status of the circular argument is only that of persuasion. It cannot be made logically or even probabilistically compelling for those who refuse to step into the circle. The premises and values shared by the two parties to a debate over paradigms are not sufficiently extensive for that. As in political revolutions, so in paradigm choice – there is no standard higher than the assent of the relevant community. (Kuhn 1970c: 94)

The argument in this passage centers on Kuhn's claim that the debate about paradigm choice has a necessarily circular character. And, so Kuhn goes on, this characteristic of the debate leads to a restriction of the power to convince the opponent: any argument cannot be made logically or probabilistically *compelling* but can only *persuade* the advocate of another paradigm. In this respect, Kuhn's argument strongly resembles the argument from the

---

[16] This is the passage that David Bloor probably has in mind when he says that "[t]here is (in Kuhn's words) no higher court of appeal than the community of acknowledged experts." (Bloor 2011: 441). I will not dwell on the difference between Kuhn's words and what Bloor takes to be Kuhn's words, however.

[17] See Kuhn 1970b: 261.

[18] See Sankey 2011a, Sankey 2012a, Sankey 2012b.

[19] See Sankey 2012b: 187.

[20] See Seidel 2013a, Seidel 2013b. See also Sankey 2013, which is a reply to the former paper

criterion: the debate about paradigm choice is necessarily circular since otherwise the opposing parties get tangled up in a looming infinite regress of justification. However, the circular justification of a paradigm will not convince the opponent; in the end, the adoption of a paradigm must rely on its bare, ultimately unjustified and dogmatic assumption. As Sankey remarks quite correctly, "to say that there is no higher standard than the assent of the scientists who adopt a paradigm is to say that there is no further justification that may be provided" (Sankey 2011a: 566).[21]

Now, although Kuhn's argument can justifiably be interpreted along the lines of the argument from the criterion, it is equally possible to interpret it as an indicator for a moderate form of relativism. This is most obvious in Kuhn's later statements on the issue, where he insists that the term "persuasion" should not be seen to indicate "that there are not many good reasons for choosing one theory rather than another" (Kuhn 1970b: 261). Therefore, some authors argue, whereas the Kuhn of SSR can safely be interpreted to announce epistemic relativism,[22] in later work Kuhn moved away from this position by invoking shared values in theory choice. But said values, the idea goes, do not provide reason to think that there is an *algorithm* of theory choice:[23] since there can be disagreement about how to weigh and how to interpret the shared values, "[t]here is no neutral algorithm for theory-choice, no systematic decision procedure which, properly applied, must lead each individual in the group to the same decision" (Kuhn 1970c: 200).

With respect to the question of epistemic relativism, I do not think that it is correct to distinguish between a more extreme, earlier Kuhn and a more moderate, later Kuhn.[24] To my mind, even the Kuhn of SSR aims at a moderate form of epistemic relativism.[25] Once this is accepted, it immediately becomes evident that there is a crucial tension in Kuhn's moderate relativism: providing an argument that undermines the invocation of good reasons in theory choice and simultaneously insisting on the co-determining power of good reasons in theory choice appears to be incoherent.

In order to see that already the Kuhn of SSR is a proponent of moderate relativism note again that Kuhn in his circularity-argument contrasts *persuasion* and *logical and probabilistic compellingness*. Can we interpret Kuhn in SSR to mean that there is no neutral algorithm for theory choice but that persuasive argumentation does not imply the absence of good reasons? Note that even there, Kuhn contrasts *techniques of persuasion* with *proofs*. Thus, he claims "that paradigm change cannot be justified by proof, is not to say that no arguments are relevant or that scientists cannot be persuaded to change their minds" (Kuhn 1970c: 152) and speaks "about techniques of persuasion, or about argument and counterargument in a situation in which there can be no proof" (Kuhn 1970c: 152). It is obvious, therefore, that already in SSR Kuhn wants to maintain that persuasion does not imply *absence* of good reasons; though there is no proof the scientists nevertheless are in a situation where there are *arguments* and *counterarguments*.[26] Therefore, I propose that Kuhn's statement that "the status of the circular argument is only that of persuasion" (Kuhn 1970c: 94) should not be

[21] It should be noted that also Williams sees a close connection between the argument from the criterion and the "fundamental argument for epistemic relativism" (Williams 2007: 94). In fact, the argument provided here by Kuhn can be found also e.g. in Wittgenstein's *On Certainty* and has been dubbed by Paul Boghossian the "argument from norm-circularity" (Boghossian 2006: 95).

[22] See e.g. Bird 2000: 241.

[23] See e.g. Sankey 2011b: 468, Sankey 2012a.

[24] Note that this is true only with respect to the question of epistemic relativism. I do not want to maintain that there is no difference between the Kuhn of SSR and the later Kuhn with respect to other issues, e.g. his treatment of semantic incommensurability.

[25] By "SSR" in the phrase "the Kuhn of SSR" I refer to the first edition of SSR from 1962. Obviously, in the Postscript to SSR from 1969 there is no doubt that Kuhn invokes shared values (see Kuhn 1970c: 184f).

[26] See also Hoyningen-Huene 1993: 252f.

seen to suggest that there are no good reasons in theory choice; persuasion here is contrasted with logical and probabilistic compellingness and should not imply the absence of arguments. Most importantly, this interpretation is sustained by a close look at the passage from which the circularity-argument has been taken. Following his denial that the circular argument can be made compelling Kuhn explains that the *shared* premises and values are not extensive enough for that.[27] This statement can surely be interpreted to be an expression of Kuhn-underdetermination; namely that the evaluation of theories is underdetermined by the shared values of the opponents. Furthermore, note that in this passage Kuhn does not claim that the evaluation of paradigms depends on a paradigm *completely* – he says that the evaluative procedures depend on the paradigm *in part*. To my mind, Kuhn's more cautious formulation here again should point to the shared values of the proponents of opposing paradigms that are – besides the paradigm – *partly* responsible for the evaluation of theories.

Therefore, I conclude, there is evidence already in SSR that the earlier Kuhn wants to propose merely a moderate form of relativism. Most importantly, we find evidence for this interpretation already in the argument for the thesis that most commentators have taken to testify Kuhn's clear-cut epistemic relativism.

However, if this interpretation is correct, we immediately are confronted with a serious problem of Kuhn's position. On the one hand, Kuhn provides us with an argument that should convince us that there is no standard higher than the assent of the relevant community. This, in effect, is the argument from the criterion. If Kuhn's own argument is correct, then we are forced – it seems – to *embrace* extreme epistemic relativism. On the other hand, Kuhn wants to avoid the conclusion of his own argument by invoking shared normative constraints, namely shared values, which should prevent us from concluding that scientific change is *wholly* a matter of assent. However, Kuhn does not give us a clue as to how the conclusion of his own argument does not apply to these normative constraints, too. How does Kuhn plan to escape the argument from the criterion without buying its conclusion? If Kuhn really accepts his own circularity-argument, he owes us an explanation of why this circularity-argument should be applied to evaluative standards in theory-choice but not to transparadigmatic values. If we assume, however, that Kuhn somehow manages to make a case for this difference, we need a reason why his explanation for exempting transparadigmatic values from the force of the argument cannot be applied to the evaluative standards *themselves*. After all, Kuhn's circularity-argument should provide us with reasons to think that the evaluative procedures are *partly* dependent on the respective paradigm. Why, in case Kuhn can make a plausible exemption for the case of values such that these are transparadigmatic, should we believe that standards are nevertheless dependent in part on paradigms? Kuhn is confronted with a dilemma, therefore: If Kuhn really wants to rely on the argumentative force of his circularity-argument, it is difficult to avoid the conclusion of the *extreme* relativist. If Kuhn wants to restrict the argumentative force of the circularity-argument, it is unclear how he can do this and why we should not use this very restriction in order to exempt also the evaluative procedures in theory-choice from its argumentative force. Why, I ask, should we not simply remain epistemic *absolutists*?

In what follows, I will show that this dilemma can also be found in the work of authors who draw on Kuhn's work and argue explicitly for a *moderate* version of epistemic relativism. I will discuss the position of Bernd Schofer and Gerald Doppelt and show that since we find the dilemma not only in Kuhn, but also in their accounts, there is no plausible form of Kuhn-

[27] Usually, commentators of this passage overlook this point. See e.g. Holcomb 1987: 468f, Sankey 2011a.

inspired moderate relativism. Thus, there is no position between relativism and absolutism in the epistemic realm – both positions are mutually exclusive and they are the only positions.[28]

## 3. Moderate Relativism: Ambiguities, Unclarities, and Inconsistency

### *3.1 Schofer's Moderate Relativism*

In his study *Das Relativismusproblem in der neueren Wissenssoziologie*[29] Bernd Schofer aims to show that "the assumption of the possibility of moderate relativism" (Schofer 1999: 24 Fn. 24 and 192) is an assumption based on plausible philosophical assumptions. However, as I will show in this section, it is entirely unclear what 'moderate relativism' is supposed to be. To my mind, Schofer's definition of relativism and absolutism does not even touch the problem of epistemic relativism at all. Furthermore, his attempt to distinguish between different forms of absolutism and relativism mixes up very different issues and theses with the result that it is not quite clear in which way moderate relativism can provide a middle way at all.

The setting of Schofer's argument is to distinguish between, on the one hand, foundationalism and anti-foundationalism and, on the other hand, absolutism and relativism. Obviously, we should be especially concerned with his exposition of the latter.[30] Absolutism and relativism, according to Schofer, make divergent assumptions about whether socio-cultural factors influence the evaluation of claims to knowledge in an epistemologically relevant sense.[31] They differ, therefore, about whether the validity of claims to knowledge and theories is independent of the socio-cultural context of the evaluation of knowledge or whether said validity is context-dependent.[32] Schofer is especially concerned about the question of the rational acceptability of theories: "a theory is rationally acceptable if it is accepted and preferred to rival theories because of good reasons." (Schofer 1999: 15). Whether theories are acceptable relatively or absolutely depends thus on whether the theory is accepted independently of socio-cultural factors that influence the evaluation or not. What, according to Schofer, are good reasons? Schofer claims that

---

[28] In this respect, I will follow David Bloor: "Relativism and absolutism are mutually exclusive positions. They are also the only positions." (Bloor 2007: 252). However, in obvious contrast to Bloor, my sympathies are with the absolutist side of the divide.

[29] See Schofer 1999. Translated, the title reads 'The problem of relativism in the newer sociology of knowledge'. The following quotes from Schofer are all my own translations.

[30] Schofer's distinction between foundationalism and anti-foundationalism is confronted with the same problem that, as I will show, affects his distinction between absolutism and relativism: it mixes up many different theses. On Schofer's account, the foundationalist assumes that valid scientific knowledge is secure, proven knowledge (see Schofer 1999: 14, German: 'sicheres, bewiesenes Wissen'), that there is a privileged access to the world-in-itself (see Schofer 1999: 15), the existence of secure/certain sentences (see Schofer 1999: 16, German: 'sichere Sätze'), the existence of an unshakable foundation of knowledge (see Schofer 1999: 16), the existence of an archimedian point from which we can deduce ('ableiten') knowledge (see Schofer 1999: 16). Furthermore, in contrast, the anti-foundationalist is said to claim that "all claims to knowledge are insecure and fallible" (Schofer 1999: 14).

Just to point out one puzzling fact about Schofer's description of the foundationalist: his description of the differences between the foundationalist and the anti-foundationalist has the consequence that "epistemological realism in the sense of assuming the knowability of the thought- and subject-independent world can be said to be a form of foundationalism" (Schofer 1999: 16 Fn. 12). However, I do not see why an epistemological realist cannot be a fallibilist: why is it impossible to have epistemic access to the independently existing world and to assume at the same time that our claims to knowledge are fallible?

[31] See Schofer 1999: 14.

[32] See Schofer 1999: 14f.

> there are good reasons for the acceptance of a theory if in the relevant scientific community there is a consensus that the reasons fulfil the scientific standards for evaluating theories. (Schofer 1999: 15)

The difference between absolutism and relativism is understood as a difference "in the assumptions *how* the consensus about the presence of good reasons is attained [...] and how to conceive the acceptability of theories accordingly" (Schofer 1999: 15). In my opinion, these definitions of what relativism and absolutism are supposed to be are quite surprising. Thinking of good reasons in terms of the *consensus* of the relevant scientific community and claiming that absolutism and relativism are different because they give different answers to the question how this consensus is generated is – at least – an unusual idea: the epistemic absolutist will complain that good reasons are *not* to be understood in terms of the consensus of the scientific community. The question of relativism or absolutism seems to be better expressed by asking the question whether there are good reasons independently of whether any community has a consensus about them or not.

The result of Schofer's setting of the relativism-absolutism-debate is that his analysis remains conceptually unclear and question-begging against the absolutist. My intention is not to nitpick, but in this case it is crucial to see that Schofer makes so many claims about what is supposed to be co-determined by social factors that the consequence of his discussion remains unclear. Schofer speaks of the co-determining influence of social factors on the acceptance of scientific theories,[33] on the acceptability of scientific theories,[34] on the validity of scientific theories,[35] on the judgements of the acceptability by the scientists,[36] on the evaluation of the acceptability of scientific theories,[37] on the consensual judgement of the acceptability of scientific theories,[38] on the evaluation of scientific theories,[39] on the application of evaluative standards,[40] on the judgement of the correct interpretation and weighing of evaluative criteria,[41] on the consensual evaluation of scientific theories,[42] on the decisions of scientists for scientific theories,[43] on the choice of scientific theories,[44] on the generation and evaluation of scientific theories (or knowledge),[45] on the evaluation of a theory as acceptable,[46] on the judgements of preference by the scientists,[47] on the development of science,[48] on the implementation of the decision of a group for a scientific theory,[49] on the formation of a consensus,[50] on the formation of the consensus about the acceptability of scientific theories,[51] and on the stability of a consensus about the acceptability

[33] See Schofer 1999: 87.
[34] See Schofer 1999: 97, 128.
[35] See Schofer 1999: 14f.
[36] See Schofer 1999: 128.
[37] See Schofer 1999: 88.
[38] See Schofer 1999: 23, 172.
[39] See Schofer 1999: 23, 96, 128, 148.
[40] See Schofer 1999: 21, 23.
[41] See Schofer 1999: 23.
[42] See Schofer 1999: 24.
[43] See Schofer 1999: 95, 97.
[44] See Schofer 1999: 96.
[45] See Schofer 1999: 57, 100, 121, 264.
[46] See Schofer 1999: 124, 165.
[47] See Schofer 1999: 128.
[48] See Schofer 1999: 165.
[49] See Schofer 1999: 172.
[50] See Schofer 1999: 173.
[51] See Schofer 1999: 150.

of scientific theories.[52] Now, the epistemic absolutist will surely object to this list by saying it mixes up *normative* and *descriptive* questions. For example, no epistemic absolutist needs to deny that e.g. the formation of the consensus of scientists, the implementation of the decision of scientists, the stability of the consensus, the (actual) evaluation of a theory as acceptable and even the acceptance of theories is co-determined by social factors – modern science is *of course* a social enterprise such that scientists must learn which theories and methods are accepted and how to apply the methods. Furthermore, of course, modern science is a social enterprise in that the individual scientist must rely on the expertise of others and come to know about social mechanisms like scientific publication and networking. However, the epistemic absolutist will deny that all this implies that the acceptability of scientific theories, the (correct) evaluation of a theory as acceptable and the validity of scientific theories is relative to social factors. To my mind, Schofer's discussion simply does not answer the question of epistemic relativism, i.e. the question whether there are any standards of justification or epistemic norms that are absolutely correct or whether all such standards are correct only relatively.

Due to the fact that Schofer mixes up claims of normative and descriptive relativism his definition of moderate relativism is unclear with respect to what exactly distinguishes the moderate form from the extreme form. Schofer claims that

> *moderate relativism* thinks of the consensual judgement about the acceptability of theories as the product of the influences from the facts of the world, the evaluative standards *and* the social factors such that moderate relativism assigns a co-determining influence of all three kinds of factors on the evaluation. (Schofer 1999: 23)

The extreme relativist, on Schofer's account,

> denies a relevant influence of the evaluative standards on the evaluation of theories and assigns their – on this view just putative – influence to the social factors. (Schofer 1999: 24)[53]

Of course, the extreme relativist will wonder whether the evaluative standards invoked by Schofer's moderate relativism are absolute or relative standards – recall that the moderate relativist inspired by the Kuhnian account must provide us with reasons not to apply the argument from the criterion to the evaluative standards themselves. The problems of Schofer's account can be seen more clearly once we focus on his criteria for an adequate sociology of knowledge.

According to Schofer, his discussion reveals that a new sociology of knowledge must meet two criteria for the assessment of whether the problem of relativism is solved: the criterion of reflexivity and the criterion of moderate relativism.[54] The latter is, as we have already seen, the claim that – in order not to fall prey to a radical irrationalism[55] – the relativistic sociology of knowledge must accept that the world, the evaluative standards and the social factors play a co-determining role in the evaluation of knowledge.[56] The criterion of reflexivity is necessary in order to meet the absolutist's reproach of self-contradiction. Schofer thinks that "the anti-foundationalist and relativist stance must be applied also to the own claims to knowledge" (Schofer 1999: 184). This reflexive requirement must be adhered to consequently

---

[52] See Schofer 1999: 150.

[53] Again, Schofer's description is unclear: moderate relativism is described as insisting on the influence of all three factors on 'the consensual judgement about the acceptability' whereas extreme relativism is described to deny the influence of – at least – one factor on 'the evaluation of theories'. It is surely debatable whether the evaluation of theories and the consensual judgement about the acceptability are really the same phenomenon.

[54] See Schofer 1999: 184.

[55] See Schofer 1999: 238.

[56] See Schofer 1999: 192.

– otherwise the thesis falls prey to self-contradiction by claiming absolute validity for itself.[57] That is, Schofer claims, the reason why, for example, Karl Mannheim's theory failed.[58] However, according to Schofer, "the thesis is not self-contradictory if the general assumption of the relativity of validity is applied also to the thesis itself." (Schofer 1999: 185). Let us grant this for the sake of argument.[59] Note, however, that with this criterion Schofer's claims about moderate relativism become quite puzzling. The idea of moderate relativism is that "values and criteria do not uniquely determine the evaluation but guide it" (Schofer 1999: 193). Since sociologists of knowledge also propose scientific theories, Schofer applies the criterion of moderate relativism to the sociologists of knowledge, too:

> In the scientific community of the sociologists of knowledge there is no consensus about the competing theories, but the reflexive moderate relativist can claim that his interpretations of the individual values correspond to their intent and can be assessed as correct, i.e. that good reasons speak for his theory. Therefore, he can promote his theory by claiming its special, not socially reducible persuasive power and argue for its excellence against other theories. (Schofer 1999: 193)

I do not see how to combine the criterion of reflexivity and the criterion of moderate relativism on this account: on the one hand, the thesis of the moderate relativist itself is supposed to be merely relatively valid, on the other hand, it claims special, not socially reducible persuasive power for itself by invoking good reasons. Schofer's moderate relativism seems to want it both ways: being based on reasons that are valid and good only relatively but providing compelling, good reasons for others. However, to my mind, that is just not coherent.

In effect, Schofer's attempt to combine moderate relativism with the principle of reflexivity reiterates the dilemma already encountered in Kuhn's application of the argument from the criterion. The principle of reflexivity is a consequence of a continuous application of the argument from the criterion: there is no reason to suppose that the argumentative force of the argument halts at the reasons invoked to argue for one's own position. That, to be sure, is the reason why the original Pyrrhonian conclusion is *suspension* of judgement. Schofer's moderate relativist, however, wants to argue for his own theory by providing good reasons that have "special, not socially reducible persuasive power" (Schofer 1999: 193). How, we must ask, is that possible if we adhere to the principle of reflexivity *stringently*?

### *3.2 Doppelt's Moderate Relativism*

In a series of papers, Gerald Doppelt, drawing on the work of Thomas Kuhn, has presented his case for a moderate form of relativism. In contrast to Schofer, Doppelt is clearer about what is at stake in the discussion: it is the question of the variability of the *normative* commitments of epistemic communities.[60] His papers shift the focus in the debate about

---

[57] Obviously, Schofer's principle of reflexivity is based on the reflexivity-postulate of the Strong Programme (see Bloor 1991: 7).

[58] See Schofer 1999: 185. Contrary to what Schofer suggests, Mannheim had a principle of reflexivity and has followed it very consequently (see Seidel 2011a, Seidel 2011b).

Schofer also argues that – though Kuhn should be interpreted as a moderate relativist (see Schofer 1999: 175) – his attempt to deal with the problem of relativism fails because of his hesitation to apply the principle of reflexivity stringently (see Schofer 1999: 176-180). In what follows, I will argue that also Schofer's attempt to combine moderate relativism with the principle of reflexivity fails in the same way. The reason, I suggest, is that a combination of moderate relativism and a continuous application of the principle of reflexivity is inconsistent.

[59] Schofer uses Mary Hesse's objection to the reproach of the self-contradictory character of relativism (see Schofer 1999: 187f, see also Hesse 1980: 42f.). See for a critical discussion of Hesse's objection: Nola 1990: 288-291, Nola 2003: 289-293, Siegel 2004: 764.

[60] See Doppelt 1982: 138, Doppelt 2001: 160.

Kuhn's concept of incommensurability from the semantical questions of reference change, untranslatability and conceptual relativism to methodological incommensurability concerning shifts in standards, methods and problems in scientific change.[61]

Doppelt does not subscribe to the theses that many in the aftermath of Kuhn have adhered to – the 'Post-Kuhnians' as Doppelt calls them either cannot argue for their case that scientific development is rational or they fall on the side of Doppelt's own 'moderate relativism'.[62] Thus, "moderate relativism is the inescapable outcome of the post-Kuhnian dialectic of argument." (Doppelt 2001: 160). Despite this straightforward argument, it is quite complicated to make out what Doppelt's moderate relativism amounts to, as in his texts he uses at least six different labels for kinds of relativism he seems to endorse.[63]

Doppelt distinguishes between different forms of relativism along two dimensions. The first dimension is *temporal* – Doppelt distinguishes between forms of so-called *short-run* and *long-run* relativism.[64] The second dimension concerns the *strength* of relativism – Doppelt distinguishes between *extreme* and *moderate* relativism. A further distinction concerns the question what exactly is relative: is it scientific knowledge,[65] scientific rationality,[66] or scientific progress[67]? In the context of this paper, I want to focus on Doppelt's version of a moderate relativism concerning scientific rationality.[68]

Doppelt distinguishes between extreme and moderate relativism concerning scientific rationality in the following way:

> (a) *Extreme Relativism Concerning Scientific Rationality:* There can never be any good reasons for judging a new paradigm in science more rational to accept than its predecessor (rival). [...]

---

[61] See Doppelt 1982: 118: "On the interpretation of Kuhn's relativism to be developed here, it is the incommensurability of scientific problems between rival paradigms and not that of meanings which constitutes the most basic premise of the argument.". See also Doppelt 1983: 109.

[62] Doppelt especially attacks Dudley Shapere (see Doppelt 1988) and Larry Laudan (see Doppelt 1986) in this way. See also Doppelt 2001: 165-176.

[63] These are: 'Moderate relativism' (Doppelt 1986, Doppelt 1988: 110, Doppelt 2001), 'Sociological Relativism' (Doppelt 1986: 225, 241, Doppelt 1988: 110), 'Cognitive Relativism' (Doppelt 1988: 110), 'Short-run-relativism concerning scientific knowledge' (Doppelt 1982: 135), 'Short-run-relativism concerning scientific rationality' (Doppelt 1982: 135), 'Short-run moderate relativism' (Doppelt 1986: 248).

[64] See Doppelt 1982: 135, Doppelt 1983: 117-120, Doppelt 1986: 246. For the case of scientific rationality, the idea of short-run relativism is that even though *single changes* in science might appear irrational since they involve epistemic losses, these changes "in the longer run turn out to be cumulative" (Doppelt 1986:246). Doppelt obviously refers to the idea of what is known as 'Kuhn-losses' in this context (see e.g. Kuhn 1970c: 148f).

[65] See Doppelt 1982: 135.

[66] See Doppelt 1982: 135, Doppelt 1983: 111.

[67] See Doppelt 1983: 114ff.

[68] Doppelt himself is unclear about the relation between his distinctions along the temporal dimension and along the dimension of strength. On the one hand, he refers to his own paper Doppelt 1982 – in which he only speaks about the distinction between short-run and long-run relativism – and claims that there he has provided "the argument for moderate relativism" (Doppelt 1986: 251 Fn. 3). Thus, so we conclude, the short-run/long-run-distinction is understood to be intimately connected to the moderate/extreme-distinction. On the other hand, he inflicts Larry Laudan that he fails to distinguish "between a short-run and long-run moderate relativism" (Doppelt 1986: 248). This seems to suggest that the distinctions along the different dimensions are independent of each other.

Furthermore, I will just focus on moderate relativism concerning *scientific rationality* because it is this form that is, according to Doppelt, most relevant (see Doppelt 1983: 132). It is also the most relevant form in the context of this paper since it is supposed to be the most powerful form of relativism to be drawn from Kuhn's work (see Doppelt 1982: 137).

> (b) *Moderate Relativism Concerning Scientific Rationality:* There are always some good reasons for judging a new paradigm in science more rational to accept than its predecessor (rival) but such reasons are never (or in some cases, not) *more rationally compelling* or stronger than the good reasons which exist in favor of its predecessor. [...] (Doppelt 1983: 111)

The idea of Doppelt's distinction is familiar from our discussion of Kuhn: instead of maintaining that scientific change is a matter *wholly* irrational, the moderate relativist insists that evaluative commitments play a role in the evaluation of theories but they do not determine a particular choice. Thus, Doppelt maintains that "the thesis of moderate relativism is that scientific change is often or typically underdetermined by good reasons" (Doppelt 1986: 225).[69] Therefore, according to Doppelt, scientific change is explicable by referring to the good reasons of scientists – thus *moderate* relativism – *and* "an ineliminable sociological component" (Doppelt 1986: 225)[70] – thus moderate *relativism*. As in Schofer's case, moderate relativism is the position that the acceptability of theories is a product of the influences of social factors *and* evaluative commitments/good reasons.[71]

Alas, Doppelt's formulations of his moderate relativism remain ambiguous. First of all, it must be noted that Doppelt's conclusion that there is "an ineliminable sociological component" (Doppelt 1986: 225)[72] in the explanation of theory-change only follows if – as Doppelt's definition of moderate relativism suggests – there is *always* underdetermination of good reasons. However, sometimes Doppelt formulates his thesis of moderate relativism such that it consists in the claim that *often*[73], *typically*[74] or *paradigmatically*[75] scientific change is underdetermined by good reasons. Therefore, it is not clear how strong moderate relativism is supposed to be: does it imply a *global* or a *local* form of Kuhn-underdetermination? Does moderate relativism allow for the possibility that there might be cases in which the reasons alone suffice to explain the choice?[76]

Secondly, and this points to exactly the dilemma we have found in the case of Kuhn, the range of the relativity of moderate relativism is unclear. Quite in accord with Kuhn's statement of *shared* premises and values in Doppelt's version of the argument from the criterion, we find the idea that

> new theories often fail to be demonstrably more rational than their predecessors, on any standards which are *mutually* acceptable and applicable to *both*." (Doppelt 1986: 225, my italics)

Thus, moderate relativism implies that we explain theory-change by reference to social factors *and* good reasons that are rational on *shared* standards. This fits quite well with Doppelt's interpretation of Kuhn, who is said to insist on "the universality of [the] epistemic values in science" (Doppelt 2005: 698). However, obviously we should ask – in tune with the argument from the criterion and the principle of reflexivity that follows from it – why the good reasons of *shared* standards and the *universal* values should have normative force that is not reducible to social factors. Note that, in other passages, Doppelt appears to claim not that the values are *universal* but that they are subject to change:

---

[69] See also Doppelt 1988: 106.

[70] See also Doppelt 1988: 110.

[71] See Schofer 1999: 23.

[72] See also Doppelt 1988: 110.

[73] See Doppelt 1986: 225.

[74] See Doppelt 1986: 225, Doppelt 1988: 106.

[75] See Doppelt 1988: 110.

[76] Note that by Doppelt's claim of the *ineliminable* sociological factor in explanations of theory-choice he comes close to the form of relativism proposed by David Bloor that there is *necessarily* a social factor in the explanation of beliefs (see Bloor 1991: 17).

> [scientific] knowledge involves not just straightforward changes in theoretical and empirical belief, but normative transformations in the very problems, aims, *values*, and *standards* taken by scientific practitioners to be essential to scientific knowledge. (Doppelt 2001: 160, my italics)[77]

In this passage it seems that the values and standards *themselves* can transform and Doppelt explicitly accepts a shift-of-standard-thesis.[78] Standards of theory evaluation, Doppelt thinks, are just "historically particular" (Doppelt 2001: 159). The only criteria for scientific rationality are internal to scientific development itself such that the philosopher of science, who aims to evaluate the progress of past scientific communities, cannot use her own criteria of what she thinks rational progress consists in: "To capture *their* scientific rationality, we need to attend to *their* standards of adequate theory, their conceptions of science." (Doppelt 1982: 139). Therefore, Doppelt's moderate relativism is just epistemic relativism *proper*: moderate relativism invoking good reasons that are just good relative to a specific historical situation *just is* extreme relativism. In a nutshell: it seems that moderate relativism wants it both ways – universal values for the explanation of why scientists provide good reasons and community-relative values that should not lead to the assumption that scientific change is irrational.

## 4. Conclusion

In this paper I have argued that a moderate relativism inspired by Kuhn's work is no tenable position. The Kuhnian version of the argument from the criterion forestalls the way to a coherent, moderate form of epistemic relativism: either the argument leads to the conclusion that *all* normative commitments in theory-evaluation are just relative to the assent of the scientific community, or it does not. If the former, the conclusion is an *extreme* form of relativism; if the latter, the moderate relativist needs to master the slippery slope of the argument of the criterion without making ad hoc exemptions in the case of shared values and standards. Thus, there is no moderate, Kuhnian position *between* relativism and absolutism with respect to the epistemic realm.

**Markus Seidel**

Westf. Wilhelms-Universität Münster
Centre for Philosophy of Science/Zentrum für Wissenschaftstheorie
Domplatz 6
48143 Münster
Germany
maseidel@hotmail.com

[77] See also Doppelt 1982,:125: "Rival paradigms can [...] exhibit fundamental disagreements irresolvable by scientific argument concerning the set of problems and data that any adequate theory must treat (only some of which they share) [...]", and Doppelt 1982: 126: "[...] the question of which paradigm better explains that data they share, like the question of how serious the failure of one or the other is with respect to these data, essentially raises a more basic normative question: with respect to the data and problems they do not share, which are more essential or important for an adequate theory to explain? The force of Kuhn's thesis that this is an irreducibly normative disagreement is that it cannot be rationally resolved by standards acceptable to rival paradigms".

[78] See Doppelt 1988: 121, Doppelt 2001: 159.

## References

Barnes, B. and D. Bloor 1982: 'Relativism, Rationalism and the Sociology of Knowledge', in M. Hollis and S. Lukes (eds.): *Rationality and Relativism*. Cambridge (Mass.): The MIT Press, 21-47.

Bird, A. 2000: *Thomas Kuhn*. Princeton: Princeton University Press.

Bird, A. 2011: 'Thomas Kuhn's Relativistic Legacy', in S. D. Hales (ed.): *A Companion to Relativism*. Oxford: Wiley-Blackwell, 456-474.

Bloor, D. 1991: *Knowledge and Social Imagery*. Chicago: The University of Chicago Press.

Bloor, D. 2007: 'Epistemic Grace. Antirelativism as Theology in Disguise', *Common Knowledge* 13/2-3, 250-280.

Bloor, D. 2011: 'Relativism and the Sociology of Knowledge', in S. D. Hales (ed.): *A Companion to Relativism*. Oxford: Wiley-Blackwell, 433-455.

Boghossian, P. 2006: *Fear of Knowledge. Against Relativism and Constructivism*. Oxford: Oxford University Press.

Carrier, M. 2008: 'The Aim and Structure of Methodological Theory', in L. Soler et al. (eds.): *Rethinking Scientific Change and Theory Comparison*. Dordrecht: Springer, 273-290.

Doppelt, G. 1982: 'Kuhn's Epistemological Relativism: An Interpretation and Defense', in J. W. Meiland and M. Krausz (eds.): *Relativism. Cognitive and Moral*. Notre Dame/London: University of Notre Dame Press, 113-146.

Doppelt, G. 1983: 'Relativism and Recent Pragmatic Conceptions of Scientific Rationality', in N. Rescher (ed.): *Scientific Explanation and Understanding. Essays on Reasoning and Rationality in Science*. Boston/London: University Press of America, 107-142.

Doppelt, G. 1986: 'Relativism and the Reticulational Model of Scientific Rationality', *Synthese* 69, 225-252.

Doppelt, G. 1988: 'The Philosophical Requirements for an Adequate Conception of Scientific Rationality', in *Philosophy of Science* 55, 104-133.

Doppelt, G. 2001: 'Incommensurability and the Normative Foundations of Scientific Knowledge', in P. Hoyningen-Huene and H. Sankey (eds.): *Incommensurability and Related Matters*. Dordrecht: Kluwer, 159-179.

Doppelt, G. 2005: 'Scientific Revolutions', in D. Borchert (ed.): *Encyclopedia of Philosophy*. Vol. 8, 2nd edition. Detroit: Macmillan Reference, 694-703.

Goldman, A. 2010: 'Epistemic Relativism and Reasonable Disagreement', in R. Feldman and T. A. Warfield (eds.): *Disagreement*. New York: Oxford University Press, 187-215.

Hesse, M. 1980: *Revolutions and Reconstructions in the Philosophy of Science*. Brighton: The Harvester Press.

Holcomb, H.R. III 1987: 'Circularity and Inconsistency in Kuhn's Defense of Relativism', in *Southern Journal of Philosophy* 25/4, 467-480.

Hoyningen-Huene, P. 1993: *Reconstructing Scientific Revolutions. Thomas S. Kuhn's Philosophy of Science*. Chicago/London: The University of Chicago Press.

Kuhn, T.S. 1970a: 'Notes on Lakatos', in *PSA – Proceedings of the Biennial Meeting of the Philosophy of Science Association* Vol. 1970, 137-146.

Kuhn, T.S. 1970b: 'Reflections on my Critics', in I. Lakatos and A. Musgrave (eds.): *Criticism and the Growth of Knowledge*. London/New York: Cambridge University Press, 231-278.

Kuhn, T.S. 1970c: *The Structure of Scientific Revolutions*. Chicago/London: The University of Chicago Press.

Kuhn, T.S. 1977: 'Objectivity, Value Judgment and Theory Choice', in *The Essential Tension*. Chicago/London: The University of Chicago Press, 224-252.

Kuhn, T.S. 2000: 'Afterwords', in *The Road Since Structure*. Chicago/London: The University of Chicago Press, 224-252.

Kusch, M. 2010: 'Kripke's Wittgenstein, On Certainty, and Epistemic Relativism', in D. Whiting (ed.): *The Later Wittgenstein on Language*. Basingstoke: MacMillan, 213-230.

Lakatos, I. 1970: 'Falsification and the Methodology of Scientific Research Programmes', in I. Lakatos and A. Musgrave (eds.): *Criticism and the Growth of Knowledge*. London/New York: Cambridge University Press, 91-196.

Nola, R. 1990: 'The Strong Programme for the Sociology of Science, Reflexivity and Relativism', in *Inquiry* 33, 273-296.

Nola, R. 2003: *Rescuing Reason. A Critique of Anti-Rationalist Views of Science and Knowledge*. Dordrecht: Kluwer.

Pritchard, D. 2011: 'Epistemic Relativism, Epistemic Incommensurability, and Wittgensteinian Epistemology', in S. D. Hales (ed.): *A Companion to Relativism*. Oxford: Wiley-Blackwell, 266-285.

Sankey, H. 2011a: 'Epistemic relativism and the problem of the criterion', in *Studies in History and Philosophy of Science A* 42, 562-570.

Sankey, H. 2011b: 'Incommensurability and Theory Change', in S. D. Hales (ed.): *A Companion to Relativism*. Oxford: Wiley-Blackwell, 456-474.

Sankey, H. 2012a: 'Methodological Incommensurability and Epistemic Relativism', in *Topoi*.

Sankey, H. 2012b: 'Scepticism, relativism and the argument from the criterion', in: *Studies in History and Philosophy of Science A* 43, 182-190.

Sankey, H. 2013: 'How the epistemic relativist may use the sceptic's strategy: A reply to Markus Seidel', in *Studies in History and Philosophy of Science A* 44/1, 140-144.

Schofer, B. 1999: *Das Relativismusproblem in der neueren Wissenssoziologie. Wissenschaftsphilosophische Ausgangspunkte und wissenssoziologische Lösungsansätze*. Berlin: Duncker & Humblot.

Seidel, M. 2011a: 'Karl Mannheim, Relativism and Knowledge in the Natural Sciences – A Deviant Interpretation', in R. Schantz and M. Seidel (eds.): *The Problem of Relativism in the Sociology of (Scientific) Knowledge*. Frankfurt (Main): Ontos, 183-213.

Seidel, M. 2011b: 'Relativism or Relationism? A Mannheimian Interpretation of Fleck's Claims About Relativism', in *Journal for General Philosophy of Science* 42, 219-240.

Seidel, M. 2013a: 'Why the epistemic relativist cannot use the sceptic's strategy. A comment on Sankey', in *Studies in History and Philosophy of Science A* 44/1, 134-139.

Seidel, M. 2013b: 'Scylla and Charybdis of the epistemic relativist. Why the epistemic relativist still cannot use the sceptic's strategy', in *Studies in History and Philosophy of Science A* 44/1, 145-149.

Seidel, M. forthcoming: *Epistemic Relativism – False, But With the Right Intuition. A Critique of Relativism in the Sociology of Scientific Knowledge*. PhD-thesis, University of Siegen, 304 pages.

Siegel, H. 2004: 'Relativism', in I. Niiniluoto et al. (eds.): *Handbook of Epistemology*. Dordrecht: Kluwer, 747-780.

Williams, M. 2007: 'Why (Wittgensteinian) Contextualism is not Relativism', in *Episteme* 4/1, 93-114.

Wray, K. B. 2011: *Kuhn's Evolutionary Social Epistemology*. New York: Cambridge University Press.

# When Is It Rational to Believe a Mathematical Statement?

Jendrik Stelling

Traditional philosophy of mathematics has long held that the fundamental business of mathematics is to ascertain the truth of mathematical propositions by finding suitable proofs for them, where "proof" means a chain of logico-deductive steps, starting from accepted premisses or axioms, and terminating in the proposition to be proved. Under this view, we are justified to believe in a mathematical statement if and only if there is a proof for it. My goal here is to call this view into question. To this end, we will discuss two examples of mathematical argumentation that suffer, *prima facie*, from the same shortcoming. We will then attempt to extract the significant differences and use them to elucidate the concept of proof that actually informs mathematical practice.

## 1. Oracles and Probabilities

Imagine that, in some remote mountain cave, there lives a mysterious mathematical oracle. If one manages to make the dangerous trek there, one is allowed to ask a single question of the form "is mathematical statement *p* true?" The oracle knows the truth value of all mathematical statements and will neither lie to us nor try to deceive us in any way. Everything the oracle says can be taken at face value. "Mathematical statement" is meant to be understood as something that might be considered a result in some publication; e.g., we might ask the oracle about the truth value of "P=NP" but not whether set theory or category theory is the more appropriate foundation of mathematics.[1] Let's also, for the moment, ignore questions about things that are in some sense indeterminate, such as the Continuum Hypothesis.

Imagine then that one day we make the trip and stand before the oracle. "Oracle," we ask, "is it true that all even numbers greater than 2 can be expressed as the sum of two primes?" The oracle pauses for a moment, and gravely intones, "Goldbach's conjecture is true if and only if, on randomly drawing a real number, the number you draw is not a natural number."

Given what we know about the oracle, what do we do with this answer? Obviously we cannot simply carry out the experiment and draw a random real number. So we need a mathematical approach: When we ask for the probability of drawing a given number from the reals, we need to come up with an account of probabilities that can handle the enormous amount of real numbers we face. Ordinarily, we would perhaps try to define probabilities as a quotient, the positive outcomes divided by all possible outcomes. But since there are infinitely many real numbers, and within them infinitely many natural numbers, this won't do. But we can use some concepts from measure theory to help us circumvent this problem. The Lebesgue measure, for instance, works by comparing smaller slices of real numbers, and seeing how many naturals we can find in them, and then extending these results across the real number system. Important in this regard is the notion of a zero measure set (meaning an interval of

[1] I presuppose that the future won't show either framework to be inconsistent; the decision between set theory and category theory is not a valid question precisely because I understand it to be a judgement call and not amenable to mathematical proof.

the real numbers that, when compared to all the reals, has Lebesgue measure zero). The idea here is quite intuitive: Anything, regardless of how we measure it, should have a measure of zero if, for every positive measure $\varepsilon$ (no matter how small), the thing to be measured is smaller than $\varepsilon$. In other words, if any given positive measure, regardless of how small we make it, is still too big, then the thing to be measured has size zero. Turning this into a definition, we get:

**Definition** Let $(a, b)$ be an open interval in $\mathbb{R}$. Define $\ell(a, b)$, called the *length* of $(a, b)$, as $b - a$. An interval of real numbers $[a, b]$ has *measure zero* if, given any $\varepsilon > 0$, there exists a (finite or countable) family of open intervals $\langle l_i \rangle_{i \in I}$ such that $[a, b] \subset \cup\langle l_i \rangle_{i \in I}$ and $\sum \ell(l_i) \leq \varepsilon$.

**Theorem** The set of naturals $\mathbb{N}$ has measure zero within the reals $\mathbb{R}$.

Proof. Given $\varepsilon > 0$, let

$$l_n = \left(n - \frac{\varepsilon}{2^{n+1}}, n + \frac{\varepsilon}{2^{n+1}}\right), n = 1,2, \ldots$$

Then

$$\ell(l_n) = \frac{\varepsilon}{2^n}, \mathbb{N} \subset \bigcup_{n=1}^{\infty} l_n,$$

and

$$\sum_{n=1}^{\infty} \ell(l_n) = \sum_{n=1}^{\infty} \frac{\varepsilon}{2^n} = \varepsilon.$$

The concept of zero measure set translates into probability theory quite straightforwardly. Given that the concept of measure we employed here implies that the full set of reals has measure one, and given that the larger an interval, the bigger its measure, it's not a big leap to simply use the size of a set as the probability of us randomly drawing a number from it. So the probability of drawing a real number from the set of reals is one, and the probability of drawing any member of a zero measure set is zero. And indeed, this is the way measure theory typically handles probabilities over the reals.[2]

So the probability of drawing an integer from the full set of reals is zero. But of course this means that our rational conviction that we will not, on randomly drawing a real number, draw an integer, should be absolute.

Taking all of this together, we come to the conclusion that it would be completely irrational to expect Goldbach's conjecture to be anything but true, based on the Oracle's statement. Our degree of belief in the conjecture ought to be one. Even though it is not *logically* impossible that the number we draw is a natural number, just as it is not logically impossible that a solid block of iron passes through a solid surface via quantum tunnelling, the probability in both cases is so utterly minuscule that even the suggestion that we should remain open to the possibility of it actually happening is preposterous.

Let's reiterate this. Based on the fact that the Oracle never lies, and the fact that it has essentially told us that the probability of Goldbach's conjecture being false is zero, we should by all means be completely convinced that it is true. This might not be very satisfying knowledge, since we still have no way of knowing *why* it is true, but we should be convinced nonetheless.

[2] See for instance Athreya & Lahiri 2006.

So what does all of this have to do with actual mathematics? My claim here is that, although there are no mathematical oracles like the one we have described, there are situations in mathematical practice that bear a striking resemblance to the case of the oracle.

## 2. The Riemann Hypothesis

Let's start with the Riemann Hypothesis (RH). RH is probably the most important unsolved problem of number theory, if not all of (pure) mathematics.[3] Stating it is fairly simple, since it is at its most basic simply a hypothesis about the behavior of a single function, the so-called zeta function. This function, defined as the analytic continuation ζ(z) of the infinite series

$$\sum_{n=1}^{\infty} n^{-z}$$

(with $z$ complex) is interesting because the distribution of its roots is just as mysterious as that of the primes. The roots of ζ(z), i.e., the solutions of the function, are complex numbers. Complex numbers are of the form $a+bi$, where $a,b$ are real numbers, and $i$ is the imaginary number $\sqrt{-1}$. Riemann himself conjectured[4] that the roots of ζ(z) all have real part $\frac{1}{2}$, i.e., for all solutions $a+bi$ to the Zeta Function, a= $\frac{1}{2}$. If you are familiar with the geometrical representation of complex numbers, this means that all solutions lie on the line x= $\frac{1}{2}$, parallel to the imaginary axis (known as the 'critical line'). What do we know about this conjecture? The Riemann Hypothesis, as it is known, remains as yet unproved. Hardy showed that infinitely many of the function's roots do indeed have real part $\frac{1}{2}$, but there is no proof that all of them do. (We do know that at least two-fifth of them do, and that almost all zeros lie arbitrarily close to the critical line.[5]) However, mathematicians are still by and large convinced that the Riemann Hypothesis is true. In the following, we will look at an argument by the French mathematician Arnauld Denjoy that could explain why.[6]

### *2.1 Three Functions and a Conjecture*

RH is famous for having far-reaching consequences for our understanding of prime numbers. So it comes as no great surprise that this will be our starting point:

> **Definition** Let $\pi(x)\colon \mathbb{N} \longrightarrow \mathbb{N}$ be the function that, for any given natural number $n$, gives the number of primes less than or equal to $n$. Thus $\pi(10) = 4$, since up to (and including) 10, there are four prime numbers: 2,3,5, and 7.
>
> **Prime Number Theorem** $\lim_{x\to\infty} \frac{\pi(x)}{x/\log x} = 1$.

In other words, for larger and larger natural numbers, the number of primes becomes closer and closer to $\frac{x}{\log x}$. This is the famous Prime Number Theorem that was proved independently by both de la Vallée-Poussin and Hadamard in 1896.[7] This theorem is important since it tells

[3] At least that's the portrayal the hypothesis receives in the official Millenium Prize problem description.

[4] The hallmark paper that contains the conjecture is Riemann 1859.

[5] The first result was proved in Conrey 1989, the second in Bohr & Landau 1914.

[6] We will come back to the reception of Denjoy's argument within the scientific community, but I want to clear something up right now: I am not claiming that Denjoy's argument is the factual reason for the mathematicians' conviction that RH is true; in fact, it certainly isn't. What I want to say is that the argument is an interesting attempt at an explanation for the conviction, rather than its historical origin. Denjoy's original argument was made in his paper Denjoy 1931.

[7] See Hadamard 1896, as well as de la Valée-Poussin 1896. For more on the history see, e.g., Kline 1990.

us something about the number of primes in general, as well as that the further we get along the number line, the less new primes turn up.

What else can we find out about the distribution of prime numbers? First of all, we need another function.

**Definition** Let $\mu(x)$ (the Möbius-Function) be defined as follows. For any natural number $x$, factor $x$ into primes. If one or more of the factors repeats, as in $18 = 2 \times 3 \times 3$, then $\mu(x) = 0$. If no factor repeats, as in $10 = 2 \times 5$, count them. If the number of factors is even, let $\mu(x) = 1$, if it is odd (as in $30 = 2 \times 3 \times 5$), let $\mu(x) = -1$.

This function has a distinct value for every natural number. Now we choose any such number, $n$, and add the values of $\mu(x)$ for all $x$ less or equal to $n$. This is a sum of +1's and -1's (well, and zeros), and it is itself a function of the number $n$ we chose. Call this third new function $M(n)$ (the Mertens-Function).

**Definition** $M(n) = \sum_{x=1}^{n} \mu(x)$

As an example:

| $n$ | prime factors | $\mu(n)$ | $M(n)$ |
|---|---|---|---|
| 2 | 2 | -1 | -1 |
| 3 | 3 | -1 | -2 |
| 4 | 2×2 | 0 | -2 |
| 5 | 5 | -1 | -3 |
| 6 | 2×3 | +1 | -2 |
| 7 | 7 | -1 | -3 |
| 8 | 2×2×2 | 0 | -3 |
| 9 | 3×3 | 0 | -3 |
| 10 | 2×5 | +1 | -2 |
| ... | ... | ... | ... |

We can now ask about the behavior of the Mertens-Function. The values of $M(n)$ vary, but without any apparent pattern. In the beginning, of course, the Mertens Function will generally be negative, as there are quite a lot of prime numbers at the start of our number sequence. But as we go further and further along, less and less new prime numbers come up. (This is by itself an implication of the Prime Number Theorem.) So suppose we pick out a very large number $n$ without repeating prime factors. Since the density of primes goes to zero as $n$ goes to infinity, $n$ is very unlikely to be a new prime number. Thus $n$ most likely has a large number of prime factors. Looking only at those numbers that do not have repeating prime factors, is there any reason to suspect that *more* of them will have an even than an odd number of prime factors? It seems that there isn't. In other words, as $\mu(x)$ goes through $\mathbb{N}$, there does not seem to be any good reason to suspect, for sufficiently high values of $x$, the next value to be a +1, rather than a -1 (given that it's not a zero). The values appear to behave like random coin flips. If this is correct, then we can derive a number of consequences from it.

– On the one hand, in a coin flip scenario we would expect to see just as many heads as tails in the long run. Mathematically speaking, we would expect $M(n)$ to have an average order of zero.

– On the other hand, extremely long runs of only heads (or only tails) seem highly improbable. So we would expect the Mertens-Function to behave in a certain steady fashion: If we really are in a coin flip scenario, then the function should not grow

extremely fast in some places, and fall extremely fast in others, since this would indicate the presence of long runs of an even (respectively, odd) number of unique prime factors.

Putting these thoughts together, we can express our conjecture the following way:

**Conjecture:** $M(n) = O(n^{\frac{1}{2}+\varepsilon})$. (As $n$ goes to infinity, the rate of growth of $M(n)$ is not higher than a constant multiple of $n^{\frac{1}{2}+\varepsilon}$, where $\varepsilon$ is arbitrary, but larger than zero. The specifics of this notation need not interest us at the moment.)

Why is this conjecture interesting to us? Because there is good reason to believe in it, and because it is provably equivalent to the Riemann Hypothesis.[8] In other words, *RH is true if and only if $M(n)$ behaves like a random walk.*

### *2.2 Arguing for the Conjecture*

We can calculate the probability that, for a randomly chosen number $x$, $\mu(x) \neq 0$, i.e., $x$ has no repeating prime factors. For this to be the case, $x$ must not be divisible by the square of a prime number. This makes sense: Since the prime factorization of 90 is $2 \times 3 \times 3 \times 5$, 90 is obviously divisible by $3 \times 3$, i.e. 9. So if $x$ is not divisible by any square of a prime number, then $x$ has no repeating prime factors and $\mu(x) \neq 0$.

What is the probability of a randomly chosen number to not be divisible by $2^2 = 4$? Imagine all the natural numbers written out on a piece of paper. Then take a pencil and cross out every fourth number—all that remain are not divisible by four. (This method is known as the Sieve of Eratosthenes.) Then, since you crossed out every fourth number, exactly $3/4$ of all numbers remain, which means that the probability that a randomly chosen number is not divisible by four is $3/4$. This method works for all the squares of primes; the probability that a random number is not divisible by $3^2 = 9$ is $8/9$, that it is not divisible by $5^2 = 25$ is $24/25$, and so forth. This leaves us with the probability of $\mu(x) \neq 0$ equal to:

$$3/4 \times 8/9 \times 24/25 \times 48/49 \times \ldots,$$

which is simply equal to $\frac{6}{\pi^2}$, or around 61%. Then the probability that $\mu(x) = +1$ is $\frac{3}{\pi^2}$, and the probability that $\mu(x) = -1$ is the same. All in all, the expected value of $\mu(x)$ is zero, the +1's and -1's should just about cancel each other out. Also note that, since there are in the long run as many +1's as there are -1's, for any given large number $n$, picked randomly from the set of all the natural numbers, the probability of $\mu(x) = +1$ should be equal to that of $\mu(x) = -1$. In fact, we expected $M(n)$ to have an average order of zero, if it behaves like a random walk. And indeed, it does: this is itself a result provably equivalent to the Prime Number Theorem.

Now suppose we pick $n$ natural numbers at random, and sum up their $\mu$-values. If we have more numbers among them for which $\mu(x) = +1$, then our sum will be positive. If for more of them $\mu(x) = -1$, the sum will be negative. But if picking a random number without repeating prime factors is like flipping a coin when it comes to its $\mu$-value, then, as $n$ approaches infinity, the probability that $M(n)$ does not grow or fall extremely fast in some places approaches one. Putting this into more precise mathematical terms, we have:

$$\lim_{n\to\infty} p\left(M(n) = O\left(n^{\frac{1}{2}+\varepsilon}\right)\right) = 1.$$

[8] The proof for the equivalency can be found, e.g., as Theorem 14.25 (C) in Titchmarsh 1986: 370.

But we saw earlier that $M(n) = O\left(n^{\frac{1}{2}+\varepsilon}\right)$ is equivalent to RH itself, so by substituting one for the other, we get:

The Riemann Hypothesis has probability one.

Have we proved our conjecture, and with it, the equivalent Riemann Hypothesis? Well, technically, we haven't. We assumed that it makes no difference whether we (a) pick $n$ numbers at random, or (b) choose a number $n$ and 'pick' the numbers one to $n$. The latter is what we need to prove our conjecture, the former is the correct starting point for our probabilistic reasoning. But substituting the random pick for the 'one-to-$n$'-pick is justified if the distribution of +1's and -1's in our $\mu$-evaluation of the numbers one to $n$ is truly random. If it is, the first $n$ numbers are nothing special; they are just as good as any random sample of $n$ numbers.

What we *have* established, however, is that RH is true with probability one. We have not established the hypothesis itself, but we have given ample reason to be completely convinced by it. In fact, the probabilistic account leaves no option even to be carefully agnostic—it would be irrational to doubt an event that has probability one.

But this is not a proof of RH, as evidenced by the fact that the community does not treat it as such. Edwards, one of the authorities on all things zeta, calls Denjoy's argumentation "quite absurd", and overall sentiment in the community seems to agree with him.[9] One might suspect that the reason for this rejection is due to the 'merely' probabilistic nature of the argument. I do not believe that this is the case. To argue for this, let's look at another example from mathematical practice.

## 3. Finite Projective Planes

Just as Euclidean Geometry takes place in a Euclidean space, so projective geometry presupposes a projective space. And just like the two-dimensional reduct of a Euclidean space is a Euclidean plane, so we have projective planes. Like its Euclidean sibling, projective geometry is most usually practiced within an infinitary such plane, but finite projective planes do exist.[10]

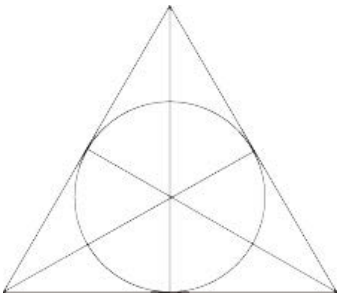
**Definition:** A *finite projective plane of order n* is a collection of $n^2+n+1$ lines and $n^2+n+1$ points such that:

1. Every line contains n+1 points,
2. every point lies on n+1 lines,
3. any two distinct lines intersect in one point, and
4. any two distinct points lie on one line.

The simplest finite projective plane is of order one: it contains three lines and three points, and looks like a triangle. The next smallest finite projective plane is of order two, contains seven lines and points, and looks like this:

[9] Edwards discusses Denjoy's argument in his book Edwards 1974: 268.

[10] Throughout this section we will follow the exposition set forth in Lam 1991.



Historically the start of the research program goes back to a paper on axiomatizations for geometry by Oswald Veblen from 1904, in which he used the finite projective plane of order two as an example. In the years that followed, Veblen published a series of papers on the topic, together with collaborators W. H. Bussey and J. H. M. Wedderburn, that established the existence of most of the finite projective planes of lower order, as well as all four non-isomorphic planes of order nine.[11] One of the missing cases was $n$=6. This case was solved in 1938 by R. C. Bose, when he tackled the problem from a combinatorial point of view, by developing the similarities between finite projective planes and so-called orthogonal Latin squares. He showed that no finite projective plane of order 6 can exist.[12] Finding out whether finite projective planes exist for any given order $n$ was by now first and foremost a combinatorial problem, and the combinatorics community quickly became interested in the search. The next open problem was the $n$=10 case.

In the 1970s, the problem of $n$=10 was still open, and it was unclear how the problem was to be handled. The defining impulse came from the mathematician E. F. Assmus, Jr., at a conference talk in Oberwolfach, Germany.[13] His idea was to extend the combinatorial side of the issue by representing finite projective planes by so-called incidence matrices. The idea here is quite simple: If you number all the points and lines, you can describe the projective plane as a matrix where a combination of a line and a point receives the value 1 if the point lies on the line, and 0 otherwise. Thus the incidence matrix for the $n$=2 case looks like this:

| $n = 2$ | P1 | P2 | P3 | P4 | P5 | P6 | P7 |
|---|---|---|---|---|---|---|---|
| L1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 |
| L2 | 0 | 1 | 1 | 0 | 1 | 0 | 0 |
| L3 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| L4 | 0 | 0 | 0 | 1 | 1 | 0 | 1 |
| L5 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| L6 | 0 | 1 | 0 | 0 | 0 | 1 | 1 |
| L7 | 1 | 0 | 1 | 0 | 0 | 0 | 1 |

These matrices can now be translated into vector spaces. The vectors under consideration here, called codewords, can be shown to determine the whole problem. The number of ones in a codeword is called its weight. Let $w_i$ be the number of codewords of weight $i$, then it can be

[11] The paper that started it all is Veblen 1904. See also the results in Veblen & Bussey 1906 as well as Veblen & Wedderburn 1907.

[12] Strictly speaking, Bose's paper (Bose 1938) does not mention the $n$=6 case, instead focussing on the examples $n$=4, 8, 9, 16, 25, 27. The $n$=6 case, however, follows from his main result in the following way: Bose pointed out that a finite projective plane of order $n$ exists if and only if there is a complete set of $n$-1 mutually orthogonal Latin Squares of order $n$. A much older result (Tarry 1900) had already established that there isn't even a single pair of mutually orthogonal Latin Squares of order 6, so the non-existence of the finite projective plane of order 6 follows as a corollary.

[13] The talk itself was an exposition of the results in Assmus & Mattson 1970.

proved that the existence of a finite projective plane of order 10 can be reduced to checking codewords of weight 12, 15, and 16. In other words, if there is a finite projective plane of order 10, it must have a corresponding incidence matrix, which in turn contains a codeword of weight 12, 15, or 16. Given such a codeword, we can consider its behavior and generate a number of submatrices of the $n$=10 incidence matrix. At least one such submatrix can be expanded to the full matrix. So if we try to find the incidence matrix itself, we can start by considering the submatrices generated by the possible codewords and check whether each submatrix can be expanded to the full matrix. If we find a way to do so, we have just solved the problem: we have successfully constructed a finite projective plane of order 10. If we cannot complete any of the submatrices, the full incidence matrix does not exist, and neither does the projective plane.

As soon as this had been worked out, a group of mathematicians proved that $w_{15}$=0, i.e., there are no codewords of weight 15, after around three hours of computer time.[14] Tackling the remaining cases proved to be more difficult. By 1974, Carter made significant headway with the $n$=16 case in his dissertation[15], thereby opening the field for the team that eventually was to solve the problem: the mathematicians Clement Lam, Larry Thiel and Stanley Swiercz.

> This was the situation when we entered the picture. While we knew $w_{15}$=0, the search for weight 16 codewords was about three-quarters done and the search for weight 12 codewords was presumed to be too difficult. (Lam 1991: 312)

Things moved along further when the three started to tackle the problem through intensive use of computers. The details of the story, leading up to the finished run of calculations, are of no specific importance to us here, interested readers might wish to consult Lam 1991, where the whole history is given in much more detail. Suffice it to say that during the 1980s, more and more cases were successfully calculated, until on November 11th, 1988, the CRAY supercomputer that was running most of the calculations by that point was finished, no full incidence matrix had been found and the finite projective plane of order ten was pronounced dead.

This would be where our story ends, if not for what happened next. So far, the search for the order ten plane is, like the Four Color Theorem, merely another computer-assisted proof. But here is where things get interesting.

One week after the CRAY finished its calculations, Nick Patterson (the deputy director of the Communication Research Division at the Institute for Defense Analyses, where the CRAY supercomputer was running) called Lam. Patterson, who had "looked after the day-to-day running of the program for over two years", and whom Lam calls "the unsung hero in the successful completion of our work" (Lam 1991:314), had checked the computation logs and found cause for concern. He reported that there had been an error code four in one of the partial computations.[16] Code four errors meant that this part of the computation was so large that the software the CRAY was running could not handle the computation. What to do? Lam, Thiel and Swiercz split the problem up into 200 subcases and used many different machines, among them the CRAY, to solve them part by part. On November 29th, after between 2000 and 3000 hours of computing time on the CRAY supercomputer alone, the problem was pronounced solved again. The authors were flooded with attention from *Science*, the *New York Times*, the *Scientific American*, and, of course, interested mathematicians.

> After the burst of publicity, we finally managed to read the magnetic tape containing the statistics. To our horror, we found another A2 with an error number four. However,

[14] The resultant publication is MacWilliams, Sloane & Thompson 1973.

[15] See Carter 1974.

[16] These partial computations, called A2's, are submatrices of the suspected incidence matrix that have been constructed in a specific way. For more details see Lam, Thiel & Swiercz 1989: 1118.

> we knew exactly what to do this time and there was no panic. It was handled exactly the same way as the previous one. By the end of January 1989, the plane of order 10 was dead a third and, hopefully, the final time. (Lam 1991: 316)

No more errors were found from this point onwards. But the fact that errors were found in the first place opened up the problem of whether or not there had been errors that slipped through the cracks, errors either due to faulty programming or undiscovered hardware malfunctions. After discussing (and discarding) the possibilities of programming errors, Lam says:

> There is, moreover, the possibility of an undetected hardware failure. A common error of this type is the random changing of bits in a computer memory, which could mean the loss of a branch of a search tree. This is the worst kind of hardware error, because we might lose solutions without realizing it. The CRAY-1A is reported to have such errors at the rate of about one per one thousand hours of computing. At this rate, we expect to encounter two to three errors! We did discover one such error by chance. After a hardware problem, Patterson reran the 1000 A2's just before the failure and the statistics have changed for the A2 processed just prior to the malfunction. How should one receive a "proof" that is almost guaranteed to contain several random errors?
>
> Unfortunately, this is unavoidable in a computer-based proof—it is never absolute. However, despite this reservation, we argued [...] that the possibility of hardware errors leading us to a wrong conclusion is extremely small. (Lam 1991: 316)

How small exactly? The worst case scenario would be this: Suppose there is in fact a finite projective plane of order 10. Then there are 24,675 codewords of weight 19, each giving rise to an A2.[17] The overall number of A2's, including the ones that do not lead to a positive result, is somewhere around 500,000. Since we are trying to conserve computing time as much as possible, the procedure begins by checking all A2's for isomorphisms between them. Because two isomorphic A2's behave the same way, we only need to check one of them, and thus the superfluous copies are eliminated at the start. So if all of these 24,675 A2's *are* isomorphic, then we really only check a single one of them, and therefore there is exactly one A2 in our list that can successfully be extended to an order ten plane. The possibility of this specific A2 being affected by a memory error, given that two to three such errors happen during the calculations, is still less than $10^{-5}$, or 0.001%. Far more likely, however, is that not all of the A2's that give rise to the supposed plane of order ten are isomorphic, in which case there would be several distinct A2's that would each have to have been affected by such a random error, in which case "the probability of hardware errors affecting all of them is infinitesimal." (Lam 1991: 317)

> Basically, the argument depends on the observation that if a plane of order ten exists, it can be constructed from many different starting points. Random hardware failures are unlikely to eliminate all of them. In other words, the fact that no one has yet constructed one is a very strong indication that it does not exist. (Lam 1991: 317)

It should be noted that the authors of this result themselves were careful in the original publication to point out the uncertainty behind their methods:

> Because of the use of a computer, one should not consider these results as a "proof", in the traditional sense, that a plane of order 10 does not exist. They are experimental results and there is always a possibility of mistakes. (Lam, Thiel & Swiercz 1989: 1120)

[17] We are skipping over some details here. Readers interested in why all of a sudden we are talking about weight 19 codewords, or how we arrived at that specific number of A2's, should consult Lam, Thiel & Swiercz 1989.

Yet the mathematical community seems to have accepted the non-existence of the plane of order 10 as an established fact, just like the proposition that four colors suffice to color any map.

## 4. Conclusion

So where does this leave us? Both arguments we encountered seem to share the same flaw—the argument for the Riemann Hypothesis in section 2 was merely probabilistic (even though the probability for RH being true was one), and the computer-aided proof for the non-existence of finite projective planes of order ten in section 3 is only accurate if no random hardware errors have given us a false solution, which in itself is highly unlikely. Why, then, do we accept the second, but reject the first argument?

One of the more common arguments here goes roughly as follows. Mathematicians operate under the premiss that acceptable proofs are those that can in principle be transformed into a formal chain of deductive, gapless arguments in some formal calculus. The conclusion of such a chain of reasoning is the established theorem, which then holds immutably true, given that the starting points of the deductive chain are true. Perhaps the most fervent supporters of this view were the group of mathematicians who collectively worked under the *nom de guerre* Nicholas Bourbaki:

> In practice, the mathematician who wishes to satisfy himself of the perfect correctness or 'rigour' of a proof or a theory hardly ever has recourse to one or another of the complete formalizations available nowadays, nor even usually to the incomplete and partial formalizations provided by algebraic and other calculi. In general he is content to bring the exposition to a point where his experience and mathematical flair tell him that translation into formal language would be no more than an exercise of patience (though doubtless a very tedious one). (Bourbaki 1968: 8)

Under this view, rejected argumentations like the one by Denjoy would fail to satisfy mathematicians because the impression is that they cannot be explicated into a formal chain of reasoning. On the other hand, since the calculations done by the computers to establish the non-existence of a finite projective plane of order ten are by their very nature Turing-computable and thus recursive, they can certainly be formalized, and hence the calculations constitute a proof.

I think there are several problems with this view. First, there is the general issue that "proof in a formal calculus" is, strictly speaking, a notion devoid of content. Formal calculi come in various shades and forms, and even under the reasonable understanding that we want formalizations that are in some form tractable, we are left with a huge variety of possibly admissible rules, ranging from the mathematical puritanism of Russian-style recursive intuitionism, through the cautiously liberal formalism in the vein of Gentzen to the hedonistic irresponsibility of higher set theory. And this does not even begin to face the problems stemming from the fact that we can derive any result we want by introducing suitable premisses and axioms.

Much more pressing, however, is the fact that Denjoy's reasoning *can* be brought into formal form—at least as much as we assume that any purely mathematical argument within analytic number theory can. We end up (presumably) with a gapless sequence of logical deductions that lead, invariably, to the "absurd" conclusion that the Riemann Hypothesis might be false, but with probability zero. On the other hand, the central theorem of section 3 can be brought into formal form if and only if no computational errors occurred. In case some random hardware errors influenced the calculations, the whole argument can no longer, even in principle, be formalized. Note that this holds even if the hardware errors are negligible in the sense that they do not lead to a false result. Even if the conclusion is factually correct and

there is no finite projective plane of order ten, an error in one of the computational branches that 'accidentally' gives the correct result will still mean that at this point of the argument the theoretical transformation into a rigorous formal deduction breaks down. And, as we have seen, the probability that such an error occurred is actually fairly substantial. So the reason we accept the computational but reject the probabilistic argument cannot lie in the underlying assumption that one but not the other is a formal proof in essence, if not in execution.

Note also that another obvious explanation doesn't work: We might be tempted to point out that the significant difference between the two cases is simply that, while Denjoy's argument is indeed a proof that RH has probability one, it is decidedly not a proof that RH is *true*. And since the latter is what we're looking for, the reason the argument isn't accepted by the community is that it proves the wrong thing. The proof by Lam, Thiel and Swiercz, on the other hand, does prove what it sets out to prove, and therefore delivers the correct content.

This objection, however, presupposes that there is a significant difference between the proposition that RH is true and the proposition that RH has probability one. And, while this may be so, simply stating that it is won't help us. Instead, we need an explanation as to *why* the two propositions are different enough to allow for the explication that Denjoy's argument simply delivers the wrong content. Needless to say, the explanation will have to be applicable to the order ten plane as well, if we want to use it to shed any light on the situation as a whole. And any such explanation will presumably at some point run into the objection that the possibility of random hardware errors having lead us to the wrong result can not logically be ruled out. But at that point the nonexistence of the order ten plane is probabilistic in nature, much like Denjoy's argument, so we have come full circle: we are still in need of an explanation as to why the two cases are different, even if they're both probabilistic.

My proposed explanation, then, is this. There is, I want to argue, a feeling of *due diligence* that forms part of the paradigm of the current state of mathematical research. Formulated as a maxim, it can be read as saying something like, "get as close as possible to the strongest solution conceivable for the problem in question." This needs some explanation. I do not want to say that mathematicians automatically strive towards general over particular solutions (though they might). Neither am I suggesting that mathematicians search for solutions that have a maximum impact across mathematics over solutions that influence only small portions of some particular field (though they might). What I mean is this:

(a) If Denjoy's argument ends up being wrong, i.e., if RH turns out to be false, we as mathematicians have made a mistake in the argument. There is nothing outside of us, the mathematical reality, and the proof as written down on the papers on our desk, and if RH turns out to be false, the buck stops with us.

(b) However, in the case there actually is a finite projective plane of order ten, we are not to blame in the same way. If it turns out that there really was an undetected hardware error that lead us astray, then hardware degradation, solar flares, compounding effects on the quantum level or any of a number of other unfortunate circumstances are at fault. In that case, there's us, the programs we wrote, mathematical reality and the outside world, the latter of which foiled our plans.

Due diligence, in this sense, is the feeling that, even on the off-chance that circumstances should have conspired to foil our plans, we have done as much to solve the problem as can reasonably be expected.[18] Mathematicians do not have influence over the reliability of

[18] It is tempting to construct an analogy to the way in which the axiom of choice, even though it leads to patently absurd conclusions when interpreted as a statement about the physical world, nonetheless finds acceptance within the mathematical community. It is almost as if there is a tacit agreement that, as long as the mathematics is beautiful, if the external world doesn't want to play along, so much the worse for the external world.

computer memory, but they do control their own mathematical actions—be they constructing proofs or writing programs. It is this area of direct influence that underlies strict scrutiny by the mathematical community. Insecurities can be excused inasmuch as they are not the fault of the practitioners, but rather a particularly mathematical form of bad luck—what the insurance business calls "an act of God".

The fundamental difference between the two cases, then, is not about the content of the statements or the reliability of the arguments brought forth to support them. It is, in the last instance, about due diligence—the feeling that we have done everything in our power to make the proof as strong and 'complete' as it can be. This could be construed as an undue influence of the extra-mathematical on our notion of proof, but I don't think it should be. Rather, I want to argue, we ought to expand our understanding of what is properly mathematical to include phenomena such as due diligence—not as a foreign sociological influence on pure mathematical concepts, but as an integral component of what forms mathematical reality.

**Jendrik Stelling**

University of Rostock
jendrik.stelling@uni-rostock.de

## References

Assmus, E. F. and Mattson, H. F. 1970: 'On the prossibility of a Projective Plane of Order 10', in *Algebraic Theory of Codes II, Air Force Cambridge Research Laboratories Report AFCRL-71-0013, Sylvania Electronic Systems*, Needham Heights, Mass.

Athreya, K. and Lahiri, S. 2006: *Measure Theory and Probability Theory*. New York: Springer.

Bohr, H. and Landau, E. 1914: 'Ein Satz über Dirichletsche Reihen mit Anwendung auf die $\zeta$-Funktion und die $L$-Funktionen', in *Rendiconti del Circolo Matematico di Palermo*, 37(1): 269-72.

Bose, R. C. 1938: 'On the application of the properties of Galois fields to the problem of construction of hyper-Graeco-Latin squares', in *Sankhyã*, 3: 323-38.

Bourbaki, N. 1968: *Elements of Mathematics, Theory of Sets*. Paris: Hermann.

Carter, L. J. 1974: *On the Existence of a Projective Plane of Order Ten*. Ph.D. thesis, UC, Berkeley, 1974.

Conrey, J. B. 1989: 'More than two fifths of the zeros of the Riemann zeta function are on the critical line', in *J. f. d. reine u. angew. Math.*, 399: 1-16.

Denjoy, A. 1931: 'L'Hypothèse de Riemann sur la distribution des zéros de $\zeta(s)$, reliée à la théorie des probabilites', in *C. R. Acad. Sci. Paris*, 192: 656-58.

Edwards, H. M. 1974: *Riemann's Zeta Function*, New York: Academic Press.

Hadamard, J. 1896: 'Sur la distribution des zéros de la fonction $\zeta(s)$ et ses conséquences arithmétiques', in *Bulletin Société Mathématique de France*, 14: 199-220.

Hardy, G. H. 1914: 'Sur la distribution des zéros de la fonction $\zeta(s)$ de Riemann', in *C. R. Acad. Sci. Paris*, 158: 1012-14.

Kline, M. 1990: *Mathematical Thought from Ancient to Modern Times, Vol. 3*. New York, Oxford: Oxford University Press.

Lam, C. W. H. 1991: 'The Search for a Finite Projective Plane of Order 10', in *American Mathematical Monthly*, 98(4), 305-18.

Lam, C. W. H., Thiel, L. H. and Swiercz, S. 1989: 'The non-existence of finite projective planes of order 10', in *Can. J. Math.*, XLI(6): 1117-23.

MacWilliams, J., Sloane, N. J. A. and Thompson, J. G. 1973: 'On the existence of a projective plane of order 10', in *J. Combinatorial Theory*, Sec. A. 14: 66-78.

Riemann, B. 1859: 'Ueber die Anzahl der Primzahlen unter einen gegebenen Grösse', in *Monatsberichte d. Berliner Akademie*, November 1859.

Tarry, G. 1900: 'Le problème des 36 officiers', in C. R. Assoc. Fran. Av. Sci., 1: 122-23; 2: 170-203.

Titchmarsh, E. C. 1986[2]: *The Theory of the Riemann Zeta-Function*, Oxford: Clarendon Press.

de la Valée-Poussin, Ch. J. 1896: 'Recherches analytiques sur la théorie des nombres premiers', in *Ann. Soc. Sci. Bruxelles*, 20: 183-256.

Veblen, O. 1904: 'A system of axioms for geometry', in *Trans. Amer. Math. Soc.*, 5: 343-84.

Veblen, O. and Bussey, W. H. 1906: 'Finite projective geometries', in *Trans. Amer. Math. Soc.*, 7: 241-59.

Veblen, O. and Wedderburn, J. H. M. 1907: Non-Desarguesian and non-Pascalian geometries', in *Trans. Amer. Math. Soc.*, 8: 379-88.

# Statistical and Non-Statistical Normality

Corina Strößner

Logics of normality which have been suggested so far interpret normality non-statistically. In this sense, something is normally the case if it is true in the most typical or plausible circumstances. On the other hand, one might think of normality as majority. In this sense something is normally the case if it is true in most circumstances. The paper argues that a descriptive explication of normality must imply statistical majority. Finally it introduces a formal logic of statistical normality and compares it to other approaches.

## 1. Principles of Normality Statements

"Normal" and "normally" are frequently used in natural language. We speak about normal days, normal members of species, normal properties of someone or something etc. In this discussion we define normality statements as nearly universal generalizations which allow for exceptions. Not any statement with the word "normal", "normality" or "normally" is a normality statement in the sense of this investigation nor is any of these expressions a necessary part of a normality statement. The word "normally" as a sentential modifier is a sure indication for a normality statement, e.g. in "Birds can normally fly" or "John normally sleeps long". However, statements like "Birds can fly", "A normal bird can fly" or "John is a long sleeper" mean often the same as a sentence with "normally" and should be regarded as normality statements too.

Normality statements are especially important to defeasible reasoning since they justify predictions which can be revised when more information becomes available. This property of normality statements can be explained by two different principles of normality statements: statistical justification and epistemic ordering.

### *1.1 Statistical Justification*

Why should one believe a normality statement like "Birds can normally fly"? Under which circumstances should we reject such a sentence? One suggestion is that the acceptance of normality statements should be justified by statistical reasons. Though the statistical reading is not popular in non-monotonic logic, even Reiter, one father of default reasoning, interpreted his defaults initially as statistical information: "Such facts usually assume the form 'Most P's are Q's'" (Reiter 1980: 82).[1]

The plausibility of statistical justification for normality statements in natural language is demonstrated by the obscurity of the following sentences:

(1) Normally, the train is on time; but mostly it is delayed.

(2) Bears are normally shy; but most bears are not shy.

Whoever utters these sentences makes a normality assumption against his statistical knowledge. This seems to be incoherent. He violates the following principle:

[1] Later Reiter strongly rejected this view. See (Reiter 1987).

**Principle of Statistical Justification:**

"If φ then normally ψ" can be understand as "If φ then mostly ψ".

### *1.2 Epistemic Ordering*

How does the acceptance of a normality statement change our beliefs? We will prefer the epistemic possibilities in which the statement is true over other world descriptions. As long as we have no information to the contrary we may expect that the world description which complies best with our normality assumptions is the right one:

**Principle of Epistemic Ordering:**

Accepting "If φ then normally ψ" is sufficient for an epistemic preference of ψ-worlds over ¬ψ-worlds among φ-worlds.

Consider the following example: You accepted *p*; *if p, then normally* $q_1$; *if p, then normally* $q_2$ and *if p, then normally* $q_3$. What should you believe? The worlds in which $q_1$, $q_2$ and $q_3$ hold are the most normal epistemic worlds. Therefore, they are preferred over other worlds. This is exactly the approach to normality chosen by Veltman in "Defaults in Update Semantics" (Veltman 1996) and Boutilier's "Conditional Logics of Normality" (Boutilier 1994).

It is out of question that these logics provide some insight into the use of normality assumptions. However, ordering seems to lead to a rejection of statistical justification. With respect to the Principle of Epistemic Ordering, the following inference is valid: *If p, then normally* $q_1$ and *If p, then normally* $q_2$ implies *If p then normally* $q_1 \wedge q_2$. But the corresponding statistical inference from *If p, then mostly q1* and *If p, then mostly* $q_2$ to *If p then mostly* $q_1 \wedge q_2$ is a fallacy. Take three cases in which *p* is true: Assume $q_1$ holds in the first and second of these cases, and $q_2$ holds in the second and third case. Most cases are $q_1$- cases and most cases are $q_2$-cases, but $q_1 \wedge q_2$ holds only in the second case. Single normality statements which are justified by statistical knowledge will not guarantee the statistical truth of the conjunctive statement.

A logical contradiction between ordering principle and statistical justification can be avoided if the requirement of statistical justification remains vague. Such step is taken for example by Adams (1974) in his analysis of "almost all" or recently with respect to counterfactuals and chance by Leitgeb (2012) and can be easily applied to normality as well. In this approach the probabilistic justification is applied to the conjunction of all normal predicates. That however makes it impossible to accept a sequence of normality statements *If p, then normally* $q_1$... and *If p, then normally* $q_n$ whenever *If p, then normally* $q_1 \wedge ... \wedge q_n$ is not justified. For a statistical reading of normality this is no plausible constraint. We feel hardly a need to check whether the conjunction of all properties we assign to bears in a normality assumption remains statistically justified when we accept a further normality statement. Thus, a statistical reading can hardly be part of the normality concept that complies with the idea of epistemic ordering.

## 2. Non-Statistical Normality Interpretation

The first step of a non-statistical interpretation can be made by stating that "normal" means the same as "typical" or "under idealized circumstances". Craig Boutilier understands his logic of normality in this way. He wants to "represent and reason with statements of normality or typicality, or default rules" (Boutilier 1994: 88). But is it plausible to interpret normality statements in such sense?

### *2.1 Stereotypes, Prototypes and Normality*

In his philosophy of language, most notably "The meaning of 'meaning'" and "Is semantics possible?" (both in: Putnam 1975), Hilary Putnam developed a view on semantics that is not focused on definitions, which determine the clear extension of a concept, but on stereotypes. According to Putnam, the stereotype is an oversimplified theory associated with the noun:

> [T]here is somehow associated with the word 'tiger' a theory; not the actual theory we believe about tigers, which is very complex, but an oversimplified theory which describes a, so to speak, tiger stereotype. It describes, in the language we used earlier, a normal member of the natural kind. It is not necessary that we believe this theory, though in the case of 'tiger' we do. But it is necessary that we be aware that this theory is associated with the word: if our stereotype of a tiger ever changes, then the word 'tiger' will have changed its meaning. (Putnam 1975: 148)

Obviously, sentences on typicality often appear as normality statements: "Tigers are normally striped" or "Normal tigers are striped". Stereotypes are, as indicated in the quote, about normal members. A stereotypical understanding of normality is based on the following thesis:

> **Stereotypical Interpretation of Normality**
>
> The sentence "S is normally P" is true iff P names an associated characteristic of the concept of S.

Putnams idea of stereotypes resembles in many respects the prototype semantics, best known by the work of Eleanor Rosch.[2] One main argument is that the recognition of membership depends largely on similarity to a typical member of the kind, i.e. a prototype (Rosch 1973), or the degree of prototypicality (Rosch 1978). It takes longer to recognize a non-flying bird (e.g. a penguin) as a bird than it takes to classify a flying bird (e.g. a robin) as bird. The degree of prototypicality is determined by the so-called cue validity which is defined as the conditional property that an entity x is member of S given it has property P: Prob(Sx/Px) (Rosch 1978). The definition of prototypicality in terms of cue validity has a peculiar relation to quantities: not the frequency of P in S but the frequency of S in P is critical; i.e. prototypicality is not about common attributes but about distinctive attributes. Understanding normality as prototypicality comes down to understanding normality in terms of specificity. The thesis of a prototypical interpretation of normality runs therefore as follows:

> **Prototypical Interpretation of Normality**
>
> The sentence "S is normally P" is true iff P names a characteristic which is specific for members of category S.

The main problem of such an interpretation of prototypes and stereotypes is that they serve other purposes than normality statements. They are especially important in recognizing membership from given properties. Normality statements, on the other hand, should help to predict properties from given membership. So, the question is how stereo- and prototypicality influence our expectations about the features a member of a category will have?

### *2.2 Communication Rules*

McCarthy (1986) suggestes a conventional interpretation of default rules when he names different ways to understand and apply non-monotonic reasoning: "As communication convention. Suppose A tells B about a situation involving a bird. If the bird cannot fly, and

---

[2] The term "prototypes" is first introduced in (Rosch Heider 1971) for colour terms. It is extended in (Rosch 1973) and further papers. Her psycholinguistics work became soon influential in philosophy of language in general. For current research on the prototype semantics see for example (Gärdenfors 2000) and (Taylor 2003).

this is relevant, then A must say so" (McCarthy 1986: 91). McCarthy emphasizes that such conversational rule "Assume S are P" doesn't presuppose any statistical facts about S and P. Statistical information cannot falsify a rule. Applied to normality statements this reflection yields another non-statistical normality interpretation, a conventional understanding.[3]

**Conversational Interpretation of Normality**

The sentence "S is normally P" is true iff it is assumed that any S is P as long as nothing else is mentioned.

Though it employs no reference to statistics this interpretation explains the expectations which can arise from normality assumption in a communication situation where all parties follow the same rules. Moreover this interpretation subsumes typicality interpretations. Stereotypes and prototypes, which are by definition commonly known in the community, will heavily influence our assumptions on what needs to be said explicitly and on what is suggested implicitly. This interpretation is also able to account for the fact that many normality assumptions are related to statistical majorities without postulating a relation. Assuming a rule which applies in many cases is *prima facie* more efficient than any other rule.

The conventional reading is the most adequate interpretation of normality from the non-statistical viewpoint. Nevertheless there are two serious objections against such a position which are deeply related to each other. First of all, normality statements, taken as conventions or typicality statements, are nearly trivial. Anybody who is a linguistically competent member of the community should know about prototypes, stereotypes as well as implicit communication rules. Only when you acquire the semantics and pragmatics of a language such normality statements are truly informative. This feature of the above normality interpretations is connected to another peculiarity, namely the non-descriptiveness of such normality readings. Understanding "Normally S are P" in these ways means to deny that such sentence is really about S. It is our language, i.e. our concepts and way of communication, that is expressed in this statement. Such conclusion might be acceptable in some cases but is counterintuitive as a general hypothesis on normality. There are a lot of normality statements, for example in biological science, which tell something about the nature of the subject term. A conventional or typicality account of normality cannot account for such statements.

### *2.3 Descriptive Normality*

There are admittedly descriptive approaches to normality which deny a strong logical relation to statistical statements. For example, Michael Thompson's (2008) characterization of natural-historical judgments indicates that these statements are normality statements but he strongly rejects that they involve majorities: "A natural-historical judgment may be true though individuals falling under both the subject and predicate concepts are as rare as one likes, statistically speaking" (Thompson 2008: 68). He understands these judgments essentially categorical and irreducible to any statement about individuals.

A somehow similar argumentation within the philosophy of biology is found in (Wachbroit 1994). Wachbroit claims that there is an indigenous biological concept of normality which is distinct from statistical normality. According to him a sentence about a normal biological

[3] Note that these communication rules are indeed conventions as defined by Lewis in his famous "Convention" (Lewis 1969). For the usage of default rules it holds that there is a common interest in conforming to the same rules. Assuming the same default will allow for a more efficient communication, while assuming different rules will lead to misinformation and confusion. A default which is sustained by large majorities is indeed better than a default which rarely applies but even a commonly used rule which is not related to statistical majorities is much better than non-conformity.

entity, or about what a biological entity normally does, is not necessarily related to a statistical majority:

> Suppose a calamity occurred in which most people's hearts failed to circulate blood so that they needed an implanted medical device for this purpose. This would hardly undermine the statement about the heart's function. (Wachbroit 1994: 580)

A problem of non-statistical accounts of descriptive normality is that they fail to indicate a way to test and falsify normality assumptions. This is exactly the problem which is addressed by Schurz (2001) who wants to provide those concepts with statistical consequences in order to make them testable by probabilistic considerations. He fleshes out the idea that such consequence can be ontologically founded by an evolution-theoretic argument.

Every natural or cultural event which involves variations, reproduction and selection is, according to Schurz, an evolutionary process. The foundation of such process is the reprotype. "Reprotype" is used as a very broad term which includes genotypes but also information on the production of artefacts. Schurz defines prototypical normality in the following way:

> T is a prototypical trait of S-members at time t iff T is *produced* by a reprotype R and from T's first appearance in the S-history until time t, there was *overwhelming* selection in *favor* of R. (Schurz 2001: 494)

He proves that his definition of prototypical traits ensures that prototypicality entails statistical majority. Schurz's arguments for the statistic consequence hypothesis are sound. But one might wonder if his definition is really non-statistical. He defines stereotypical traits in terms of "overwhelming selection". Though the statistical part is pushed on the side of the reprotype, it is obvious that Schurz's definition of prototypicality has a quantitative spirit. Hence, his arguments don't prove that statistics can be extracted from typicality. Rather it is shown that typicality can be founded on statistical normality. More generally, it is hard to see that any descriptive understanding of normality can abandon statistical justification.

As mentioned in the opening section, there is no logical contradiction between statistical grounding and non-statistical logic as long as the quantifying terms (in this case "overwhelming" and "most") are understood in a certain vague way: Almost every member of the class fulfils the basic traits so that, even if much less fulfil all of them, it is still enough to ensure a statistical justification even for the strongest consequences from the conjunction rule.[4] Though the compatibility of statistical justification and conjunctive closure is indeed attractive it raises the problem of a low acceptability of normality statements. For example in case of Schurz's statistical definition of prototypicality we expect that most evolutionary systems fulfil one certain basic prototypical trait but that only some have all of them.

What logic results if one gives less weight to ordering and deductive closure and more weight to statistical justification? The next section will present a logic which is mainly based on the statistical interpretation of normality

## 3. Logic of Statistical Justification

According to the principle of justification the acceptance of "If φ then mostly ψ" is necessary for claiming "If φ then normally ψ". What we need is therefore a formalization for "mostly".

### *3.1 "Most" and "Mostly"*

The semantics of the natural language term "most" is studied in the theory of generalized quantifiers (GQT) and is usually interpreted to mean "more than a half". For formalizing

[4] A formal theory in this spirit was presented by Leitgeb at the CLMPS 2011 in Nancy.

statements like “Most S are P” a binary quantifier can be added to a predicate logic.[5] In a first step, the set of individuals for which φ is true need to be defined:

*Def. 1:* $\varphi_{M,g,\chi}$ – the set of entities that fulfil φ – is the set of all individuals d in the domain such that $V_{M,g[\chi/d]}(\varphi)$ – the truth value of φ in model M under the assignment g[χ/d] in which χ denotes the individual d – is 1: $\varphi_{M,g,\chi} = \{d : V_{M,g[\chi/d]}(\varphi) = 1\}$.

This definition creates a set of things for which some open sentence is true. For example with respect to “It is human” we would put Socrates in the set since the statement is true if “it” refers to Socrates.

The set $\varphi_{M,g,\chi}$ is needed for a definition of a predicate logic with the binary quantifier MOST:

*Def. 2:* (MOSTφ)ψ is a well formed formula if φ and ψ are well formed formulae. $V_{M,g}((\text{MOST}\varphi)\psi) = 1$ iff $|\varphi_{M,g,\chi} \cap \psi_{M,g,\chi}| > |\varphi_{M,g\chi} - \psi_{M,g,\chi}|$.

The definition says that a statement “Most S are P” is true if and only if there are more individuals which fulfil “it is S” and “it is P” than individuals for which “it is S” is true but not “it is P”. The resulting logic PL+MOST is more expressive and, obviously, more complex than PL.[6]

For a statistical interpretation of a normality statement like “It normally rains in London” one needs to use “mostly” instead of “most”. Thus, we need an approach to the meaning of “mostly”. The adverb “mostly” is, according to David Lewis (1998), a quantification over cases where “a case may be regarded as the ‘tuple of its participants; and the participants values of the variables that occur free in the open sentence modified by the adverb” (Lewis 1998: 10).We accept that “mostly” refers to cases. However, as Lewis admits, there are numerous statements with “mostly” in which further variables, e.g. time points or different circumstances, need to be introduced. Therefore, we rather use possible worlds to represent alternative cases. We apply the semantics of MOST in a propositional modal logic where the binary operator MOSTLY doesn’t refer to individuals which fulfil an open sentence but to possible worlds in which a proposition holds. The formal definition of MOSTLY runs therefore in the same way as the definition for the quantifier. First, we define an expression which denotes all φ-worlds:

*Def. 3:* Let φ be a well formed formula. $[\varphi]_M$, the set of worlds that fulfil φ in the model M, is the set of every possible world w such that $V_{M,w}(\varphi)$, the truth value of φ in w with respect to the model M, is 1: $[\varphi]_M = \{w : V_{M,w}(\varphi)=1\}$.

Using this definition we determine the semantics of “mostly” in a modal logic framework. Modal logics usually work with an accessibility relation. In the definition of MOSTLY we need therefore to make a restriction to the set of accessible worlds from the actual world w, i.e. to $\{w' \in W : wRw'\}$:

*Def. 4:* (MOSTLYφ)ψ is a well formed formula if φ and ψ are well formed formulae. $V_{M,w}((\text{MOSTLY}\varphi)\psi) = 1$ iff $|\{w' : wRw'\} \cap [\varphi]_M \cap [\psi]_M| > |\{w' : wRw'\} \cap [\varphi]_M - [\psi]_M|$.

This definition can be used to extend any propositional modal logic ML to a more expressive logic ML+MOSTLY.

[5] There are different ways to approach quantifying determiners in a formal way. An extensive overview on the history of quantifiers and the developments of GQT is given in (Peters and Westerstahl 2006).

[6] For example the application of MOST to infinite domains is rather tricky. However, for the interpretation of normality statements this is less problematic since we usually apply normality statements in finite domains.

### *3.1 A Statistical Logic of Normality: SN*

The given formal characterization of "mostly" can be applied to determine a statistical logic of normality SN. Sentences of the structure "If _ then normally _" (formally: _ ~> _) will be understood as "if _ then mostly _". In the following definition of SN the previous explication of "mostly" is employed in the determination of φ ~> ψ. Apart from that SN resembles the well-known modal logic S5.

*Def. 5:* SN is a modal logic determined by the following definitions:

– Formulae of SN are the atoms p, q, r... as well as ¬ φ, φ ∧ ψ, □ φ and φ ~> ψ if φ and ψ are formulae of SN.

– A model M is a triple ⟨W,R,V⟩, containing a set of possible worlds W, an accessibility relation R on W that is an equivalence relation, and a valuation V for each possible world w assigning the truth values 1 or 0 to atomic formulae of SN.

– The formula φ is true in Model M at world w ∈ W – formally written: M, w |= φ – according to the following clauses:

M,w |= φ iff $V_w(\varphi) = 1$ where φ is an atomic formula,

M,w |= ¬φ iff not M,w |= φ,

M,w |= φ ∧ ψ iff M,w|=φ and M,w |= ψ

M,w |= □ φ iff M,w′ |= φ for all w′ ∈W such that w′Rw,

M,w |= φ ~> ψ iff $|\{w':wRw'\}\cap[\varphi]_M\cap[\psi]_M| > |\{w': wRw'\} \cap [\varphi]_M - [\psi]_M|$, where $[\varphi]_M = \{w' \in W : M, w \models \varphi\}$.

– The inference from $\varphi_1, \varphi_2, ..., \varphi_n$ to ψ is valid – formally written: $\varphi_1, \varphi_2, ..., \varphi_n \models \psi$ – iff for every M at every world w∈W it holds: if M,w |= $\varphi_1$, M,w |= $\varphi_2$ ... and M,w |= $\varphi_n$, then M,w |= ψ.

– The formula φ is a tautology – formally written: |= φ – iff for every model M at every world w∈W it holds: M,w |= ψ.

An operation for unary normality statements is useful: *normally* ψ abbreviates (φ ∧ ¬ φ) ~> ψ: Since φ ∧ ¬φ is a tautology we refer at all worlds. Thus *normally* φ is true if and only if φ is true in most worlds.

## 4. SN and Probability

Veltman's default logic has, besides the normality conditional, the additional unary operator *presumably*. This expression is of vital importance to his logic. It expresses the epistemic state of an agent without adding any new information. Roughly speaking, the statement *presumably* φ must be accepted if the most normal possible worlds which are not excluded by factual information are φ-worlds. This allows to formalize *normally* φ |= *presumably* φ and *normally* φ, ¬φ |≠ *presumably* φ. There is, by now, no way of rendering such defeasible conclusions in SN. To gain this possibility the quantitative statement must be related to probability.

The popular example from logic books "All men are mortal. Socrates is a men. Therefore Socrates is mortal" is without doubt a valid inference. "Most men are right-handed. Plato is a man. Therefore, Plato is probably right-handed" is not logically valid in the classical sense but it is plausible: The information that Plato is a man and that most men are right-handed confirms the assumption that Plato is right-handed more than that he is left-handed, at least as long as no more specific information is given.

### *4.1 Carnap's Logical Probability*

In "Logical Foundations of Probability" (Carnap 1962) Carnap fleshes out the concept of logical probabilities. In his theory probability is used as a measure of the degree of confirmation of a hypothesis by given evidence. In a first step, probabilities are assigned to state descriptions, which give a complete characterization of a possible state of affairs. The probability measure m of a sentence is the addition of the probability of all state descriptions to which the sentence belongs. Finally, the confirmation c(e/h) of hypothesis h by evidence e is defined as conditional probability $m(e \wedge h)/m(e)$.

Carnap's theory of probabilistic confirmation is twofold. One part is reasoning from the entire population to a smaller sample: the direct inference. Another part is the reasoning from one sample to the population or to another sample. We will restrict our considerations to the direct inference. Let c be a confirmation function for the evidence e and the hypothesis h. Confirmation functions for direct inferences need to be regular and symmetric (Cf. Carnap 1962: VIII):[7]

Def. 5: For every regular and symmetric confirmation function c(h/e) it holds:

– Regularity: c(h/e) = 1 iff e |= h. The best possible confirmation is entailment.

– Symmetry: c(h/e) is defined as conditional probability $m(e \wedge h)/m(e)$ on a measure that gives the same value to state descriptions which do not differ in the number of individuals that are in the extensions of all possible predicates.

Now, let us assume r exclusive and exhaustive predicates $M_1$, $M_2$, ..., $M_r$ and the following statistical distribution for the population of n individuals: $n_1$ individuals are $M_1$, $n_2$ individuals are $M_2$, ..., and $n_r$ individuals are $M_r$. This is our evidence. Our hypothesis for a sample of s individuals is that $s_1$ individuals will be $M_1$, $s_2$ are $M_2$, ... and $s_n$ individuals are $M_n$. Then, as shown by Carnap, every regular and symmetric confirmation satisfies the following equation:

$$c(h/e)=\frac{(n-s)!}{(n_1-s_1)!(n_2-s_2)!\ldots(n_r-s_r)!}\times\frac{n_1!n_2!\ldots n_r!}{n!}.$$ (Carnap 1962: 495)

If we restrict ourselves to two exclusive and exhaustive predicates (e.g. being right-handed and not being right-handed) and a sample of only one individual (e.g. Plato), the equation above can be simplified:

$$c(h/e)=\frac{(n-1)!}{(n_1-1)!(n-n_1)!}\times\frac{n_1!(n-n_1)!}{n!}=\frac{n_1}{n}.$$

This shows that, as one would expect, the confirmation of the hypothesis that an individual will have a property is identical to the relative frequency of the property in the population.

### *4.2 MOST, MOSTLY and Probability*

The semantics of "most" in "Most S are P" is expressible by the number of S which are P and the number of S which are not P. In a finite universe "Most S are P" translates to a disjunction of statistical distributions, which have the form "$n_1$ S are P and $n - n_1$ are not P". The disjunction contains all statistical distributions in which $n_1$ is greater than $n - n_1$. All of these distributions confirm the proposition that some arbitrary individual S is P better than the statement that this individual is not P. That means that the thesis that an individual which is S (e.g. the man Plato) has a property (being right-handed) is always better confirmed by the evidence that most individuals of that kind have this property than a contradictory

[7] Note that this symmetry is related to individuals only. The rather problematic symmetry for Q-predicates, predicates characterizing an individual completely, is not required for the results on direct inference.

hypothesis. *If most men are right-handed and Plato is a man then it is probable that he is right-handed* is convincing as long as no other evidence is given.

This argument can be applied to formulae of SN in finite models. The SN formula $\varphi \sim\!> \psi$ corresponds to the following metalogically formulated disjunction:

$$\bigvee_{2n_1>n} |\{w':wR'\}\cap[\varphi]_M\cap[\psi]_M|=n_1\wedge|\{w': wRw'\} \cap [\varphi]_M - [\psi]_M| =n-n_1$$

Every of the statistical distributions, together with the evidence that φ holds in w, confirms the hypothesis that ψ holds in w to a higher degree than 0,5.[8] Hence, in terms of Carnap's confirmation c the following holds for SN formulae in any finite model:

$c((\varphi \sim\!> \psi) \wedge \varphi) > 0,5$

### *4.3 A sketch of Probabilistic Inferences for SN*

One can easily use the preceding results on confirmation to define a probabilistic inference that can be applied to SN formulae. The probabilistic argument should be valid if the confirmation of the conclusion by the premises is higher than 0,5 or undefined (because the evidence is inconsistent) in every finite model. The probabilistic inference should also be valid if confirmation remains undefined since any conclusion is logically implied by inconsistent premises. In order to ensure that every logically valid inference is also probabilistically valid it is stipulated that the probabilistic inference is valid in this case as well.

*Def. 6*: $\psi_1, \psi_2, ..., \psi_n \mid\!=^{prob} \varphi$ iff $c(\varphi/\psi_1\wedge\psi_2\wedge...\wedge\psi_n) > 0,5$ or c is undefined for every c in a finite domain.

The results we get by the application of the probabilistic inference to SN include:

(1) If $\psi_1, \psi_2, ..., \psi_n \mid= \varphi$, then $\psi_1, \psi_2, ..., \psi_n \mid\!=^{prob} \varphi$

(2) $\varphi \sim\!> \psi, \varphi \mid\!=^{prob} \psi$

(3) $\varphi \sim\!> \psi, \varphi, \neg\psi \mid\!\neq^{prob} \psi$

(4) $\varphi_1 \sim\!> \psi, \varphi_1 \wedge\varphi_2 \sim\!> \neg\psi, \varphi_1 \wedge\varphi_2 \mid\!=^{prob} \neg\psi$

## 5. Statistical and Non-Statistical Logic of Normality

The four results for the application of probabilistic inferences to SN comply with results for the operator *presumably* in (Veltman 1996): [9]

(1) If ψ1, ψ2, ..., ψn |= φ, then ψ1, ψ2, ..., ψn |= presumably φ

(2) $\varphi \sim\!> \psi, \varphi \mid=$ presumably ψ

(3) $\varphi \sim\!> \psi, \varphi, \neg\psi \mid\!\neq$presumably ψ

(4) $\varphi 1 \sim\!> \psi, \varphi 1 \wedge \varphi 2 \sim\!> \neg\psi, \varphi 1 \wedge\varphi 2 \mid=$ presumably ¬ψ

[8] One needs to make the additional background assumption that wRw, i.e. that R is reflexive. This is, however, guaranteed by the fact that the accessibility relation of SN is an equivalence relation.

[9] (1) is rather trivial: ψ1, ψ2, ..., ψn |= φ holds iff φ is true in all worlds which are not excluded by the premises. This implies that φ is true in the most normal worlds which are not excluded by the premises, which is sufficient for presumably φ. In (2) ψ-worlds are preferred among the φ-worlds. Since all non-φ-worlds are excluded all preferred worlds are ψ worlds. In (3), however, ψ-worlds are excluded and count no longer as the most normal worlds that are not excluded. Result (4) shows that the most precise normality statement overrules the less precise normality statements. The background of this rule is rather sophisticated. The technical details are explained in (Veltman 1996: 253ff).

Veltman's results, however, are not grounded on probabilistic considerations which follow from a statistical reading of normality. His logic yields predictions because his normality conditionals order epistemic possibilities. Thus, both principles of normality, statistical justification and ordering, can be logically used to yield defeasible predictions. But, as one could expect considering the different basic principles which are used in both approaches, there are some differences in the two different accounts of normality.

The following examples give a final overview of some commonalities and differences of SN and Veltman's system.[10]

| | | | |
|---|---|---|---|
| (1) | Contrariety I: | $\textit{normally}\ \psi, \textit{normally}\ \neg\psi \models \bot$ | Veltman & SN |
| (2) | Contrariety II: | $\neg\ \textit{normally}\ \psi, \neg\ \textit{normally}\ \neg\ \psi \not\models \bot$ | Veltman & SN |
| (3) | Conjuntion: | $\varphi \sim\!> \psi, \varphi \sim\!> \chi \models \varphi \sim\!> (\psi \wedge \chi)$ | Veltman |
| (4) | Weakening: | $\Phi \sim\!> \psi \models \varphi \sim\!> (\psi \vee \chi)$ | SN |

The first two lines show that normality statements in both systems are strong enough to have contrary oppositions. At most one of two exclusive propositions can be normally the case. The difference in the rule of conjunction is the most important one: If you accept typicality statements you must accept the conjunction of them as a normality statement. The same will not hold for statistical data. Weakening of the consequent with an arbitrary statement is not valid in Veltman's default logic. The disjunction should give two normal alternatives. This seems to be plausible for the use of normality statements in natural language. However, the invalidity of weakening is a rather unique feature of Veltman's approach. Other logics with ordering, e.g. Boutilier's "Conditional Logics of Normality", allow weakening. SN does not require relevant alternatives in disjunctions, either, and leaves the problem of relevance to pragmatics.

**Corina Strößner**

Universität Konstanz
corina.stroessner@uni-konstanz.de

## References

Adams, E. W. 1974: 'The Logic of "Almost All"', *Journal of Philosophical Logic* 3, 3 – 13

Boutilier, C. 1994: 'Conditional Logics of Normality: a Modal Approach', *Artificial Intelligence* 68, 87-154

Carnap, R. 1962: *Logical Foundations of Probability*. Chicago: The University of Chicago Press

Gärdenfors, P. 2000: *Conceptual Spaces: The Geometry of Thought*. Cambridge (MA): MIT Press

Leitgeb, H. 2012: 'A Probabilistic Semantics for Counterfactuals. Part A', *The Review of Symbolic Logic* 5, 26-84

Lewis, D. 1969: *Convention. A Philosophical Study*. Cambridge (MA): Harvard University Press

[10] For contrariety I see (Veltman 1996: 17). For conjunction and weakening see (Veltman 1996: 37). Strictly speaking, Contrariety II cannot be stated in TN because *normally* can only occur as main operator in Veltman's system. But it is quite obvious that it is coherent in Veltman's semantics not to believe one of the two normality statements.

McCarthy, J. 1986: 'Applications of Circumscription to Formalizing Common-Sense Knowledge', *Artificial Intelligence* 28 (1), 89-116

Peters, S., and Westerstahl, D. 2006: *Quantifiers in Language and Logic*. Oxford: Clarendon Press

Putnam, H. 1975: *Philosophical Papers, Vol 2: Mind, Language and Reality*. Cambridge: Cambridge University Press

Rosch Heider, E. 1971: '"Focal" Color Areas and the Development of Color Names', *Developmental Psychology* 4, 447-455

Rosch, E. 1973: 'Natural Categories', *Cognitive Psychology* 4, 328-350

— 1978: 'Principles of Categorization', in E. Rosch, and B. Lloyd (eds.): *Cognition and Categorization* 1978, Hillsdale NJ: Lawrence Erlbaum Associates, 27-48

Schurz, G. 2001: 'What is "Normal"? An Evolution-Theoretic Foundation for Normic Laws and Their Relation to Statistical Normality', *Philosophy of Science* 68, 476-497

Taylor, J. R. 2003: *Linguistic Categorization*. Oxford: Oxford University Press

Veltman, F. 1996: 'Defaults in Update Semantics', *Journal of Philosophical Logic* 25, 221-261

# 4. Philosophie des Geistes

# Theory of Mind as Gradual Change Guided by Minimalisms

Gerhard Chr. Bukow

The paper assumes that an important ability of Theory of Mind is the ability to follow changes in thoughts. Agents are assumed to develop theories and schemas about thought-following and they do apply different constraints on such theories like minimalisms. However, what does it mean to follow and exactly "what" is it that agents follow? How are changes in agents and changes in domains related? By considering thought-following in the domain of qualitative spatial reasoning in cognitive psychology, different minimalisms and schemas of thought-following are discussed. Finally, the paper considers the consequences of the approach for our view on Theory of Mind in philosophy and psychology.

## 1. Qualitative Spatial Reasoning and Theory of Mind

Consider the following situation: two engineers work on the problem how to place some very huge elements such that they fit into a given environment and fulfil a certain function. Think about huge mechanical devices for a ship or so. They use some special cars or so that enable them to move the devices. However, both of them cannot directly see what the other can see. From their own point of view, they must reason to take the perspective of the other. Even if they could directly see what the other can see, they must reason with respect to the function the elements should do. In the situation, the elements are first placed wrongly so they must move them step by step. And step by step, each of them must reason what the other sees and consequently does in order to move the elements adequately. To sum up: one person has to "follow" or "track" the other person step by step. But how is this tracking done, what does perspective-taking mean in the sense of agents that track each other? Let us explore this.

Cognitive psychology has explored in detail how humans qualitatively reason about spatial relations. The best known approach (in terms of successful prediction and explanation in terms of cognitive architectures) uses conceptual neighborhood graphs to analyze the gradual change of spatial relations between objects. Such graphs allow us to say something about the conceptual distance between different situations describing the objects' spatial relations. And it allows us to say what (perhaps single) operations are needed to get from one situation to another one. In the figure below you can see such a graph and the different possible "moves" two objects could do (the original context was the use of such a graph to say something qualitatively about the similarity of two spatial situations).

In each case, the one worker could (at least in theory) say how much distance between two situations is, or how much distance there is between "his own" situation and the situation of the co-worker. Of course, he does not so by directly "seeing" the others mental representations, for example the mental models representing the spatial situation. But he can make a plan how he would gradually change the situation such that the other can see it.

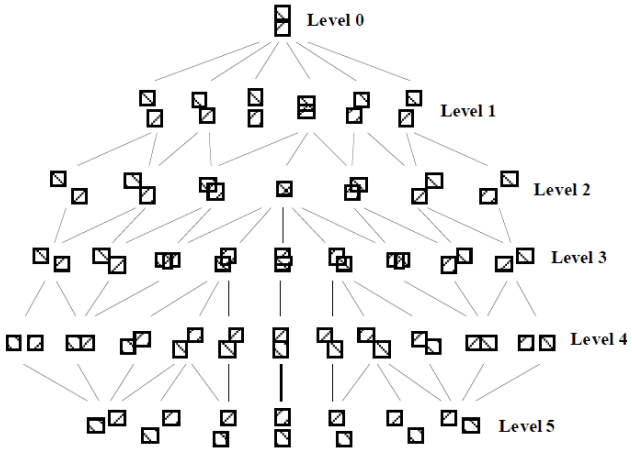



**Figure 7**: Similarity network.

Figure 1. Gradual change on a conceptual neighborhood graph given by an example that judges the similarity between two spatial scences by Bruns & Egenhofer (1996).

Now let us make that issue more adequate to our environment with respect to the change of information. Again, consider that humans reason about spatial relations. Perhaps, one worker gets the information that one container is so-and-so positioned and that the other containers are so-and-so positioned. However, the other worker then informs him that the last information was wrong – he should revise his beliefs about the situations. This has been investigated experimentally for example in such very simple and controllable settings like the one below.

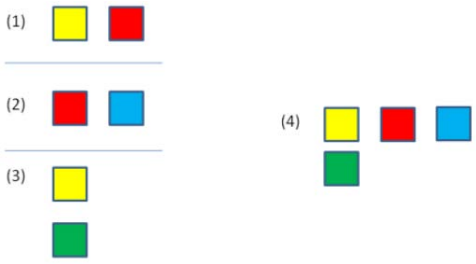



Figure 2. Based on (1), (2), and (3), the (4) is constructed.


Figure 3. Left: After having constructed (4), the figure (a) (or the figure (b)) will be presented to you. Now (4) has to be varied. Right: the variation has succeeded in integrating (a) (the right figure) or (b) (the left figure).

Now, put these two issues together: people have to reason about possible steps gradually (for example in spatial relations of objects) and people have to reason what other people reason about this issue. This is just the dynamic application of theory of mind – namely perspective taking – in contexts of spatial relations in this example.

Of course, such situations are not restricted to spatial aspects. Consider two philosophers reasoning about a problem. However, philosophers are strange people and while one philosopher does accept Modus Ponens, the other philosopher does not. He does not accept Modus Ponens because there is no final legitimization for Modus Ponens in research yet. So,

the philosopher that accepts Modus Ponens strongly investigates in a chain of thoughts he could present the other one – such that gradually, thought after thought, the other will finally accept Modus Ponens. In a sense, the philosopher "measures" the distance between both of them and how much argument he must give to convince the other one. There are surely more problems like this one.

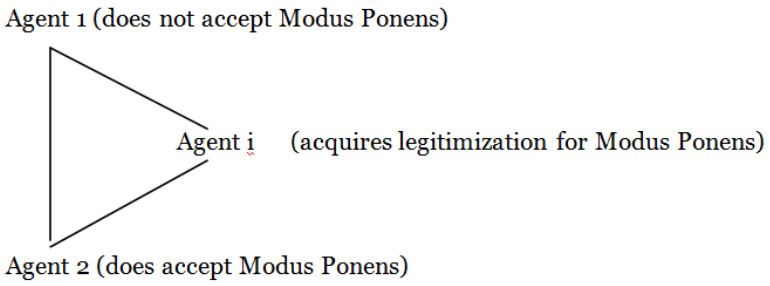



Figure 4. The figure shows an example for a difficult, but quite usual gradual change of agents doing science. There may be no best straight forward route from Agent 1 to Agent 2 due to multi-criteria driven solutions, multiple evidence, etc.

We handled synchronic problems. But the issue can also be seen diachronically: you may just think about one philosopher thinking about what he must argue to convince himself – in a sense, this philosopher plans his own epistemic development. We also handled the problems as if they were always conscious or as if philosophers would really do so. Of course, this is not the case. Nobody must propose that rational agents always consciously apply theory of mind (though some of them can do it, in fact). However, what is commonly proposed is that we (or every other agent) should do so in a "minimal way" – because we just are limited (and every other actual agent, too.)

## 2. Minimalisms

In both theory of mind and rationality theory, there are at least two issues of minimalism. First, such ideals like simplicity or minimalism are often called for. Gradual change shall be in accord with minimalism. For example, the change of belief shall be economically done and minimal changes are the result (for example belief revision or mental model variation). This leads to the second more general point, because it is argued that such a minimalism is because actual agents are not god-like. There may be actual agents that are not "maximally rational", or do not have unlimited computational resources, or with restricted reasoning domains, etc. The following two tables give a rough overview about different forms of such minimalisms.

Table 1. Some minimalisms in rationality theory.

| **Label** | **What is to be minimized?** |
|---|---|
| Computational limits (Cherniak (1986)) | Capacities of memory and processing speed |
| Denying logical omniscience | Capacity of actually (or potentially) being able to believe all the consequences of one's beliefs (not only due to computational limits) |
| Ecological rationality in a Gibsonian style | Minimizing agent's internal structures, maximizing the external environmental structures |

| | |
|---|---|
| Probabilistic rationality in a Bayesian style | Bayesian approaches focusing on prediction of rational behavior (while neglecting explanation) |
| Blind spots | Refraining from the thesis that all meaningful beliefs can be believed truly (or that all meaningful beliefs just are true ones) |
| Domain-bounded inference | Minimizing the potential content of inferences to just some contentual domains – that is "inferring" via content and not via form |
| Limited core principles of cognitivism | Minimizing systematicity and compositionality (against classic assumptions of cognitivism dealing with unlimited forms of these features) |

Table 2. Minimalisms in theory of mind

| **Label** | **What is to be minimized?** |
|---|---|
| Syntactical agents (Butterfill & Apperly (2011)) | Semantics comes for free – semantics is minimized to syntax |
| Computational limits of syntactical thought-following (Alecchina& Logan (2010)) | Following an agent from its conclusion (plus the way to it) to the next conclusion (plus the way to it) is minimized to just following from conclusion to conclusion |
| Limited access to concepts | Agents only have access to some but not all concepts, for example only to perceptual concepts (as it is worked out in the literature of theory of mind of animals) |
| Simple concepts | Complex concepts are breaked down to concepts with less granularity, details, connections to other concepts etc. |
| Domain-bounded understanding | Understanding is only possible in some domains or niches |
| Protologic (Bermudéz (2003)) | Inference is driven by contents of evolutionary adapted domains (ecological validity) |

## 3. The Measurement-combination Problem

Again, let us put the things together: theory of mind is applied to gradually "reach" another agent. For example, one worker applies theory of mind and initiates a model from his own position to reach his co-worker's model mentally. Minimalism shall guide the gradual change, or even stronger, the agent may be somehow minimal with respect to his reasoning. Now, one can argue that one minimalism just is enough. For example, the AGM model of belief revision just applies minimal change principle. Or, scientific theories shall be simple. One could argue, the application of theory of mind is a mono-criteria problem.

But this does not seem to be the case: often, different minimalisms are somehow combined. But then, the application of theory of mind is essentially a multi-criteria problem! For example, we could describe our workers by very different minimalisms in parallel: limited computational resources, limited reasoning domains, applications of different simplicity-measures in their perception and spatial reasoning, or there may be several blind spots of reasoning.

**Multi-criteria problems** usually do not have simplified one-value-solutions like "Just let consistency govern the process of change." Or set another value instead of consistency, for example: simplicity, coherence, or truth. And there may be not *the one and only* minimalism and *the one and only* combination of minimalism. This leads us to the measurement combination problem: how can and should we combine different minimalisms?

There are two commonly known possible solutions to such a problem. First, develop a unified theory such that all measures come from one theory and you can say how each measure is related to the other one. But there is no such unified theory and we do not know how to construct it. The other possible solution is to develop a theory about the relations between different theories of cognition. There are some proposals about relations, for example in relational data base theory, but nothing we could use here.

Another problem is that it is difficult to conceive how measures of rationality could fulfill what measures should fulfill. Just think about additivity.

**Additivity**: M (a) + (b) = M (a+b)

It is not at all clear why additivity should hold *if* a and b come from different epistemic theories. What M(a) and M(b) do deliver are only seemingly "numbers" that just wait to be added up. We always have to care for the way and history of the numbers, they are theory-laden so to say. If the numbers are delivered by different theories, finally, the combined semantics of the number M(a+b) is not just given by addition. Still, semantics for measurements is not automatically extensional so that we could just apply a union to two sets of semantics (that is the semantics of M(a) and the semantics of M(b)).

So, where do we stand? Theory of mind has been considered as gradual changing of rational agents. Gradual change shall be guided by (the combination of different) minimalism(s). But to do so we must have a solution for the measurement combination problem. Of course, this solution must respect the ABC of actual agents (the next table).

Table 3. The ABC of actual agents.

| **ABC** | **Content** |
|---|---|
| Agent | Agents as structured entities with theories (c.f. Gopnik (1988)) |
| Belief | Features of belief systems and the representational format |
| Cognition | Effects of cognitive significance like transparency, opacity, reference, scope of belief operators, indexicals, globalism/localism, ... |

This means that our measure method should fulfill some important requirements concerning actual agents. I argue for these four points.

**Change**: Change of beliefs, for example revision. This placeholder is legitimated because the ability to "follow" other agents over changes is a typical theory of mind-ability. Furthermore, as research in cognitive psychology has shown, the underlying representation formats and other cognitively significant features are relevant for change (concerning the ABC).

**Relations**: Derivation of norm systems from norm systems, or relations between involved theories. In general, agents will accept a norm system with respect to a given task at hand, and this will also involve theories. Note, that the relation between involved theories is not the same (big) task as a general theory of relations between theories.

**Standards of formation, proof and inference**: The agent will accept certain standards of belief formation, proof types, types of inference, evidence standards, or in fundamental way specific evidence.

**Mapping between cognition and norms**: Every pair consisting of accepted norms and a given problem should be mapped to the cognitive level (consisting of representation formats, mechanisms, etc.).

The same placeholders – now filled up for our situation of two workers from above:

**Change**: Principles of model variation (especially, if the AGM-framework following Alchourròn, Gärdenfors, & Makinson (1985) is suitable for this task, what we can deny generally)

**Relations**: Principles of spatial relations in calculi (naïve geography, or approaches like region connection calculus)

**Standards of formation, proof and inference**: Inferences like Modus Ponens or the specific belief transformation norm "Believe that p, iff p." (Standards of belief formaten are due to the correspondence theory of truth accepted by Johnson-Laird's mental model theory.)

**Mapping between cognition and norms**: The cognitive level is reflected due to the algorithms that are given for problem solving processes and how they work on specific representation formats. The "output" delivered by these algorithms can be projected onto the representation level of belief systems (in a many to one manner, because of course the problem is solvable in many different ways or by many algorithms).

## 4. The Third S0lution: The Application of Method Schemas in Agent Theory

A third solution proposes that we could have a theory of measurement methods. That is, we theorize how agents measure other agents in synchronic situations or themselves in diachronic development. Note, this does not mean to propose a theory of how we just have a look at other agents and directly see what is going on there.

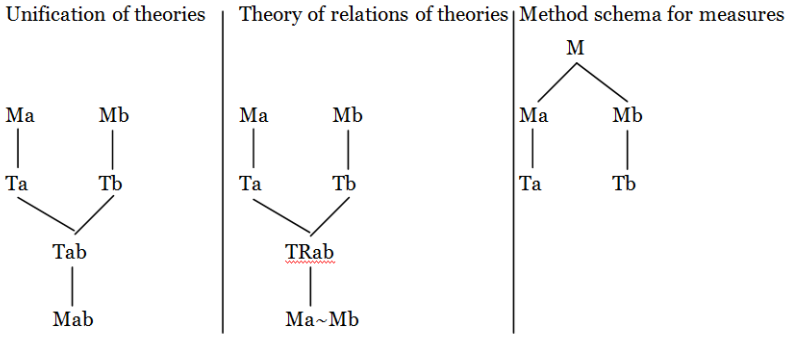



Figure 6. Three potential solutions to the measurement combination problem.

The first step towards such a theory makes an assumption concerning the conceptualized structure of agents. In the following, I will take agents to be structured by theories in accord with the theory-theory-approach and its conceptual reference to scientific theory. However, I think, a completely parallel investigation could be done by considering simulations. As we will see, this does not matter with respect to the principles of the suggested approach.

The second step however loans some concepts from the new science of method engineering. Method engineering (following Ralyté (2002)) considers how we could construct methods that do construct methods. For example, what methods do construct measurements that can

be used to determine the distance between two rational agents (think about the two workers "measuring" each other)? Such methods finally provide blueprints or schemas that can be filled up by agents. They deliver a classification of the agent.

Typically, first a map is constructed consisting of intentions and strategies how to reach such intentions. Every intention is the state of another agent (or oneself in future). Every strategy is the strategy of how to change or develop until one reaches the other agent. So let us map this in the following picture.

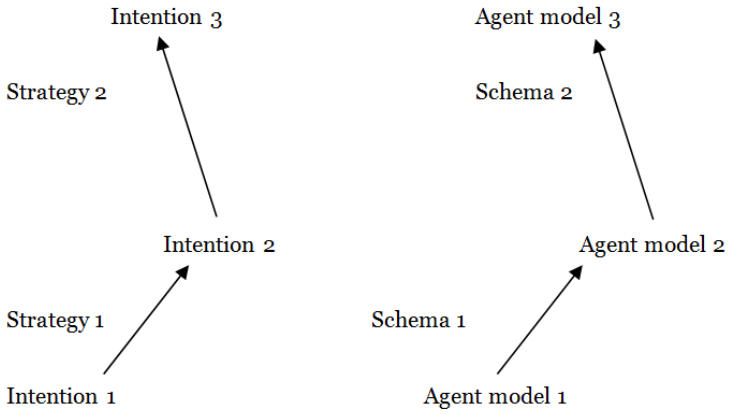



Figure 7. Intention map (intention, strategy) and Schema map (agent model, schema-development).

With respect to the ABC of actual agents and the four requirements, I propose the following "enriched" schema. It provides placeholders for commitments to theories and consists of a core and several extensions. Note, this is not a concrete schema or measurement – it says what agents have to take into account if they measure and where they can make changes to reach another agent model.

Within such a schema, different minimalisms can be combined in accord with different theoretical commitments. This combination is not mysterious, but it is the application of methods to get solutions for a multi-criteria-problem. Whatever concrete solution this may be, there cannot be in principle the one combination preferred over all other combinations.

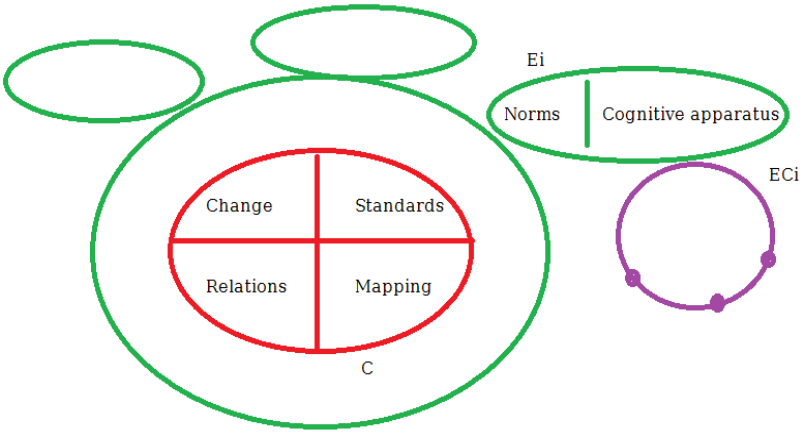



Figure 8. The "placeholders" of the suggested enriched measurement schema.

In a structural core (red, C) there is place for different theoretical commitments to the four already mentioned features. So, the core also contains norms or standards (or a system of norms). From these norms, extensions can be derived (green, Ei), whereas the norms in the extensions are grounded by the norms in the core. Additionally, every extension Ei also has

cognitive elements (lila, EiC), which are element of the pairing of a given core + extension and the task at hand. Because the meta-schema already respects change, a timeline is integrated (that is every concrete measure will also make measures at a time point).

We can even do more: we can consider the concrete scene and its corresponding concrete mental model, and its corresponding agent, and the agent model behind it with its applied schema. We can do so by abstracting from the concrete scence.

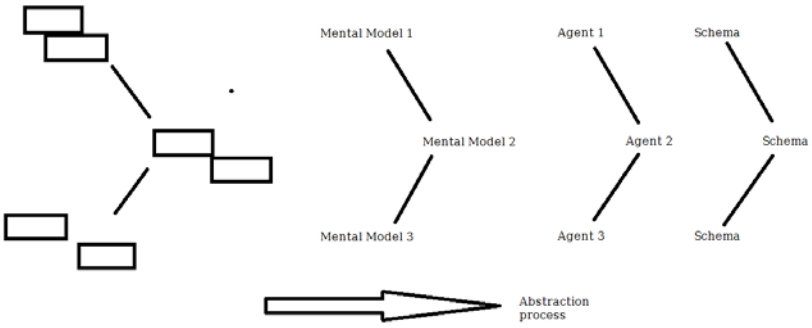



Figure 9. Abstracting from gradual change in spatial scenes to mental models to agents to schemas (or agent models). Because I do not explore change of schemas themselves in this paper, I label them just "schema", though these schemas may be different ones.

## 5. Agent Models as Nodes

Now let us make the most important step: **now, we consider every node of the conceptual neighborhood graph as a positioned agent model (characterized by its schema).** It is positioned in accord with its measurement methods, and the task or problem at hand that has to be solved. This finally gives us a glimpse what "reach" does mean: it means to develop along the conceptual neighborhood graph guided by the schema applied to classify agents which itself is guided by a combination of minimalisms. This also allows us to discuss several possible cases of distance measurement with respect to the cores and extensions in the schema.

Table 4. Distances between agent models and cores/extensions.

| **Distance** | **Both cores are … / Both extensions are …** |
|---|---|
| Trivial case: "0 distance" | Both cores and extensions are "identical" |
| Easy case: "Small distance" | Equal / equal |
| Hard case: "Big distance" | Equal / different |
| Very hard case: "Very big distance" | Different / different |
| Hardest case: "Infinite distance" | God - nothing can be known about unreachable cores |

Agent models will have their place as nodes in the graph. Between these models, distances shall be measured. As one could see, this implicates something like a metric that allows us to measure distances that can be used then to define minimal rationality, minimal understand, minimal theory of mind etc. But as one could also see, this is essentially a multi-criteria-problem. There will be not only one minimalism as the label "Minimal theory of mind" suggests. Instead, there will be many minimalisms in parallel – just depending on what minimalisms we accept and want to combine. This means that our "operationalization" of distance can be very different. For example, one could (in a classic fashion) measure distance

by the number of (or weight of) operations like "adding norms" needed to reach another agent model, but one could also try to combine this "operation-cost" and costs in terms of how many domains or niches of knowledge must be added. There is no final answer concerning this issue – it essentially depends on how we argue in the debate about methods and solutions of multi-criteria-problems.

From the discussion of gradual change and its related issues like distance and understanding, I first want to give some definitions grounded in the discussion: definitions of minimal rationality, theory of mind, minimal theory of mind, and minimal understanding. Some figures will illustrate how different minimalisms lead to different preferred gradual change, if one prefers minimal change. Then, I want to show what further issues can be explored in this framework in future work.

An agent X is named "**minimal rational**" from the position of another agent Y in accord with this agent's specific distance definition (resulting from the combination of minimalisms) etc. **Ineffectiveness**: if X is either suspected to take an ineffective route of gradual change to Y. **Asymmetric change abilities**: if Y can specify the route of gradual change to X but not vice versa, i.e. X (presumably) cannot specify the route of gradual change to Y. A set of minimal rational agents then can be given by minimal distance (from X): all the agent models that can be reached within minimal distance starting from the current agent (or agent model), but not vice versa. For this reason, minimal rationality never focuses only on one agent – minimalism is not a property of an isolated agent. Measuring rationality does not give us an *absolute position* of some very poor agent models on the map of possible rational agent models as long as we do not take an absolute (god-like) position on that map. It is even quite likely that there can be very different rational agent models that can be named minimal. For this reason, we first dealt with the problem of combining different minimalisms and measures.

**Theory of mind of rational agents** is defined as the application of meta schemas on the (structuring) theory the agent has of itself. The aim of that application is to reach another agent model gradually, in order to be able to understand another agent, or to solve a specific problem at hand, or else. In a sense, Theory of mind then is self-assembly (though, of course, not necessarily in a conscious modus).

Theory of mind can be specified in accord with the "map" of rational agent models (i.e. the graph) as the set of nodes and vertices that can be "used" by the agent.

**Minimal theory of mind** then specifies the set of reachable agent models in a minimal (distant) way. In a certain sense this means that Minimal Theory of mind is not a very poor form of human theory of mind realized in animals or robots. Instead, we should use the label "minimal" to characterize the effort necessary to understand other agents. We understand other agents like ("normal") humans surely much easier and with much less (minimal) effort than for example chimps, robots etc. This also refines the notion of minimal rationality: it is not a very poor form of maximal rationality (of god, for example). Instead, with minimal effort we can rationally understand those agents consisting of agent models that can be reached by us. But because we always combine several measures, our position on the map or in the graph is neither fixed nor absolute with respect to something like the rational god. Multi-criteria problems like measuring rationality or applying theory of mind do not have "absolute" solutions. This gives us a definition of minimal understanding.

**Minimal (rational) understanding** is specified by the set of agent models where theory of mind can be applied to in a specified minimal sense. Because understanding is a multi-criteria problem, there can be several different sets of agent models we are able to understand in a minimal way. Intuitively, this first seems to be contrary to what we usually label minimal understanding: only being able to understand somebody in a very poor or distant way. But here also intuitions about rational understanding may not be right. Otherwise, we ourselves

would be the ones we would understand maximally and best. But this would implicate something like “full introspection” or “full transparency”, which is logically possible, but empirically for *this world* certainly wrong (see for example Carruthers (2011) for an actual overview).

If you want to hold up some intuitions, then this one may be more interesting for understanding: knowing someone’s epistemic history and gradual changes let us understand him in a better way than without knowing it. However, this is only right to a certain degree: it is questionable if we could extrapolate such future “epistemic histories” in a plausible way to very high numbers (up to infinity) of gradual changes. Though this would need an own essay, I want to give a first impression what the matter could be with this issue: Do you agree that agents essentially do something like computation if they (cognitively) change? And do you think that this computation can be modeled by a Turing machine (Fodor, 1975)? If so, then maybe you agree that we would have to solve something like the halting-problem: do the changes lead to a “viable” agent model or does the agent model crash, runs infinite loops, or else? So, maybe we cannot be sure in a very fundamental way that we can “extrapolate” an agent model’s future changes (or even the changes of the meta-schema, if it is applied to itself in case of “evolution”, see future directions). Of course, there are many other options that should be explored: Non-Turing-computation for example. Or even no computation at all. But this is a topic for future research.

To sum up the given definitions, the following (obviously quite artificial) example is given by figure 10. Assume that Agent A has modeled a specific mental model representing spatial relations (the red bricks), and A wants to have the relations that can be modeled by M3. Agent A gradually changes to Agent M3. But the M-Series do not know (have eliminated) the principle of transitivity (for example, because of some minimalism). Then, A develops itself to M3, but the way back to A may not be the same way – because M3 cannot reconstruct that way by transitivity. Though M3 is rational, it may be minimal rational from the viewpoint of A. This also implicates that A cannot be understood easily by M3, though the opposite may hold: A understands M easily. For A, M is within the focus of minimal Theory of mind, but possibly not vice versa. A can take the perspective of M3 to A, but M3 cannot easily take the perspective of A to see how A sees M3. However, both agent models could (in a synchronic fashion) interoperate, though M3 does not work effectively for A, but seems to try out instead of “knowing”.

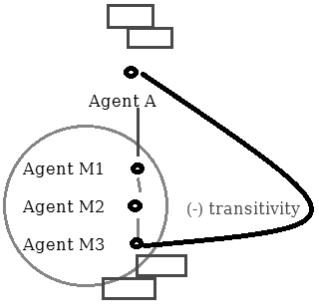



Figure 10. A’s gradual change from A to M3, and the difficult way back from M3 to A. A develops from A to M1 ... M3 and back from M3 to A but loses the transitivity rule (for what reason ever). However, if the Ms do not accept transitivity, it is not easy to take the same route back to A.

## 6. Conclusion and Future Directions of Research

In this paper, theory of mind was developed as an agent's method of rational self-assembly by gradual change. This has been abstracted from theory and experiment in the cognitive sciences, especially spatial cognition and belief revision, though not limited to these areas. Agents do so by "measuring" other agents according to multiple criteria, i.e. they measure the distance between their own position and another agent's position on the map of rationality. Then they apply methods following blueprints or schemas of methods. We can conzeptualize that map as a conceptual neighborhood graph. Based on this graph, several notions can be given without much effort concerning distance, minimal rationality, minimal theory of mind, and minimal understanding. So, theory of mind can be characterized as a multi-criteria problem that is solved essentially by gradual change in terms of self-assembly. For this reason, there neither is the one minimalism nor is it easy to propose how to combine several minimalisms. There are very different possible directions of future research in this area, just to name a few and show the potential of the approach:

(1) Losing/tightening the formal constraints to model specific "maps" of rationality

Notions like consistency, inconsistency, or circularity are very important with respect to actual real agents. "Consistency"/"inconsistency" are hot topics for the development of theory of mind, i.e. in terms of false belief tasks that inherently assume for example young children to apply inconsistent theories about persons and the world. It is also possible, for example, to disallow circularity to get chains and directions on the graph that model histories of epistemic moves (in a diachronic perspective), or communities of similar agent models (in a synchronic perspective).

By such constraints, one can explore what epistemic moves are secure or allowed with respect to consistency, for example. And what does it mean for possible rational understanding to change such constraints like consistency? The reconstruction of another agent's schema can depend on consistency or inconsistency assumptions – for example, for the same "arithmetical surface" (doing arithmetical operations), one can always have a consistent projective arithmetic or an inconsistent arithmetic (e.g. Priest (1994), Priest (2000)). This is a kind of theory choice: which theory do we choose that analyzes the other agent in the best way? Then, differences in consistency assumptions may lead to different accepted norms in gradual development, because different "epistemic moves" are allowed depending on allowed grades of inconsistency.

It is obviously possible to map such issues on the approach by changing conceptual parameters of method schemas or of the graph. Then, it is also an interesting project to compare the resulting graphs with classical approaches of belief fixation and belief change like Levi (1991). He postulates a structure based on Boolean algebra/Boolean lattice and conceptualizes ultra-filters and filters in order to characterize "secure epistemic moves". From a mathematical point of view, such algebras/lattices are just special cases in that graph (namely every chain is an algebra/lattice, if we want to introduce them with the necessary commitments to order theory and at least local consistency).

(2) Developing the "method"-approach in theory of mind and rationality research

In this paper, we only had a first glimpse of "method-engineering" and what it could do for the conceptualization of agents. Future research could work on building blocks of agents consisting of method schemas, for example.

(3) Applying gradual change to gradual change itself, i.e. having a developmental perspective

As we know from actual agents, theory of mind develops. With respect to the building blocks, one could model developing theory of mind for example by letting building blocks of method

schemas evolve. This application by the agent itself to itself is especially interesting, if it loses some of its fundamental principles on the way from one agent model A to another agent model M3 (see figure 10).

(4) Interoperability between different agents

If we can conceptualize different agent models and maybe even different agent models have evolved in nature, it is interesting to see how they can interoperate. Just consider human-animal-interaction or human-robot-interaction. If we allow for more structure and restriction (for example by chains and directions), and thus can identify communities of agent models sharing some method schemas, we can integrate questions of interoperability. For example, we can ask: How very different existing agent models in other communities can be reached – f given our one agent model? And how can agent models gradually change (to interoperable but not necessarily equal agent models) to solve cooperatively a given problem? For a technical view on this issue of interoperability between (very simple) cognitive agent models see for example Doerr et al. (2009) (cognitive radio).

For a rich picture of agents, minimalisms should not be parallel but should be combined in the analysis of agents. After all, my aim was to sketch a starter for such a framework. The positions of this sketched approach can be summed up finally:

- Theory of mind is considered as gradual change by applying instances of method schemas – development is not mysterious and does not come in arbitrary stage models.
- The method approach allows us to consider blueprints of possible schemas used to initiate gradual change (for example schemas of agent models or rationality measures).
- The application of theory of mind in such a framework can be given a bunch of examples and brings in new (and old) problems concerning the legitimization of ingredients and acceptable change of agent models (norms, basic inferences, cognitive apparatus, consistency vs. inconsistency, etc.).
- The combination of minimalisms in gradual change is considered as a multi-criteria problem. It means that not one issue (like consistency) is the primary issue, and this naturally reflects the dimension of the measurement combination problem. This may shed new light on issues like false belief and actual agents.
- Grounded definitions of distance and understanding can be given. Different rational agents do not stand loosely and lonely in isolated spaces, but can be positioned with a distance to each other on a map of rationality. Essentially, this map is not “absolute” or from god's point of view, but it depends on the accepted measures and minimalisms.
- Future work should show how some theories of belief change are just special cases of a more general theory of rational change of agents.

**Gerhard Chr. Bukow**

Institute of Philosophy, University of Magdeburg
Institute of Psychology, University of Giessen
bukow@ovgu.de

## References

Alchourròn, C. E., Gärdenfors, P. and Makinson, D. 1985: On the logic of theory change: Partial meet contraction and revision functions. Journal of Symbolic Logic, 50, 510–530.

Alecchina, N., & Logan, B. 2010: Belief ascription under bounded resources. Synthese, 173, 2, 179-197.

Bermudéz, J. L. 2003: Thinking without words. Oxford University Press.

Bruns, T. and Egenhofer, M. 1996: Similarity of spatial scenes. Seventh International Symposium on Spatial Data Handling (SDH '96), Delft, The Netherlands, M.-J. Kraak and M. Molenaar (eds.), pp. 173-184, August 1996.

Butterfill, S. and Apperly, I. 2011: How to Construct a Minimal Theory of Mind. Online version:
http://butterfill.com/papers/minimal_theory_of_mind/minimal%20theory%20of%20mind%2007c.A4.share.pdf (31.05.2012)

Carruthers, P. 2011: *The opacity of of mind: an integrative theory of self-knowledge*. Oxford University Press.

Cherniak, C. 1986: *Minimal rationality*. MIT Press.

Fodor, J. 1975: *The language of thought*. Harvard University Press.

Gopnik, A. 1988: Conceptual and semantic development as theory change. *Mind and Language*, *3*, 163-179.

Levi, I. 1991: *The Fixation of Belief and its Undoing: Changing Beliefs through Inquiry*. Cambridge University Press.

Ralyté, J. 2002: 'Requirements Definition for the Situational Method Engineering', in: Proceedings of the IFIP WG8.1 Working Conference on Engineering Information Systems in the Internet Context (EISIC'02), Kanazawa, Japan, September 2002. C. Rolland, S. Brinkkemper, and M. Saeki (eds.), Kluwer Academic Publishers, 127-152.

# Mechanistische Erklärung: Reduktiv oder nicht?

Bettina Gutsche

Ist der Ansatz der mechanistischen Erklärung (ME) ein ausschließlich reduktiver Ansatz? Hat er reduktive Komponenten oder muss er als völliger Gegenentwurf zum Reduktionismus aufgefasst werden? Die Antworten hängen davon ab, wie Reduktionismus und Reduktion verstanden werden und wie die Vertreter von (ME) ihren Ansatz beschreiben. Während in einigen Publikationen William Bechtels und seiner Kollegen (Bechtel 2001; Bechtel 2007; Bechtel & Hamilton 2007) der mechanistische Erklärungsansatz durchaus als ein reduktionistischer Ansatz verstanden wird, so ist dies bei den Arbeiten von Carl Craver und seinen Kollegen (v.a. Machamer, Darden & Craver 2000; Craver 2005; Craver 2007) und auch bei neueren Publikationen von William Bechtel und Kollegen (Bechtel & Abrahamsen 2008; Bechtel 2009; Bechtel 2010) nicht der Fall. Im Folgenden wird anhand der angeführten Texte gezeigt, inwiefern (ME) als reduktiver Ansatz aufgefasst werden kann. Danach wird beschrieben, inwieweit (ME) dem Reduktionismus entgegengesetzt ist. Schließlich werden mit Ernest Nagel, dem Begründer der klassischen Reduktion, die verbleibenden reduktiven Komponenten von (ME) beleuchtet. Genauer: die „reduktive Sicht der Vereinheitlichung“, die Craver (2007) angreift und der er für die Neurowissenschaften eine alternative Form von Vereinheitlichung entgegensetzt, kann im Sinne von Nagel (1961) rehabilitiert werden. Damit wird die Nagel-Reduktion in Teilen als mit (ME) kompatibel erachtet.

## 1. Einleitung

In diesem Aufsatz geht es darum, ob der Ansatz der mechanistischen Erklärung, bei dem kausale Mechanismen für bestimmte Phänomene v.a. in den Neurowissenschaften gefunden werden, ein reduktiver bzw. reduktionistischer Ansatz ist. Eine Hauptfrage dabei ist, ob der Ansatz reduktive Komponenten hat oder gar als völliger Gegenentwurf zum Reduktionismus aufgefasst werden muss. Dies hängt natürlich davon ab, wie Reduktionismus und Reduktion verstanden werden, jedoch auch, wie die Vertreter der mechanistischen Erklärung ihren Ansatz beschreiben.

Während in einigen Publikationen William Bechtels und seiner Kollegen (Bechtel 2001; Bechtel 2007; Bechtel & Hamilton 2007) der mechanistische Erklärungsansatz durchaus als ein reduktionistischer Ansatz verstanden wird, so ist dies bei den Arbeiten von Carl Craver und seinen Kollegen (v.a. Machamer, Darden & Craver 2000; Craver 2005; Craver 2007) und auch bei neueren Publikationen von William Bechtel und Kollegen (Bechtel & Abrahamsen 2008; Bechtel 2009; Bechtel 2010) nicht der Fall. Das heißt, hier wird der mechanistische Erklärungsansatz als eine Alternative beschrieben, die die Vorzüge der Reduktion beibehält und die Nachteile ausmerzt. Mehr noch, ein Vergleich mit dem „alten“ Reduktionsmodell scheint nicht mehr angebracht, da der Ansatz der mechanistischen Erklärung (nachfolgend auch (ME) genannt) eine eigene Betrachtungsweise bietet, die nicht mit einem Konkurrenzmodell verglichen werden muss, um sich dagegen abzuheben. Vielleicht markiert auch das Jahr 2007 mit dem Erscheinen von Cravers *Explaining the Brain* einen Wendepunkt in der Beschreibung des mechanistischen Erklärungsansatzes, da Craver (2007) so prägnant und scharfsinnig die Vorzüge des mechanistischen Ansatzes erklärt, z.B. seine empirische Plausibilität (d.h. dass in den Neurowissenschaften wirklich Forschung nach diesem Modell betrieben wird und nicht nach dem Modell der Reduktion), sowie die Kritikpunkte am klassischen Reduktionsmodell herausstellt.

Im Folgenden möchte ich anhand der angeführten Texte zunächst zeigen, inwiefern (ME) als reduktionistischer Ansatz verstanden werden kann (Abschnitt 3), danach beschreibe ich, inwieweit (ME) dem Reduktionismus entgegengesetzt ist (d.h. die Kritik am Reduktionismus, Abschnitt 4), um am Ende jedoch mit Ernest Nagel, dem Begründer der *klassischen Reduktion*, wieder die verbleibenden reduktiven Komponenten von (ME) zu beleuchten (Abschnitt 5). Genauer heißt das: die „reduktive Sicht der Vereinheitlichung", die Craver (2007) angreift und der er für die Neurowissenschaften eine alternative Form von Vereinheitlichung („intralevel/interlevel integration") entgegensetzt, kann im Sinne von Nagel (1961) rehabilitiert werden. Im Fazit wird die Nagel-Reduktion in Teilen als mit dem mechanistischen Ansatz kompatibel erachtet.

Zuvor soll jedoch eine kurze Charakterisierung des Ansatzes der mechanistischen Erklärung gegeben werden und anhand des Beispiels der Weiterleitung eines Aktionspotentials veranschaulicht werden, wie mechanistische Erklärung funktioniert (Abschnitt 2).

## 2. Mechanistische Erklärung

Was macht den mechanistischen Erklärungsansatz (möglicherweise im Gegensatz zur klassischen Reduktion) aus? Darin, d.h. in einer ersten kurzen Definition, unterscheiden sich Bechtel und Craver und jeweilige Kollegen kaum. Schauen wir uns drei entsprechende Zitate an, die auf die Frage antworten, was ein Mechanismus ist. Ein Mechanismus ist „a set of entities and activities organized such that they exhibit the phenomenon to be explained." (Craver 2007: 5) Mechanismen sind

> collections of entities and activities organized in the production of regular changes from start or setup conditions to finish or termination conditions (Craver 2002: S84, ähnlich in Machamer, Darden & Craver 2000).

Ein Mechanismus ist

> a structure performing a function in virtue of its component parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena (Bechtel & Hamilton 2007: 405; aus Bechtel & Abrahamsen 2005: 423).

Die wichtigsten Bestandteile eines Mechanismus sind also seine (relevanten) Teile bzw. Entitäten, entsprechende Operationen/Aktivitäten und eine entsprechende Organisation. *Entitäten* sind Komponenten (d.h. relevante Teile) im Mechanismus mit bestimmten Eigenschaften: sie sind lokalisierbar, haben eine bestimmte Größe, Struktur und können auch eine bestimmte Ausrichtung haben. *Aktivitäten* sind die kausalen Bestandteile des Mechanismus (vgl. Craver 2007: 6). Aktivitäten sind *produktiv* in dem Sinne, dass sie einen (kausalen) Unterschied machen (gehen also über Korrelationen, pure zeitliche Sequenzen hinaus und können für „Manipulation und Kontrolle" genutzt werden). Die Entitäten und Aktivitäten sind zudem zeitlich, räumlich, kausal und hierarchisch *organisiert* und mit einem Mechanismus wird ein *zu erklärendes Phänomen* beschrieben.

Es geht also augenscheinlich im mechanistischen Erklärungsansatz darum, ein Phänomen dadurch zu erklären, dass ein „zugrunde liegender" Mechanismus angegeben wird, bei dem auf die Teile des Phänomens sowie auf deren Zusammenspiel rekurriert wird. Anders scheinbar als bei der klassischen Reduktion (siehe auch Unterabschnitt 3.1) geht es nicht darum, die Beschreibung des Phänomens aus der Beschreibung der Prozesse auf einer niedrigeren Ebene *logisch* abzuleiten.

Beispiele für Mechanismen finden sich zahlreich in den Bio- und Neurowissenschaften, z.B. die Entstehung eines Aktionspotentials (und deren Weiterleitung, das heißt elektrische Signalweiterleitung am Axon bzw. chemische Signalübertragung an der Synapse), DNA-

Transkription und Translation, das Phänomen der Langzeitpotenzierung (LTP = *long term potentiation*), das mit Lernen und Gedächtnis in Verbindung gebracht wird, Prozesse der visuellen Wahrnehmung etc. Als ein Beispiel soll in den folgenden drei Unterabschnitten die Weiterleitung eines Aktionspotentials am Axon beschrieben werden (vgl. Birbaumer & Schmidt 2006: Kapitel 3; Schandry 2003: Kapitel 4).

### *2.1 Weiterleitung des Aktionspotentials am Axon – das Ruhepotential*

Dieser Vorgang nimmt das so genannte Ruhepotential einer Nervenzelle zum Ausgangspunkt. Das Ruhepotential der Nervenzellen liegt durchschnittlich bei etwa -70mV und kommt dadurch zustande, dass die Konzentration von innerhalb und außerhalb der Zellmembran befindlichen positiv oder negativ geladenen Ionen sich derart verteilt, dass im Inneren der Zelle eine negativere Ladung vorliegt. Positiv geladene Ionen sind z.B. Kalium- (K+), Natrium- (Na+) und Kalzium-Ionen (Ca2+); negativ geladene Ionen sind z.B. Chlorid-Ionen (Cl-) und Eiweiß-Anionen. Die Zellmembran ist nicht für alle Ionen gleichermaßen durchlässig. Dies ist ein Grund, warum es nicht zu einem Ladungsausgleich kommt und das Ruhepotential aufrechterhalten wird (entgegen der Diffusionskraft sowie der elektrischen Anziehung verschieden geladener Ionen). Ein weiterer Grund ist die unter Energieausnutzung (also durch Anlagerung von ATP/Adenosintriphosphat) funktionierende Natrium-Kalium-Pumpe, ein Ionenkanal, der Natrium-Ionen aus der Zelle hinausbefördert und Kalium-Ionen wieder in die Zelle hineinbringt (dabei werden mehr positiv geladene Na+ Ionen hinaus als positiv geladene K+ Ionen in die Zelle hinein befördert).

### *2.2 Weiterleitung des Aktionspotentials am Axon – das Aktionspotential*

Ein Aktionspotential kann nun derart beschrieben werden, dass die Zelle (z.B. durch verschiedene Signale von benachbarten Zellen meist über Synapsen und über die Dendriten der Zelle übertragen) eine Depolarisation über einen bestimmten Schwellenwert hinaus (z.B. -40mV) erfährt und damit „feuert“. Die charakteristische Spannungskurve eines Aktionspotentials ist gekennzeichnet durch einen steilen Anstieg des Potentials in den positiven Bereich hinein (das Maximum liegt etwa bei +30/+40mV) sowie danach einen etwas flacheren Abfall des Potentials (Repolarisation) über ein Nachpotential in die negative Richtung (Hyperpolarisation) wieder zurück zum Ruhepotential. An der Zellmembran wird das Aktionspotential durch verschiedene Ionenkanäle und den Austausch und die Wanderung von Ionen realisiert: wird der Schwellenwert erreicht, so öffnen sich spannungssensitive Natrium-Kanäle und in sehr kurzer Zeit strömen viele Na+ Ionen in die Zelle hinein, das Membranpotential wird positiv. Nach etwa 1 ms schließen sich die Kanäle wieder und es kommt zur Öffnung von Kalium-Kanälen, durch die K+ Ionen rasch aus der Zelle hinauswandern, womit das Potential wieder ins Negative abfällt. Die dadurch veränderten Konzentrationen der Natrium- und Kalium-Ionen werden durch die Natrium-Kalium-Pumpe wieder ins Gleichgewicht gebracht.

### *2.3 Weiterleitung des Aktionspotentials am Axon – Ausbreitung des Potentials*

Ruhe- und Aktionspotential, wie bisher beschrieben, sind nun die Grundlagen dafür zu verstehen, wie sich ein Aktionspotential vom Axonhügel der Zelle aus entlang des Axons weiter ausbreitet (diese Ausbreitung erfolgt in der Regel immer nur in eine Richtung) hin zu den nachgeschalteten Nervenzellen. Die Ionenströme erfolgen nicht nur zwischen Zellinnerem und dem extrazellulären Raum, sondern die Ionen können auch innerhalb der Zelle entlang des Axons wandern. Durch die spannungsgesteuerten Natrium-Kanäle wird an einer Stelle des Axons ein Aktionspotential generiert, welches wiederum benachbarte Natrium-Kanäle in Ausbreitungsrichtung stimuliert und damit aktiviert. So kann sich das

Aktionspotential entlang des Axons ausbreiten (in etwa vergleichbar mit dem „Abbrennen“ einer Zündschnur).

In dieser mechanistischen Erklärungsskizze wurden z.B. folgende *Entitäten* benannt: die Zelle, ihre Teile wie Dendriten, Axonhügel und Axon; Natrium- und Kalium-Ionen; verschiedene Arten von Ionenkanälen wie spannungsgesteuerte Natrium- und Kalium-Kanäle, die Natrium-Kalium-Pumpe, etc. Die *Aktivitäten* im beschriebenen Mechanismus sind z.B. Depolarisieren, Öffnen, Schließen, Einströmen, Ausströmen, etc. Die zeitliche und räumliche *Organisation* wurde ebenso angedeutet: z.B. das Schließen der Natrium-Kanäle nach einer kurzen Zeit von 1 ms, die Signalweiterleitung in eine Richtung vom Zellkörper und Axonhügel weg zum Ende des Axons hin.

Die Auseinandersetzung mit dem mechanistischen Erklärungsansatz wirft einige Fragen auf, die ich hier kurz andeuten, denen ich jedoch im Folgenden nicht weiter nachgehen möchte (einen interessanten Beitrag dazu leistet m.E. Fazekas & Kertész 2011). Eine der Fragen ist diejenige nach dem Zusammenhang zwischen dem Mechanismus und dem zu erklärenden Phänomen. Der Mechanismus erklärt das Phänomen, also scheint er nicht damit identisch sein zu können. Weiterhin scheint der Mechanismus als Ganzer (mitsamt seiner Organisation) auf einer höheren Ebene ansässig zu sein als die entsprechenden Teile des Phänomens bzw. Teile im Mechanismus (man spricht auch davon, dass die organisierten Teile den Mechanismus *konstituieren*). Man fragt sich hier beispielsweise, auf welcher Ebene sich die „organisierten“ Teile eines Mechanismus befinden: auf der Ebene der Teile, auf der (höheren) Ebene des Mechanismus oder auf der (ebenfalls höheren) Ebene des Phänomens? Diese Fragen zu beantworten scheint relevant für die Bewertung des mechanistischen Ansatzes als ein reduktiver oder nicht-reduktiver Ansatz zu sein (siehe Fazekas & Kertész 2011). Hier möchte ich jedoch einen anderen Weg einschlagen und explizit eine Kritik von Craver (2007) zurückweisen (siehe Abschnitte 4 und 5).

Nach dieser ausführlichen Illustration, wie eine mechanistische Erklärung aussieht, komme ich zum nächsten Abschnitt.

## 3. Mechanistische Erklärung als ein reduktiver Ansatz

### *3.1 Modelle der Reduktion*

Zuerst soll ein kurzer Überblick über die *klassische Reduktion* und ihre Ableger gegeben werden. Bei klassischen Reduktionsmodellen handelt es sich um Varianten der Theorienreduktion, d.h. es werden verschiedene Theorien aufeinander reduziert. Die beiden wichtigsten formalen Prinzipien der Nagel-Reduktion (Nagel 1961: Kapitel 11) sind Verknüpfbarkeit (*connectability*) und Ableitbarkeit (*derivability*), d.h. wenn es der Fall ist, dass einige von den Begriffen der zu reduzierenden Theorie nicht in der reduzierenden Theorie enthalten sein sollten (heterogene Reduktion), so kann über begriffliche Verbindungen (die viel zitierten Brückengesetze) das fehlende Vokabular in die reduzierende Theorie eingeführt werden (*Verknüpfbarkeit*). Verfügen beide Theorien dann über die gleichen Begriffe (bzw. die Basistheorie muss über die (wahren) Begriffe der zu reduzierenden Theorie verfügen, Nagel spricht von homogener Reduktion), so können auch idealerweise die Gesetze der zu reduzierenden Theorie aus den Gesetzen der reduzierenden Theorie abgeleitet werden (*Ableitbarkeit*). Aus der Ableitbarkeit folgt die Verknüpfbarkeit, aber nicht umgekehrt.

Eine eher metaphysische Abwandlung dieses Ansatzes ist in dem Manifest (wie es Craver (2007) nennt) von Oppenheim und Putnam (1958) zu finden – in ihrem Programm der *Mikroreduktion*. Hier werden reduktive Ebenen vorausgesetzt und Reduktion wird als Mittel begriffen, eine Vereinheitlichung in den Wissenschaften herzustellen, d.h. es werden nur reduktive Beziehungen *zwischen den Ebenen* als vereinheitlichend gewertet.

Diese beiden Modelle (das von Nagel und das der Mikroreduktion) werden oft zusammen unter das Etikett „klassische Reduktion“ subsumiert (vgl. z.B. Fodor 1974; McCauley 1986: 180; obwohl sich die Modelle durchaus unterscheiden) und als starke Ansätze der Reduktion aufgefasst. Hier sollte m.E. jedoch – wie sich im Folgenden auch in diesem Aufsatz zeigen wird – eine schärfere Trennlinie gezogen werden.

Neuere Reduktionsmodelle gestehen auch zu, dass Teile der reduzierten Theorie verworfen werden können, solange sich die wahren Teile nichtsdestotrotz *annähernd* aus der Basistheorie ableiten lassen (einige dieser Ansätze findet man in der Literatur unter dem Stichwort „New Wave Reduktionismus“, z.B. Hooker 1981; Bickle 1996; Bickle 1998). Modelle der *approximativen Reduktion* (z.B. auch Schaffner 1967) „allow the fit between reduced and reducing theory to be less than exact“ (Craver 2007: 229).

Als letzte abgeschwächte Variante der Reduktion wird in der Debatte vielfach behauptet, dass Theorienreduktion in den Neurowissenschaften (und als ein Anwendungsfall der *mind sciences* für das Körper-Geist-Problem) nicht erreicht werden kann. Stattdessen könne man jedoch Phänomene höherer Ebenen immer noch *reduktiv erklären*. Schwache Varianten der Reduktion haben das Prinzip der Ableitbarkeit komplett aufgegeben und Reduktion besteht nur noch darin, dass Phänomene höherer Ebenen durch fundamentale Mechanismen oder Gesetze erklärt werden sollen/können: „All that remains of reduction in these cases is a commitment to the primacy of downward and fundamental explanation.“ (Craver 2007: 230) Wir haben im letzten Fall also noch eine Art „Perspektive-nach-unten“ (*downward looking perspective*), der ein Vorrang eingeräumt wird.

Die Reduktionsmodelle können in abnehmender Stärke (in Anlehnung an Craver 2007: 229) wie folgt aufgelistet werden:

(1) Klassische Reduktion (Nagel-Reduktion und Mikroreduktion),

(2) Approximative Reduktion, z.B. New Wave Reduktion,

(3) Reduktive Erklärung.

### *3.2 Mechanistische Erklärung ist reduktiv*

Inwiefern kann nun der mechanistische Erklärungsansatz als ein reduktionistischer Ansatz verstanden werden? Am ehesten sicher als Reduktionismus in seiner schwächsten Variante. Dies wird z.B. an dem Titel eines Papers von Bechtel deutlich: „Reducing Psychology while Maintaining its Autonomy via Mechanistic Explanations“ (Bechtel 2007). Wir können einerseits die Psychologie (bzw. ihre Theorien) reduzieren, aber dennoch ihre Autonomie aufrecht erhalten (ein oftmals vorgebrachter Vorwurf gegen die klassische Reduktion, d.h. das Problem, dass die Nagel-Reduktion kontraintuitive Konsequenzen für alle nicht-fundamentalen („speziellen“) Wissenschaften und deren Gegenstände habe, vgl. Fodor 1974), und zwar mithilfe von mechanistischen Erklärungen. Laut Bechtel (2007) sind mechanistische Erklärungen sowohl reduktionistisch in diesem schwächeren Sinn als auch kompatibel mit der Vorstellung der Autonomie höherer Ebenen.

Es fragt sich dennoch, was *reduzieren* in diesem Kontext noch bedeuten kann. Denn Bechtel (2007) bezieht sich auch auf *Ebenen von Mechanismen* (ähnlich wie Craver 2007: Kapitel 5). Diese seien lokal definiert (bzw. unter Rückgriff auf Mechanismen), so dass ein umfassender „Blick nach unten“ damit gar nicht gewährleistet werden könne, da Entitäten nur *innerhalb* eines Mechanismus bezüglich ihrer Ebenen verglichen werden können. Was bei (ME) an Reduktion zu bleiben scheint, ist allein die „Perspektive nach unten“, die jedoch keine Priorität zu haben scheint. Somit haben wir es hier mit einem noch schwächeren Begriff von Reduktionismus zu tun. Laut Bechtel sei der *reduktive Aspekt* (den Prinzipien der *Dekomposition* und *Lokalisierung* folgend) allein nicht hinreichend, um das Verhalten des

Mechanismus zu erklären. Wichtig seien nicht nur die Teile und ihre Operationen, sondern auch ihre *Organisation*.

In Bechtel und Hamilton (2007) findet sich z.B. folgendes Zitat:

> A central feature of mechanistic explanations, and the one that makes them reductive, is that they involve decomposing the system responsible for a phenomenon into component parts and component operations. (Bechtel & Hamilton 2007: 405)

Durch den Rekurs auf zugrunde liegende Komponenten und Operationen bleibe der reduktionistische Anspruch gewahrt. Jedoch wird nicht davon ausgegangen, dass die Teile *allein* die entsprechenden Phänomene hervorbringen, sondern der Mechanismus als Ganzer.

Ähnliche Argumentationsstränge finden sich auch in Craver und Bechtel (2007). Die Autoren nehmen an, dass es Ursachen höherer Ebene gibt, die jedoch durch konstitutive Mechanismen (niedrigerer Ebene) vollständig erklärt werden können. In Craver und Bechtel (2007) wird die mysteriöse Rede von Verursachung zwischen den Ebenen (*between-level causation*) analysiert. Der Mechanismus könne zwar kausale Eigenschaften haben, die seine Teile nicht haben, aber „[w]e do not assume that the mechanism has causal powers over and above the organized collection of their parts." (Craver & Bechtel 2007: 548, Fußnote 2)

Schauen wir uns noch einmal die Auflistung mit den reduktionistischen Positionen in 3.1 an. Es scheint, als müssten wir eine vierte Position hinzufügen, die aber so schwach zu sein scheint, dass sie womöglich in eine anti-reduktionistische Perspektive „umkippt" (vgl. Abschnitt 4):

(1) Klassische Reduktion (Nagel-Reduktion und Mikroreduktion),

(2) Approximative Reduktion, z.B. New Wave Reduktion,

(3) Reduktive Erklärung,

(4) (ME) ist reduktiv in dem Sinne, dass es einen *reduktiven Aspekt* gibt, eine *downward-looking*-Perspektive.

Diese „Perspektive-nach-unten" scheint aber verglichen mit derjenigen im Modell der reduktiven Erklärung keine Priorität zu haben. Ebenso wichtig scheint die *upward-looking*-Perspektive zu sein (vgl. dazu auch den eindrücklichen Titel von Bechtel (2009): „Looking Down, Around, and Up: Mechanistic Explanation in Psychology").

Dieses vorläufige Ergebnis bedarf einer weiteren Kommentierung. Bliebe es bei dieser Diagnose, dass der Ansatz der mechanistischen Erklärung nur reduktiv im Sinne von (4) sei, so hieße dies, dass sich in (ME) reduktive und nicht-reduktive Komponenten mischen und man nicht letztgültig sagen könnte, ob (ME) nun reduktiv sei oder nicht. Die Antwort wäre ein Kompromiss, d.h. mechanistische Erklärung scheint beides zu sein, sowohl reduktiv als auch nicht-reduktiv.

Fraglich bleibt dabei jedoch, was mit der „Perspektive-nach-unten" bzw. „Perspektive-nach-oben" genau gemeint ist. Sicherlich gilt auch für die klassische Reduktion, dass es Phänomene höherer Ebene gibt, dass damit auch verschiedene Perspektiven einhergehen können. Im Standard-Beispiel für die Nagel-Reduktion, der Reduktion der Thermodynamik auf die statistische Mechanik, wird die Theorie der Thermodynamik unter Zusatzannahmen auf die Theorie der Mechanik reduziert bzw. aus ihr abgeleitet. Eine der Zusatzannahmen bezeichnet dabei die Verknüpfung des Begriffs „Temperatur (eines Gases)" aus der Theorie der Thermodynamik mit dem Begriff „mittleren kinetischen Energie (der Moleküle des Gases)" aus der Theorie der Mechanik. Das heißt, ein Phänomen höherer Ebene wird mithilfe von Prozessen niedrigerer Ebene *erklärt*. Dennoch bleibt das Phänomen höherer Ebene bestehen. Auch wird die Theorie der Thermodynamik durch die Reduktion gerechtfertigt und damit weiterentwickelt. Es ließe sich also sagen, dass auch in der klassischen Reduktion

verschiedene Perspektiven – nach oben, nach unten, zur Seite – eingenommen werden (vgl. auch Abschnitt 5).

Dennoch scheint gerade die klassische Reduktion deshalb reduktiv zu sein, weil das Phänomen höherer Ebene *durch* Prozesse niedrigerer Ebene *erklärt* wird. Die niedrigeren Ebenen haben also eine Erklärungspriorität, und es lässt sich dafür argumentieren, dass diese Erklärungspriorität auch im mechanistischen Ansatz besteht, dass (ME) also reduktiv im Sinne von (3) ist.[1]

Nachdem ich nun angedeutet habe, worin das Reduktionistische am Ansatz der mechanistischen Erklärung liegt, komme ich zum nächsten Abschnitt.

## 4. Mechanistische Erklärung als ein nicht-reduktiver Ansatz

### *4.1 Mechanistische Erklärung ist nicht-reduktiv auf vielfältige Weise*

Warum ist der Ansatz der mechanistischen Erklärung *kein* reduktiver Ansatz? Dazu findet man bei Craver (2007) in nahezu jedem Kapitel relevante Aussagen. Es lassen sich vier Hauptkritikpunkte am klassischen Reduktionsmodell ausmachen, welche (ME) versucht zu umgehen.

Erstens muss daran erinnert werden, dass die Nagel-Reduktion eine Verallgemeinerung des deduktiv-nomologischen Modells der Erklärung (D-N-Modell) ist, weshalb die zahlreichen Kritikpunkte gegen das D-N-Modell auch gegen die Nagel-Reduktion angeführt werden können. Logische Ableitung aus gesetzesartigen Verallgemeinerungen ist *weder notwendig noch hinreichend* dafür, dass das entsprechende Explanandum *erklärt* wird. Das D-N-Modell ist nicht notwendig (d.h. es ist zu restriktiv bzw. lässt manche Schlüsse, die Erklärungen sind, nicht als solche gelten), weil es z.B. in den Neurowissenschaften keiner (strikten) Gesetze bedarf, um erfolgreiche Erklärungen abzugeben. Stattdessen sind die Generalisierungen in den Neurowissenschaften eher *ceteris paribus* Gesetze bzw. „fragile Generalisierungen“ – ein Begriff von Craver (2007). Das D-N-Modell ist auch nicht hinreichend (d.h. es ist zu liberal, es lässt Erklärungen als Erklärungen durchgehen, die keine sind), da eine Ableitung, die dem Schema genügt, noch Fragen offen lassen kann. So kann mithilfe zufälliger Generalisierungen etwas als Erklärung angeführt werden, das jedoch nicht relevant für das Explanandum ist, da z.B. keine genuinen Ursachen benannt werden. Das D-N-Modell kann generell nicht zwischen relevanten und irrelevanten Erklärungen bzw. zwischen möglichen und anzunehmenden Erklärungen unterscheiden (vgl. für eine umfassende Kritik und einen Alternativ-Ansatz für ätiologische Erklärungen, Craver 2007: Kapitel 2 und 3, für konstitutive Erklärungen, Craver 2007: Kapitel 4).

Ein zweiter grundlegender Kritikpunkt kann für das Ebenenmodell von Oppenheim und Putnam (1958) formuliert werden (vgl. Craver 2007: Kapitel 5). Craver zeigt, dass die Ebenen der Natur nicht mit den Ebenen der Theorien/Wissenschaften (wie es Oppenheim und Putnam behaupten) korrespondieren. In Wirklichkeit gebe es keine monolithischen Ebenen der Welt, so Craver (2007: 191).

Ein dritter damit zusammenhängender Kritikpunkt bezieht sich auf die fundamentale Sichtweise, „nur“ mithilfe der niedrigeren Ebenen (also nur mithilfe konstitutiver

[1] Im Folgenden werde ich dieser Argumentationslinie jedoch nicht nachgehen. Nur zwei Hinweise dazu: Erstens, als ein Indiz für die Zentralität der reduktiven Erklärungen innerhalb des mechanistischen Ansatzes kann die Tatsache gelten, dass die so genannten „konstitutiven Erklärungen“ (d.h. im Wesentlichen reduktive Erklärungen) auch bei Craver (2007) einen enormen Stellenwert und Raum einnehmen. Diese scheinen das Kernstück des mechanistischen Ansatzes zu bilden. Zweitens scheint kein Reduktionist bestreiten zu wollen, dass auch die *Organisation* der Teile für die Erklärung eine Rolle spielt.

Mechanismen) bzw. mit der schrittweisen Reduktion auf eine Basistheorie/Basiswissenschaft hin können wir „echte“ Vereinheitlichung in den Wissenschaften erreichen (Craver 2007: Kapitel 7).

> The reduction model is focused exclusively on explanations that appeal to lower-level mechanisms, and so does not accommodate [important, B.G.] aspects of the explanatory unity of neuroscience. (Craver 2007: 231)

Diese Art des Fundamentalismus könnte man „vereinheitlichenden“ Fundamentalismus nennen.

Eine metaphysische Ausprägung dieses Fundamentalismus (den man „metaphysischen“ oder „kausalen Fundamentalismus“ nennen könnte) und damit ein vierter Kritikpunkt ist in den aktuellen Debatten innerhalb der Philosophie des Geistes anzutreffen. Hier wird behauptet (z.B. Kim 2005), dass es echte kausale Kraft nur auf der niedrigsten Ebene geben könne, weshalb höhere Ebenen keine eigene Kausalkraft besäßen (und damit weniger real seien). Mithilfe solcher Überlegungen wird oft für einen *reduktiven Physikalismus* argumentiert (Kim 2005). Craver (2007: Kapitel 6) versucht zu zeigen, dass diese Art des Fundamentalismus nicht angenommen werden muss. Er schreibt:

> I defend the view that higher mechanistic levels are explanatorily relevant. I also show that realized phenomena (that is, phenomena at higher levels of realization) are often causally, and so explanatorily, relevant for many of the explanantia of interest to neuroscientists. (Craver 2007: 195)

Nach diesem Überblick über die wichtigsten Kritikpunkte am Reduktionismus will ich auf den dritten Kritikpunkt („vereinheitlichenden Fundamentalismus“) näher eingehen (vgl. Craver 2005; Craver 2007: Kapitel 7).

### *4.2 Die Einheit der Neurowissenschaft folge nicht aus der „reduktiven Vereinheitlichung“*

Craver beschreibt die alternative Art der Vereinheitlichung von (ME) (d.h. alternativ zum „vereinheitlichenden Fundamentalismus“ des Reduktionismus) als „intralevel“ und „interlevel integration“ zwischen verschiedenen Feldern. Die Neurowissenschaften erhalten Input u.a. aus folgenden Feldern: Anatomie, Biochemie, Informatik, Molekularbiologie, Elektrophysiologie, experimentelle Psychologie, Pharmakologie, Psychiatrie, etc. (vgl. Craver 2007: 228) Die verschiedenen Felder, die an der Integration beteiligt seien, sind gemäß Craver autonom, haben ihre eigenen wichtigen Probleme und operieren mit unterschiedlichen Techniken und Hintergrundannahmen. Damit könne ein möglicher Mechanismus unabhängige Evidenz aus den verschiedenen Feldern erhalten.

Es gibt laut Craver drei Eigenschaften der Reduktion, die nicht zur Mosaik-artigen Einheit der Neurowissenschaften passen, weshalb der reduktive Ansatz keine adäquate Beschreibung liefere.

Erstens, Reduktion könne nicht mit „aufwärts-schauenden“ Aspekten umgehen, denen in den Neurowissenschaften eine wichtige Rolle zukomme (Craver 2007: 232). Craver zeigt, wie Erklärungen in den Neurowissenschaften verschiedene Perspektiven einnehmen (*multilevel explanations*): top-down, bottom-up (vgl. auch die Aufsatz-Titel von Bechtel & Abrahamsen 2008; Bechtel 2009; Bechtel 2010). Die Perspektiven zwischen den Ebenen sind auch in kontrollierten Experimenten anzutreffen (z.B. gibt es *Interferenz-Experimente*: diese sind bottom-up hemmend, z.B. Läsionsstudien; *Stimulationsexperimente*: diese sind bottom-up stimulierend, z.B. Transkranielle Magnetstimulation (TMS); *Aktivierungsexperimente*: diese sind top-down aktivierend, z.B. fMRI-Studien). Da die Mikroreduktion nur eine fundamentalistische Perspektive nach unten einnehme, fehle hier die nach oben gerichtete Sichtweise:

> Oppenheim and Putnam recommended reduction as a working hypothesis for building the unity of science. To support this thesis, they appeal to historical evidence of reductive trends in science. But their argument is flawed because they overlook evidence of upward-looking trends. (Craver 2007: 246)

Zweitens, Formen der Vereinheitlichung auf einer Ebene (intralevel) würden ignoriert. Da die Mikroreduktion nur auf Reduktion *zwischen* den Ebenen als vereinheitlichend fokussiert, kann sie die ebenfalls stattfindende „intralevel integration" nicht erklären. Ein Beispiel von Craver:

> hippocampal synaptic plasticity was not discovered in a top-down, reductive search for the neural correlate of memory; rather, it was noticed during an intralevel research project in which anatomical and electrophysiological perspectives were integrated. (Craver 2007: 240)

Drittens, im Beispiel zur Erforschung des Phänomens der Langzeitpotenzierung (LTP), welches mit Prozessen des Lernens und des Gedächtnisses in Zusammenhang gebracht wird, wurde Reduktion laut Craver als Ziel aufgegeben (entgegen der empirischen These, dass die Wissenschaft nach dem reduktiven Ansatz verfahre, vgl. Craver 2007: 237, 245). Mit der darauf folgenden Suche nach Mechanismen gelangte man zu fruchtbareren Thesen und Erkenntnissen. LTP wurde im Verlauf der Forschung nicht mehr als identisch mit Lernen und Gedächtnis angesehen, sondern eher als Komponente des Mechanismus für Lernen und Gedächtnis. Insgesamt lässt sich mit Craver auch sagen, dass Reduktionen in der Neurowissenschaft selten zu finden sind.

Die drei Kritikpunkte von Craver gegen die „reduktive Sicht der Vereinheitlichung" seien nochmals zusammengefasst:

(a) Reduktion könne nicht mit „aufwärts-schauenden" Aspekten umgehen.

(b) Formen der Vereinheitlichung auf einer Ebene (intralevel) würden ignoriert.

(c) Anhand des Beispiels der Langzeitpotenzierung lasse sich zeigen, dass Reduktion als Ziel aufgegeben wurde.

Als abschließende Bemerkung zum „vereinheitlichenden Fundamentalismus" – bevor ich zum nächsten Abschnitt komme – noch folgendes Zitat:

> What seems right about this view of the unity of science is that higher-level (and higher-order) phenomena can *often* be explained in terms of lower-order phenomena. But this is not an argument for the thesis that the unity of science is achieved by reduction to a common lowest level. (Craver 2007: 268, Hervorhebung B.G.)

## 5. Klassische Reduktion ist mit vielen Ideen der mechanistischen Erklärung kompatibel

Im letzten Teil dieses Aufsatzes möchte ich einige Ideen von Nagel (1961) einbringen, dem *locus classicus* der Theorienreduktion. Dabei möchte ich gegen die obigen zwei Kritikpunkte (a) und (b) aus 4.2 argumentieren.

Zuerst zu (b), der Kritik, dass „intralevel" Formen der Vereinheitlichung ignoriert würden: Zwar sagen Oppenheim und Putnam (1958), dass nur reduktive Beziehungen *zwischen* den Ebenen der Vereinheitlichung dienen. Sieht man sich jedoch entsprechende Textstellen in Nagel (1961) an, so muss man feststellen, dass es keine Beschränkung des Modells diesbezüglich gibt, d.h. Reduktion ist unspezifisch und zunächst nicht auf Ebenen bezogen. Es muss erwähnt werden, dass Nagel keine metaphysischen Aussagen machen wollte und kein so generell vereinheitlichendes Modell wie Oppenheim und Putnam aufgestellt hat. Jedoch

kann man aus seinen Texten verschiedene Vorstellungen, wie *intralevel integration* funktioniert, generieren.

Oft wird in der Diskussionsliteratur zu Nagel zwischen synchroner und diachroner Reduktion unterschieden. Bei der *synchronen* Reduktion werden Theorien verschiedener Ebene zur gleichen Zeit verglichen (ein Beispiel wäre die viel diskutierte Reduktion der Thermodynamik auf die statistische Mechanik). Bei der *diachronen* Reduktion werden zeitlich aufeinander folgende Theorien eines Gegenstandsbereichs (also einer Ebene) verglichen (ein Beispiel wäre die Reduktion der geometrischen Optik auf die Maxwellsche Theorie des Elektromagnetismus oder die Reduktion der klassischen Mechanik auf die spezielle Relativitätstheorie). Eine Vereinheitlichung auf einer Ebene kann so aussehen, dass Theorien für ähnliche, benachbarte Phänomene entwickelt wurden (und zwar zunächst unabhängig voneinander), die dann in ein einheitliches Modell integriert werden.

Nagels Beispiel für eine homogene Reduktion (bei der die wesentlichen Begriffe sich zwischen den Theorien nicht unterscheiden; vgl. Nagel 1961: 338) ist die Theorie der Mechanik, die zuerst nur für die Bewegungen von Punktmassen formuliert worden war. Letztlich wurde sie dann auch auf Bewegungen von starren sowie verformbaren Körpern ausgedehnt. Es kann dadurch eine Vereinheitlichung erreicht werden, dass mehrere Phänomene (anfangs als zu verschiedenen Arten gehörend gedacht) später mit einem einheitlichen Modell beschrieben werden können (und somit unter die gleiche Art Phänomen subsumiert werden). Dazu Nagel:

> A theory may be formulated initially for a type of phenomenon exhibited by a somewhat restricted class of bodies, though subsequently the theory may be extended to cover that phenomenon even when manifested by a more inclusive class of things. (Nagel 1961: 338)

Das heißt, die klassische Reduktion schließt so etwas wie *intralevel integration* nicht aus.

Nun zu (a), der Kritik dass Reduktion nicht mit *upward-looking* Aspekten umgehen könne. Dazu lässt sich Folgendes erwidern: Nagel betont bei der Besprechung seines Reduktionsmodells, dass nicht nur die beiden formalen Bedingungen der *Verknüpfbarkeit* und der *Ableitbarkeit* erfüllt sein müssen, sondern ebenso verschiedene informelle Bedingungen. Beispielsweise müssen die Brückengesetze durch empirische Evidenz möglichst aus mehreren unabhängigen Quellen *gut gestützt* sein. Weiterhin muss durch die Reduktion der einen Theorie auf die umfassendere Theorie eine *fruchtbare Weiterentwicklung* der Theorien in Aussicht stehen, damit also auch der *reduzierten* Theorie. Dies klingt nun nicht so, als würde hier die *upward-looking* Perspektive hin zur reduzierten Theorie höherer Ebene vernachlässigt, denn durch die Integration in die reduzierende Theorie wird die reduzierte erweitert und auf eine fruchtbarere Grundlage gestellt. Durch die Reduktion der Thermodynamik auf die statistische Mechanik beispielsweise wurden eine Menge bis dahin als unabhängig geglaubter Gesetze der Thermodynamik und anderer Teile der Physik in ein einheitliches System integriert. Dazu zwei Zitate von Nagel:

> The reduction of thermodynamics to mechanics [...] paved the way for a reformulation of gas laws so as to bring them into accord with the behaviors of gases satisfying less restrictive conditions; it provided leads to the discovery of new laws; and it supplied a basis for exhibiting relations of systematic dependence among gas laws themselves, as well as between gas laws and laws about bodies in other states of aggregation. (Nagel 1961: 359)

> In consequence, the reduction of thermodynamics to kinetic theory not only supplies a unified explanation for the laws of the former discipline; it also integrates these laws so that directly relevant evidence for any one of them can serve as indirect evidence for the others, and so that the available evidence for any of the laws cumulatively supports various theoretical postulates of the primary science. (Nagel 1961: 361)

Es lässt sich also festhalten, dass die Nagel-Reduktion sehr wohl mit der *upward-looking* Perspektive umgehen kann und sie mit berücksichtigt. Eine Weiterentwicklung dieser (beibehaltenen) Züge der klassischen Reduktion finden wir im New Wave Reduktionismus, einer Position, die explizit die Ko-Evolution von Theorien verschiedener Ebene betont (also in der synchronen Reduktion die Ko-Evolution von reduzierter Theorie höherer Ebene und reduzierender Theorie niedrigerer Ebene; vgl. Hooker 1981; Churchland 1986; Bickle 1996; Bickle 1998).

Insgesamt kann Nagels Bedingung der Verknüpfbarkeit für das Ziel von Vereinheitlichung verteidigt werden und ist mit der *intralevel/interlevel integration* des mechanistischen Erklärungsansatzes vereinbar.

## 6. Resümee

Im vorliegenden Aufsatz wurde Folgendes gezeigt: In einigen Texten, in denen der mechanistische Erklärungsansatz beschrieben wird, wird dieser so dargestellt, dass er ein reduktiver bzw. reduktionistischer Ansatz ist. In anderen Aufsätzen wird (ME) als Gegenentwurf zum Reduktionismus bestimmt. In diesem Text habe ich zu zeigen versucht, dass einerseits die Ideen der *intralevel* und *interlevel integration* in den Neurowissenschaften, welche mit (ME) zusammengehen, und andererseits einige Ausführungen von Nagel als klassischer Referenz zur Reduktion gar nicht so weit auseinander liegen. Die Nagel-Reduktion scheint in manchen Punkten ein viel liberalerer Ansatz zu sein als weithin behauptet (und kann durchaus vom Programm der Mikroreduktion abgegrenzt werden). Diese Passung der Nagel-Reduktion mit dem mechanistischen Erklärungsansatz kann nun als ein Indiz dafür gelesen werden, dass es sich beim mechanistischen Ansatz um einen reduktiven Ansatz handelt, auch wenn die Titelfrage des Aufsatzes nicht letztgültig beantwortet werden kann. Dazu sind noch weitere Argumentationsschritte nötig.

**Bettina Gutsche**

Universität Mainz
bettinagutsche@gmx.de

## Literatur

Bechtel, W. 2001: „The Compatibility of Complex Systems and Reduction: A Case Analysis of Memory Research“. *Minds and Machines* 11, 483–502.

— 2007: „Reducing Psychology while Maintaining its Autonomy via Mechanistic Explanations“, in M. Schouten und H. Looren de Jong (Hrg.): *The Matter of the Mind: Philosophical Essays on Psychology, Neuroscience and Reduction,* Oxford: Basil Blackwell, 172–98.

— 2009: „Looking Down, Around, and Up: Mechanistic Explanation in Psychology“ *Philosophical Psychology* 22, 543–64.

— 2010: „The Downs and Ups of Mechanistic Research: Circadian Rhythm Research as an Exemplar“, *Erkenntnis* 73, 313–28.

Bechtel, W. und Abrahamsen, A. 2005: „Explanation: A Mechanist Alternative“, *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, 421–41.

— 2008: „From Reduction Back to Higher Levels“, Proceedings of the 30th Annual Meeting of the Cognitive Science Society.

Bechtel, W. und Hamilton, A. 2007: „Reduction, Integration, and the Unity of Science: Natural, Behavioral, and Social Sciences and the Humanities“, in *T. A. F.* Kuipers *(Hrg.): General Philosophy of Science: Focal Issues, Amsterdam;* Heidelberg: Elsevier, 377–430.

Bickle, J. 1996: „New Wave Psychophysical Reduction and the Methodological Caveats“, *Philosophy and Phenomenological Research* 56, 57–78.

— 1998: *Psychoneural Reduction: The New Wave*. Cambridge (MA): MIT Press.

Birbaumer, N. und Schmidt, R. F. 2006: *Biologische Psychologie*. 6. überarb. Aufl. Heidelberg: Springer.

Churchland, P. S. 1986: *Neurophilosophy: Toward a Unified Science of the Mind-Brain*. Cambridge (MA): MIT Press.

Craver, C. F. 2002: „Interlevel Experiments and Multilevel Mechanisms in the Neuroscience of Memory“, *Philosophy of Science* 69, S83–S97.

— 2005: „Beyond Reduction: Mechanisms, Multifield Integration and the Unity of Neuroscience“, *Studies in History and Philosophy of Biological and Biomedical Sciences* 36, 373–95.

— 2007: *Explaining the Brain*. New York: Oxford University Press.

Craver, C. F. und Bechtel, W. 2007: „Top-Down Causation without Top-Down Causes“, *Biology and Philosophy* 22, 547–63.

Fazekas, P. und Kertész, G. 2011: „Causation at Different Levels: Tracking the Commitments of Mechanistic Explanations“, *Biology and Philosophy* 26, 365–83.

Fodor, J. A. 1974: „Special Sciences (Or: The Disunity of Science as a Working Hypothesis)“, *Synthese* 28, 97–115.

Hooker, C. A. 1981: „Towards a General Theory of Reduction“, *Dialogue* 20, 38–60, 201–36, 496–529.

Kim, J. 2005: *Physicalism, or Something Near Enough*. Princeton: Princeton University Press.

Machamer, P., Darden, L. und Craver, C. F. 2000: „Thinking about Mechanisms“, *Philosophy of Science* 67, 1–25.

McCauley, R. N. 1986: „Intertheoretic Relations and the Future of Psychology“, *Philosophy of Science* 53, 179–98.

Nagel, E. 1961: *The Structure of Science*. London: Routledge.

Oppenheim, P. und Putnam, H. 1958: „Unity of Science as a Working Hypothesis“, in H. Feigl und M. Scriven (Hrg.): *Concepts, Theories and the Mind-Body-Problem. Minnesota Studies in the Philosophy of Science. Vol. 2.* Minneapolis: University of Minnesota Press, 3–36.

Schaffner, K. 1967: „Approaches to Reduction“, *Philosophy of Science* 34, 137–47.

Schandry, R. 2003: *Biologische Psychologie*. Weinheim: BeltzPVU.

# Phenomenal Concepts - Still Battling the Bewilderment of Our Intelligence

Max Mergenthaler Canseco[1]

In this paper I defend the *Phenomenal Concept Strategy* (PCS) against two recent attacks. First (2) I briefly sketch the relation between the PCS and the Knowledge Argument, highlighting the idea of conceptual fineness of grain. Afterwards (3) I introduce a relevant consideration about public language and go on to explain the deferentialist attack on the PCS (4). Based on the publicity of language and demonstrative reference (5) I claim that the worries that externalism expresses can be explained away. By arguing that Phenomenal Concepts refer demonstratively (6) I consider some final objections and go on to draw the conclusion that the PCS is far from being refuted but instead offers a promising line of research (7).

## 1. Introduction

Since the 80's, many different anti-physicalist arguments have been proposed. Often they depart from a conceivable epistemic or conceptual gap between physical and qualitative experiences and go on to refute physicalism on *a priori* grounds. Physicalists have come up with various strategies to counter these arguments. One promising and prominent defense argues, that these gaps can be explained through a correct understanding of how phenomenal concepts work. And further, that the puzzlement is rather the result of a conceptual and not, as is suggested, an ontological difference. In short, the strategy offers a pleasingly deflationary account of what are probably the main problems in contemporary philosophy of mind. If this strategy proves to be successful, it offers a powerful reply to many strong anti-physicalist arguments.[2] Following Stoljar we will call this line of thought the *Phenomenal Concept Strategy* (PCS).

I will concentrate on the PCS as an answer to *Knowledge Argument*[3] and the purpose of this paper will be to defend the claim that this strategy is the best alternative despite some recent critiques. I will therefore discuss and reject some recent objections that claim to refute the PCS. I will however argue that the recent critiques are based on a wrong conception of PC. I will propose that if we conceive of PC as referring demonstratively and we clearly distinguish fine-grained senses from coarse-grained extensions, we can conserve an appealing version of the PCS that resolves positively the strong anti-physicalist intuitions that the KA creates.

[1] Thanks to David Papineau, Jonathan Cohen, and Holm Teten's research colloquium for very helpful comments.

[2] Besides the Knowledge Argument (Jackson, 1982, 1986) which this paper is about, I'm thinking about the explanatory gap (Levine, 1983; Chalmers, 2007) and the problem of the appearance of contingency between the relation of phenomenal and physical states (Kripke, 1981; Chalmers, 1997).

[3] The KA was originally proposed by Jackson (1982), however the recent literature offers many new formulations between others Jackson (see 1986); Nida-Rümelin (see 1998, 2002); Stoljar (see 2005).

## 2. PC, KA and Fine-grained Concepts

Let us begin by briefly stating what a Phenomenal Concept (PC) is supposed to be, and how the Phenomenal Concept Strategy (PCS) is supposed to save physicalism from the Knowledge Argument (KA). A PC is the concept of a particular type of sensory or perceptual experience, where the notion of experience is understood phenomenologically (see Stoljar 2005), PC's are concepts of our qualitative experiences. PC's are normally thought to be the concepts that we use to think about our qualitative states in a 'direct' way. Although there are many different definitions and versions of what constitutes a phenomenal concept, the following definition is compatible and pretty general.

**Phenomenal Concept Criterion (PCC)**

The Concept C is a Phenomenal Concept iff:

1. There is some phenomenal experience type E, and some property P, such that experience tokens fall under E in virtue of their relation to P.

2. C refers to P.

3. Under normal circumstances, a human being can possess C only if she has had an experience of type E.

The third clause of the PCC is sometimes referred to as the *Experience Thesis* (ET). It states that to possess a Phenomenal Concept in the right way we need necessarily to undergo a certain qualitative experience which is, in this case, normally caused by seeing colourful objects.[4] This thesis is held to be what distinguishes, in some sense, normal concepts from PC. Throughout this paper I will argue that a slight modification of ET holds.[5]

Let us now turn to the *Knowledge Argument*, to which this sort of concept should give an acceptable answer for physicalism. For the sake of simplicity we will deal with a version inspired by Nida-Rümelin (2002).

**Knowledge Argument (KA)**

**P1** Mary has complete physical knowledge before her release. (Including complete knowledge about human color vision)

**C1** Therefore, Mary knows all the physical facts about human colour vision before her release.

**P2** There is some (kind of) knowledge concerning human colour vision that Mary could not have before her release.

**C2** Therefore (from P2), there are some facts about human colour vision that Mary could not know before her release.

**C3** Therefore (from C1 & C2), there are non-physical facts about human colour vision.

Lets notice that *C3* implies directly that physicalism is wrong, for physicalism is the thesis that the phenomenal, or experiential truths supervene with metaphysical necessity on the physical truths. A well studied consequence of physicalism is that: if $P$ is a statement summarizing all the physical truths of the world and $P_{phen}$ is some phenomenal truth, then the conditional $P \rightarrow P_{phen}$ must necessarily be true. This conditional is sometimes called the *physicalist entailment thesis (Balog, forthcoming)* and sometimes referred as the Psychophysical Conditional (Stoljar, 2005) we will adopt the second term and formulated it as follows:

---

[4]

[5] The modification from ET that I will defend acknowledges that PC refer demonstratively

**Psychophysical Conditional**

$\forall T \Box (P \rightarrow T)$ [6]

The decisive anti-physicalist step is to infer *C2* from *P2*. That is, the core of the Knowledge Argument implies that one can validly infer from the fact that Mary learned something new, that there are non-physical facts. If this inference is valid, physicalism would indeed be refuted. However, as mentioned before, the PCS offers a convenient way of blocking this inference. The idea is to argue that someone can learn something new, as Mary does, in an unproblematic way for physicalism, i.e. something along the lines of getting to know and old fact under a new mode of presentation. And this brings us to the core consequence of the PCS: one can come to know new contents without coming to know new facts.

This conclusion is manly a consequence of the fregean distinction between the extension (*Bedeutung*) and the content (*Sinn*) of concepts (Frege, 1892). According to which it is possible to know new contents without coming to know new facts about the physical world. Considering that contents are fine-grained while facts are coarse-grained explains why identity statements can be informative and have cognitive significance. To name one example, someone can know the fact that "Mark Twain is a writer" and then learn the new content that "Samuel Clemens is a writer". But since Mark Twain and Samuel Clemens are the exact same person, our literature student would not come to know a new fact about the world but just a new content. Frege explains the speaker's grasp of the sense of a singular term as having its referent presented in a particular way, or as having a disposition to think about the referent in a certain way Byrne (see 2011).

And the PCS claims that the same fregean consideration applies *mutatis mutandis* for the Mary case. When she gets out of the monochromatic room she comes to know a new content but not a new fact, for the PC in question could refer to a brain state identical to a qualitative state which is now given under a new *mode of presentation* (Loar, 2007; Frege, 1892). But, the PCS argues, in order to be able to deploy the PC that conforms to the new knowledge she has to undergo a certain experience. So even if the identity $Experience_{red}$=$BrainState_{red}$ holds, Mary could come to learn that 'That is what it is like to see red', although she already knows the fact that 'Brain $State_{red}$ is what it is like to see red'. This invalidates the problematic inference from P2 to C2, for it shows that one can come to gain new knowledge without there being non-physical facts, showing that the problem is conceptual and not metaphysical. It is important to understand that the PCS is however not proposing a comprehensive answer to the question of what phenomenal states are, but is rather thought to be a flexible strategy to counter the anti-physicalist intuitions.[7] Notice that the PCS is not committed to *identity theory* about mental states, it is in fact compatible with Functionalism and other naturalistic theories of mind, and notice furthermore, that the PCS does not offer an argument for physicalism, but rather assumes it's plausibility and defends it against prominent anti-physicalist arguments by showing that the problems posed are nothing more than linguistic confusions.

## 3. Publicity of Colour Words

As anticipated before one common mistake while critizing the PCS is the confusion between true phenomenal concepts and public language words like the physical concept *red* and the psychological concept *pain*. The meaning of those words, as Wittgenstein famously explained,

[6] Where T are all the possible true sentences.

[7] Overlooking this fact has brought Hill (2009) to take the strategy to be claiming that undergoing a qualitative experience always requires some kind of conceptualization. However, as we will se later, this claim is unjustified, and will not be treated here.

is not essentially related to the qualitative experiences of our inner life. It seems clear that outside the philosophy seminar we would agree that the sky is blue and that granny smith apples are green, regardless of our inner qualitative experiences. Even border-line cases satisfy this intuition. A subject suffering from color blindness, for example, would not say that the green things are rather and in reality yellow, but he will rather admit that he experiences them as yellow but they are really green. In this case the person has troubles applying the right concept. Notice furthermore that qualitative experiences are not necessary for deploying language of feelings and perception; Philosophical Zombies for example, are completely capable of deploying correctly colour words and identifying when someone is very likely in pain. The case of the inverted spectrum also confirms this line of thought, for even if our inner qualitative experiences were different, the meaning of the words we use to talk about colors would stay the same and would apply to the same objects.[8]

The above confirms what we said before about public language. As we said before, we can learn that the meaning of the word *'red'* applies to red things disregarding our qualitative experience. Depending on our preferred semantic theory we can say that the meaning of the predicate red is the set of objects that we call red, disregarding how they appear to us or that the meaning is determined by the usage of the predicate in a linguistic community, also independently from private qualitative experiences. To stress the point I have been making, let me quote a famous passage of Wittgenstein:

> Look at the blue of the sky and say to yourself "How blue the sky is!"–When you do it spontaneously–without philosophical intentions–the idea never crosses your mind that this impression of colour belongs only to you. And you have no hesitation in exclaiming that to someone else. And if you point at anything as you say the words you point at the sky. I am saying: you have not the feeling of pointing-into-yourself, which often accompanies 'naming the sensation' when one is thinking about 'private language'. Nor do you think that really you ought not to point to the colour with your hand, but with your attention. Wittgenstein (1973, §275)

When we speak we do not detach the color-impression from the object. Only in very special scenarios, for example in Mary's release or in philosophy conferences we are tempted to think that we use one word to mean at one time the color known to everyone–and another word for the 'visual impression' which we get while staring at colorful things. And this suggests that the meaning of the word 'red' is public and partially independent of qualitative experiences. It also concedes that we have the ability to refer to the *private* qualitative experiences; the point is that that is not normally how we use color words. In fact, the concepts we use to refer to our qualitative experiences are PC and even if sometime people misuse the public color words to refer to inner experiences, we should distinguish both concepts.

But what then is special about phenomenal concept? Which concepts refer not to the things outside to us but to the qualitative experiences of our inner life and necessitate, in order to be deployed, that we undergo a certain experience, i.e. what is a true PC? I will argue that the special way in which PC refer to phenomenal experiences is through demonstratives. E.g. "*That* is what is like to see red". We could also introduce a new term $\text{Red}_{\text{phen}}$ to refer to them. However, I will claim that although both concepts have the same extension, they have different senses. As we will see, this fact seems to be ignored very often.

---

[8] J. Cohen (in conversation) pointed out, that this consequence is compatible with two possibilities: i.) Meanings are understandable independently of qualitative experiences, ii.) meanings are rather closely tied to qualitative experiences but the qualitative experiences are equal across the population. It seems to me, that i.) is the most plausible alternative. I showed that even subjects with varying qualitative experiences like Zombies and people with Color Blindness are fully capable of understanding and using words like red and pain. Besides, the best explanation for the fact that qualitative experiences don't vary across the population in a significant matter is best explain through the physical similarly of human kind.

## 4. Deferentialism about PC

Before going into more details about demonstrative reference of qualitative experiences, let us go through the arguments that question the existence of phenomenal concepts and claim that the *Experience Thesis* is false by arguing that Mary could have possessed PC by means of deference prior to her release. This critical strategy has been recently developed by Ball (2009) and Tye (2009) and is based in the results of *semantic externalism*, which was famously defended by Putman (1975) and Burge (1979). Oversimplified, this theory states that *'meanings' just ain't in the head* but that the semantic content is rather constituted externally by the social and factual nature of the external world, by means of social, cultural and/or linguistic interactions. In order to discuss the argument against the PCS it suffices to define what the externalist understands under *content*, *concepts* and *conception*. *Contents* are the objects of *de dicto* propositional attitudes such as beliefs, desires, and thoughts. *Concepts* under this account are those mental representations of which internal beliefs and other mental representations with an internal structure say propositional attitudes, are composed. And the *Conception* of a concept is the collection of beliefs associated with it.

One of the most significant consequences of semantic externalism, is that someone may possess a concept although her conception is not completely right or exact. One does not need *concept mastery* to possess a concept. Although I agree generally with this conclusion of semantic externalism I will argue that some concepts, i.e. demonstrative PC, are significantly different. Let us state, however, the externalist thesis about concept possession which grounds the critique against the PCS:

**Concept Possession (CP)**

S possesses the concept *C* if s is able to exercise (even if vaguely or incorrectly) *C* in her thoughts.

CP amounts to the claim that it is sufficient to possess a concept if one is able to grasp propositions that contain the concept, or think contents of which the concept is a component. Concept possession is rather liberal. To exemplify CP let us go over the famous case of Alfred (Burge, 1979). Alfred possesses the concept ARTHRITIS, but Alfred does not know what is medically common knowledge, namely, that arthritis is pain in the joints and not in the limbs, as Alfred thinks when he claims that he has arthritis in the thigh. Since ARTHRITIS is a medical term, doctors have a more sophisticated and complete conception of the concept ARTHRITIS. The ''experts'', to which laymen like Alfred defer, often possess conceptual mastery or at least show a more accurate conception of the concept. However one can possess a concept and still be grossly wrong about its extension, and even about its constitutive a priori truths. On the contrary, *conceptual mastery* excludes such a vague conception (see Sundström, 2008). If CP holds for every concept without exception, it is obvious that the PC Strategy is condemned to fail.

At this point I would like to make a general criticism of CP. It seems intuitively clear that Alfred possesses the concept Arthritis. However one could think of many cases where it is not clear if a subject possesses a concept. For example if someone says that Beeches are pebbly and sandy, we would not agree that S possess the concept Beech. The criterion to decide if someone possesses a concept is vague. It seems that if the subject is able to give a critical amount of default inferences we would agree that he possesses the concept. However it seems unclear what this critical mass is. Notice further that there are some conceptions (default inferences) that S would be able to make about his concept Beech (he actually means Beach) that are correct for the actual concept Beech. For example *Beech belongs to nature, Beech is a study object of some scientists, If something is a Beech then it has a mass, Beech is an essential part of some ecosystems*. For lack of a better name, let's call this kind of inferences

*Kind-Inferences*[9], for they are the product of identifying some object with a Kind whose members share many properties. Clearly, Kinds can be more and more specific. However notice the important fact that S would not be able to produce informative non-default inferences entailing for example definite descriptions or even more concrete specifications like *Beech is part of the ecosystem Forest*, *That there is a B*eech. I think Ball is too quick in saying that CP is just invalid in cases of extreme confusion or insanity.

This point is not irrelevant because as it turns out Mary is going to be able to deploy phenomenal concepts like Red*phen* in *Kind-Inferences* or via linguistic report. However she would not be able to deploy the co-extensional but fine-grained demonstrative PCs, because the required *mode of presentation* would not be available.

*There are no PC*

Following this line of reasoning and extending CP to PC, the mentioned opponents of the PCS will claim to have established the truth of the following thesis:

**Anti PC**

Mary does possess PC in her room prior to the release.

She could, they argue, for example come to posses PC in her room through interaction with her experienced colleagues or through the lecture of different reports of normal speakers.[10] In order to see if this thesis holds, let us analyze what Mary learns when she comes out of the room and sees a red object. Mary would express her alleged new knowledge in the following way:

(1) That is what it's like to see red. (Where 'that' refers to an experience of red, to which Mary is attending in introspection, or to some feature of such an experience.)

The debate concerning whether PC are deferential or not will be highly dependent of what we claim PC are in the above sentence. As(Ball, 2009) mentions the plausible candidates in (1) are '*that*', '*red*', or '*what it's like to see red*'. Surprisingly he very quickly dismisses the demonstrative '*that*', arguing that Mary could possess such an indexical concept in her room. He then runs an argument that is supposed to refute the PCS claiming that the PC in question is the word *red*. I will argue that his refutation of the PCS does not apply if we sustain that the PC in question is the demonstrative '*that*' and not the public color word red[11]. Deferentialists give the following sentences as proof by demonstration of PC that Mary could have possessed prior to her release:

(1) Ripe tomatoes typically cause experiences of red.

(2) What it's like to see red resembles what it's like to see black more than it resembles what it's like to hear a trumpet playing middle C.

(3) If x is a number then x is not what it's like to see red.

[9] This concept is not supposed to be concrete and simple definable. Indeed it is very vague too. However it will do the work since it requires just a simple understanding of what a Kind is. If the reader is extremely bothered with this concept, he might use *a priori* inferences in the sense proposed by (Stoljar, 2005), I claim that this conception does also the work.

[10] Notice that according to this view she could also possess PC prior to her release through scientific investigations. However we will just concentrate on acquisition through the community. My critique is however applicable to the other cases *mutatis mutandis.*

[11] However, I grant him that if we hold that 'red' is the PC in question, we would have to give up PCS. Since Mary could have indeed posses that concept prior to her release as we explained before. But it as we showed above red is not a PC.

(4) That is not what it's like to see red. (Where 'that' refers to some experience which the speaker is introspecting, or to some feature of such an experience.)

(5) Seeing red is a phenomenal state.

There is no obvious reason to deny that Mary could express the sentences 1, 2, 3 and 5 prior to her release. But we still owe an account of how exactly Mary came to possess the PC (let's say $Red_{phen}$) prior to her release that does not invalidate the *ET*. Arguing that in this case what happens is that Mary possesses some other concept THRED and the false metalinguistic belief that "'red' expresses normally the concept RED" and that therefore what Mary possesses prior to her release is a non-phenomenal concept is invalid. For arguing that, because RED lacks some features that the concept $Red_{phen}$ has, there is no significant concept type of which RED and $Red_{phen}$ are both tokens is absurd. Remember Alfred's case. His inaccurate possessed concept Arthritis and the real concept ARTHRITIS have different features. However this would never allows us to infer that there is no significant concept type of which Alfred's arthritis and the doctor's arthritis are both tokens, since both refer to ARTHRITIS.

However it also seems that this is not problematic since it confirms our prediction about sentences Mary could know via report or Kind-Inference. And we have a coherent way of claiming that the special concept in turn is not Red or even $Red_{phen}$ but rather a concept which refers demonstratively. Let's go over the sentences to see that indeed our predictions got confirmed. It seems that Mary could know perfectly well that the qualitative experience that she has when she sees a number or hears a sound is not what it is like to see red. But she would know because she has seen numbers, heard sounds and knows the experiences of red are not properties of those experiences.[12] Notice that even Zombies, which lack phenomenal character, could express 1-6. However, maybe we are too quick. Then it could be the case that Mary suffers from synesthesia but does not know it. In this case she would come to realize after her release that the proposed instantiation of 4 is false and not true as she thought. However this argument seems to violate *P1* since, if she knows all the physical facts, she would certainly be able, via analyzing her own brain, to know for example that her grapheme-recognition area is cross-activated with V4. And that therefore she is a Grapheme-color synesthete.

To motivate the demonstrative account of PC and the difference it bears to this one, consider that sentences with indexicals are not acquirable via report since sentences including demonstratives are not disquotational and they also pose a problem for Kind-Inferences as we showed above. Notice however that the New-Knowledge sentence 1 entails such a demonstrative. So, the question arises again: could Mary have, as the deferentialist claims, entertained 1 prior her release?[13] I strongly belief she could not. For, what would be the demonstrative referring to if Mary can't demonstrate neither to the qualitative experience nor to the physical fact identical to it? It seems rather, that Mary simple does not posses the demonstrative in the right way before seeing colors.

To conclude however from the deferentialist case discussed that there are no such things as concepts that necessarily require a particular experience to be possessed is wrong. What we are allowed to conclude is that, at least in the case of red, we have solid grounds to claim that red is deferentially acquirable. But this is something we accepted and even motivated from the beginning. Our claim is that the special PC are in fact demonstratives. In order to show why the critique does not work when demonstrative are involved, let us briefly sketch what a demonstrative is.

---

[12] She knows somehow the general kind "phenomenal experience" and the more particular kind "accustic" experiences as well as the kind "numbers"; whose members obviously do not have colors.

[13] Other sentences she could not have enteratin are "That is what I felt when I saw red the first time" or "I never felt that before". Where 'that' refers to a phenomenal experience in a special way

## 5. Demonstrative Reference

Going into the rich philosophical analysis of indexicals goes far beyond the boundaries of this paper. Let us state, however, the things that are necessary for the point I want to make. First as Braun (2010) explains, indexicals are those linguistic expressions whose reference shifts from context to context: some examples are 'I', 'here', 'now', 'today','he', 'she', and 'that'. Famously Kaplan argues that indexicals have two sorts of meaning. Following the classic terminology, we will distinguish the static *linguistic meaning* or single character of the concepts and the *varying content* which is relative to the contextual factors such as time, location, and intentions of the speaker. Furthermore, Kaplan (1989) distinguishes between two different sorts of indexical, *pure indexicals* and *true demonstratives*.[14]

The demonstratives include words like 'he', 'she', 'his', 'her', and 'that', while the pure indexicals include words like 'I', 'today', 'tomorrow', 'actual', 'present'. The difference between the two types of indexicals is how their references and contents are fixed. The reference and content of a pure indexical in a context is fixed independently of the speaker intentions. Vaguely speaking the reference and context of a pure indexical is automatic (see Braun, 2010).

This sort of indexicals is not going to be thematized any further in this paper. We are rather concerned with demonstratives, where the reference and content is not independent from the intention or demonstration that accompanies the speaker's utterance in a certain context. For example, the reference and content of 'that' in a context is partly determined by the pointing gestures of the speaker or by the speaker's intention to refer to a particular object(Braun, 2010).

One relevant peculiarity of indexicals is that sentences that contain them are not disquotational. This is important since that will reaffirm the expected result that Mary could not entertain sentences with demonstrative PC. To illustrate what it means to say that demonstratives are non-disquotational consider the following two sentences:

(P) The cat is on the mat.

(P') That is a cat.

Where the difference is that P' but not P includes a demonstrative. Now assume that we hear Mary uttering *P*. In this case we are normally[15] entitled to infer:

(2) Mary beliefs 'the cat is on the mat'.

However, if Mary would utter *P*', we would not be entitled to infer:

(3) Mary beliefs '*that is a cat*'.

The validity of the inference from Mary says that 'P' to (2) is a valid inference because the content of the terms are independent from any demonstration and we assumed Mary is not trying to fool us when she utters a proposition. However, this is not the case of (3), because as we explained the content of 'that' varies depending on the context and the speakers intention. The converse is also the case, given that Mary is in the right epistemic relation, we can deduce *P* from (2). However it would be wrong to infer *P*' from (3). This actually holds for any sentence, if the sentence includes demonstratives reference, disquotation is not always allowed.

To offer a more formalized version of the invalidity of disquotation involving demonstrative, lets us quickly go over a possible formal analyses.[16] Let $\phi$ be a proposition where no indexicals

[14] In this paper referred to as demonstratives

[15] Granted that her utterances reflect what she holds to be the case.

occur. Let furthermore $\nabla$ be any intensional operator of the form "*believes*, *knows*, *fears*, *desires*". Given the sentence $\ulcorner S\nabla\phi \urcorner$ we can trivially infer $\ulcorner S\nabla\phi \urcorner$. Because the meaning of all the logical operators stays fix such an inference is completely unproblematic for every $\phi$ .

If however $\phi'$ entails a demonstrative, the inference from $\ulcorner S\nabla\phi' \urcorner$ to $\ulcorner S\nabla\phi' \urcorner$ is invalid because the meaning of the non logical term 'that' is not fixed. Given that V is a function that assigns a truth value to every non-demonstrative proposition of a formal language: V($\phi$ )= 1 iff $\phi$ is the case. However, if $\phi'$ contains a demonstrative this schema is not longer valid. The truth function of V regarding a language where propositions can entail demonstrative would have to look more like: V($\phi$ )=1 iff $\phi$ is the case & (if $\phi$ entails a demonstrative, then the necessary $c$ is satisfied).[17] Where $\phi$ is every possible well formed proposition and $c$ is a n Tuple of the form $\langle w, t, p, a \ldots n\rangle$ , where $w$ is a world, $t$ is a time index, $p$ a 3 dimensional index of the form $\langle x, y, z\rangle$ , an $n$ other possible factors that would determined the demonstration and need not to be established here.[18]

This shows that we are just able to disquotate propostions including demonstratives if c is satisfied. It also shows that we are able to disquotate from intensional contexts like, 'S justified beliefs "$\phi$ ", to $\phi$ or from S says that 'P' to S beliefs that 'P'. This is important because one can acquire knowledge that does not entail demonstratives by report, but not knowledge that includes demonstratives. As mentioned this analysis reaffirms the wanted results concerning the above 6 sentences where Mary supposedly possessed PC's and Mary could have come to know by report.[19] Given that demonstratives are not disquotational, this would not be possible.

After this short introduction to demonstratives. we will re-evaluate the arguments against the existence of PC's and, as anticipated, if we accept that PC's are demonstrative the critique will be disarmed. While Mary can learn the use of the public concept 'red' deferentially she will not be able to deploy correctly the indexical because she lacks what is the necessary context given by a particular *mode of presentation* (see section 2).

## 6. *That* is a Non Deferential Phenomenal Concept!

So, we are able to claim that the 6 sentences do not represent a threat to the claim that there are PC's or to the Experience Thesis. We can also conclude that the claim that PC's are individuated differently than normal concepts is, contrary to what Ball (2009) thinks, not *ad hoc* at all, but rather the natural consequence of viewing PCs as demonstratives. While color words are indeed public and acquirable in the relevant way through deference, demonstrative PC's refer to something rather private, and are just correctly deployed when the necessary context, in this case, undergoing the experience is satisfied. And obviously one cannot undergo the necessary experience through deference.

The distinction between private words, demonstrative PC's and private qualitative experiences is highly important. Objecting that the sentence Mary utters when she sees colors contains public language words and that therefore all entailed concepts could have been rightly acquired through deference ignores this distinction. For what we are claiming is not

---

[16] This brief formalization has just an illustrative purpose. For a complete construction of logical systems containing indexicals see Kaplan (1979).

[17] Clearly this is naive, but it satisfies our current needs. If the $\phi$ is non-demonstrative then the conditional is vacuously satisfied and V is as always, if however $\phi'$ entails demonstrative it takes $c$ into consideration.

[18] The nature of this needed "extra something" is controversial and we will not discus it here. However two obvious candidates are pointing gestures and speakers' intentions. (For a longer discussion see Kaplan (1979); Braun (2010)

[19] Or Kind-Inference or deference.

that the words of the sentence, say the demonstrative 'That' is private, but rather that the private feeling that accompanies color vision is private and that undergoing that qualitative experience is necessary to deploy correctly the demonstrative in the sentence '*That is what it's like to see Red*'. The only thing that Mary could possess via deference is the *linguistic character* of indexicals. She could indeed be a successful user of indexicals, she could, for example, deploy correctly sentences of the form 'That is a table', 'That is what it is like to feel pain', but not the sentence 'That is an object that I cannot point at' or 'That is what it is like to see red'.

At this point it is important to evaluate one possible objection to the claim I have been defending. Every Indexical, the objection goes, can be substituted by a non indexical term.[20] 'That dog there' can be substituted by 'Fido', if the name of the dog is indeed Fido. So, the objection continues, the demonstrative in turn could be substituted by a non demonstrative term *Q*[21] and *Q* can be learned deferentially. So, again, the experience thesis would be challenged. In order to answer to this objection, let us recall that we established that concepts are fine-grained. So, surely, Mary is going to be able to learn something about Q by interaction with her experienced colleagues. She is going indeed going to possess knowledge involving Q. But the important issue is that she is just going to possess knowledge of a certain sort, namely sentences like 1-6. That is, sentences with a low informative character, that are produced via competent language possession, report or Kind-Inference.

However, as we noted before this knowledge is going to be limited to sentences that do not include demonstratives. And although we want both concepts to refer to the same object, namely a type brain state, they can indeed have a different sense. It seems that Ball ignores for one second that co-extension does not mean identity of concepts. Then although Mary is going to possess certain knowledge of color vision as brain state and certain knowledge of phenomenal concepts per report, she necessarily needs to undergo the qualitative experience to acquire the concept that is co-extensional to *Q* and '*Type Brain State*' but has a different sense because it has a different mode of presentation. Again, until she undergoes the qualitative experience she is not going to be able to refer correctly using the demonstrative 'that', since she lacks the relevant context and intention. This form of direct pointing is wanted while answering to the KA, for it permits that she knows all the physical facts but can still learn a new content. So although she has the concept *Q* she is not going to be able to entertain the thought '*That is red*' until after her release. If indeed different PC can share a referent, then it will be proved that PC are not exhausted by their referents. Mary's new experience is what allows her to know the fine grained content that referred to the old fact she knew.

Let's consider a relevant example of co-extensional terms that have different senses. Alfred could possess the concept Arthritis, (and even achieve concept mastery) without coming to know that Arthritis = inflamed Arthrons in standard human beings.[22] However, although both concepts refer to the same thing they are different. So we can conclude that Mary possesses prior to release a co-extensional concept but she is not able to use the demonstrative PCs. In order to possess them, she needs to undergo a certain experience.

---

[20] Byrne (2011) offers a similar response to this criticism. He agree that subjects can introduce terms to express concepts they entertain which are not expressed by public language terms but he also explains the notion of expression in question is very thin: "In particular it seems that for Mary to be able to conceptualize phenomenal colour qualities in the same was as a normally-experienced person, she needs to do a lot more than to read a term with which another thinker expresses it".

[21] Or Red*phen* for example

[22] This identity statement is of course fictional, however it is sufficient to think that it is possible that science discovers that Arthritis occurs iff X (Where X is a natural kind). If this case does not seem convincing the paradigmatic case can do the work too. It is clear that 'Hesperus' and 'Phosphorus', and 'Mean kinetic energy' and 'heat' are different concepts, although the respective identities are necessarily true and they respectively refer to the same object or natural kind.

We see that the objections that accuse PCS of sustaining a private language or ignoring that demonstratives can be substituted by singular terms, which could be possessed deferentially, are mislead and can be countered: it is simply wrong to follow that since Mary's new sentence is not in a private language she could have possessed PC in the relevant way before leaving the room. For although indexicals can point to an object which can have another name, Mary needs to undergo the experience to correctly deploy demonstrative PCs and learn an old fact in a new way.

*Some Final Objections*

Now we will go quickly over some objections that supposedly show the implausibility of the PCS. We will see that our considerations can disarm this argument too. For example Ball (2009) gives the following argument against the PCS. He invites us to consider the following two sentences

(4) I don't know what it's like to see red.

(5) I know what it's like to see red.

Where (12) is expressed by Mary before she leaves the monochrome laboratory, and (5) after. Ball (2009) then argues that a PC strategist could not affirm that (12) is the negation of (5)if he considers that *red* is the PC in question, for both sentences would involve different concepts. And he continues, how could

(6) I used to wonder what it's like to see red, but now I know.

be possibly true if Mary's new knowledge did not involve already possessed concepts. However, this argument is clearly based on a confusion. In both cases the meaning of red stays the same, as we explained before. It is possible that Mary even possesses the complex concept 'what it's like to see red' and still be surprised that "'that' is what it's like to see red" (where that refers to the brain state as qualitative experience). As an example imagine the following scenario: I point with my hand to the back of the room and refer to something that I don't see by using the demonstrative 'that there'. In this case I will be able to make some inferences about the thing I'm pointing at. Moreover a lot of those inferences will be correct. I could for example know the following sentences:

(7) That there possesses some color. (Given that even transparent objects have some coloration)

(8) That there is located inside the room.

(9) That there has a mass .

(10) If that $x$ there has a mass $m$ and is being pointed at, $x$ is not a number.

(11) $x$ resembles the desk more than it does resemble the sound of rain on a tin roof. (Where $x$ is the $x$ that is being referred to with that there.)

Notice that in these cases, since I'm using the concept correctly in some sentences, this would give us license, according to the externalist position Ball represents, to infer that I possess the concept in question. However, it is clear that it is wrong to believe that I'm not entitled to express the following sentence

(12) I don't know what that is.

And then, after I turn around and see that what I was pointing at was a chair I'm not entitled to express this one

(13) I used to wonder what that is, but now I know it's a chair.

In both cases I somehow possess the concept 'that', although it is obvious that in the first case I do not really know what 'that' is. Then although I'm a competent user of indexicals and also possess the concept chair, before I turned around I could not know if the sentence *That is a chair* was true or not. I could surely do a priori inferences before I turn around, but I'm not capable of saying something relevant about 'that'. Just after I turn around I gain knowledge of what "that" really is.

Some might claim that we are misusing the demonstrative, since it could be actually the case that I did not refer to anything before I turned around. However we can imagine different cases with different grades of indexical abuse. Imagine first a case of complete misuse: someone stands in a room with no Bears (or representations of Bears) and says *I know what that Bear is like.* Even if the weird person in question is normally a competent user of demonstratives, the 'that' in question refers to nothing.

But imagine now a case of minor misuse. For lack of a better example imagine this scenario: In the party game Truth or Dare someone, who chose to be dared, has to engage in some daring activity with someone else. However to follow the strict rules of the game he has to choose his partner randomly. All players stand in line by the wall (whose length is known to everyone) while the dared one is turned around, then he randomly without turning around has to point at someone and say 'You'. In this case, the speaker knows that he is definitely pointing at someone, and he would be able to know the many truths that apply to all persons and that he knows by report, categorical inference or just as a competent speaker. However when he turns around he might well be very surprised. And although he somehow possessed the demonstrative 'You', he was pointing blindly. And he would have the right to express (12) before he turns around and (13) after he realized who we randomly picked.

If someone wanted to argue that a further shortcoming of the PCS is that it would falsely predict that Mary would be surprised about far too many things, for example about the fact that Red*phen* is not a number although she should not. Our distinction would allow us to counter that she would be indeed be surprised to learn that ""That' is what it's like to experience visual information sent to the brain from retinal ganglion cells via the optic nerve to the optic chiasma." But she would not be surprised about the sentences she could know by report that do not involve demonstrative, i.e. the relevant way to refer to qualitative experiences via PC's. To continue with the above suggested scenario, I'm not surprised that 'that' or 'you' did not turn out to be a rational number. Furthermore if she turns out to be synesthetic (and did not know it) she might be surprised that PC refers to red objects and some graphemes. In that case she could say "'That' is what it's like to see red and it is just like seeing some numbers." But given the appropriate understanding of PC, the PCS does not over generate predictions.

So Mary can indeed come to know new contents by correctly deploying demonstrative concepts she could not use before. Given that the demonstrative must not refer to a non-physical object, Mary does not learn any new fact. And this version of the PCS is a coherent and sufficient way to disarm the KA.

## 7. Two Not Worrisome Worries and a Conclusion

There remain basically two worries. The first arises from a thought experiment proposed by Wittgenstein (1973). He invites us to imagine that everyone has a box with a beetle inside. However nobody can see the beetle of the others. The beetle could stand for the private qualitative experience to which we refer via demonstratives. But the problem is that it could be the case that there actually is no beetle in the box. This worry ignores one fact that we have been stressing throughout the paper, that we are not claiming that the demonstratives are private, but maybe just their referents. Like the box possessor we are equipped with a public

language that allows us to make comparisons and descriptions about the things in our private box. Furthermore, this question is subject to the claim that it ignores the dialectical situation. The argumentative burden lies on the shoulders of those skeptics that claim that everyone else could be a zombie, and not with the ones that think that their friends have an inner life similar to theirs.

The second worry argues that since the PCS has not been able to give a precise definition of PC involving demonstratives, it must be mistaken. The answer to the first one brings clarity to our understanding of public demonstratives that refer to a special kind of subjective experience. The answer to the second is an open invitation to continue the research tradition started by Loar two decades ago. This paper can be seen as motivating this line of thought, by refuting the attacks that claimed to show that the PCS was lethally wounded. We evaluated the recent worries and were able to disarm them. The clarification and formulation of a unified demonstrative version of PC is not to be seen as a flaw but a sign that the strategy has a promising future in being the preferred line of response to the famous anti-physicalist arguments. The PCS, therefore, remains the best alternative to disarm the KA.

**Max Mergenthaler Canseco**
Freie Universität Berlin
University of California San Diego
max.mergenthaler@fu-berlin.de

## References

Alter, T. and S. Walter (Eds.) (2007). *Phenomenal concepts and phenomenal knowledge: new essays on consciousness and physicalism*. Philosophy of mind series. Oxford University Press.

Ball, D. (2009). There are no phenomenal concepts. *Mind 118*(472), 935–962.

Balog, K. (1999). Conceivability, possibility, and the mind-body problem. *The Philosophical Review 108*(4), pp. 497–528.

Balog, K. (forthcoming). In defense of the phenomenal concept strategy. *Philosophy and Phenomenological Research*.

Braun, D. (2010). Indexicals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2010 ed.).

Burge, T. (1979). Individualism and the mental. *Midwest studies in philosophy 4*(1), 73–121.

Byrne, D. (2011). Phenomenal senses. *Draft*.

Chalmers, D. J. (1997, November). *The Conscious Mind: In Search of a Fundamental Theory* (1 ed.). Oxford University Press, USA.

Chalmers, D. J. (2007). Phenomenal concepts and the explanatory gap. See Alter and Walter (2007).

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik 100*(1), 25–50.

Hill, C. S. (2009, December). *Consciousness*. Cambridge University Press.

Jackson, F. (1982). Epiphenomenal qualia. *The Philosophical Quarterly 32*(127), pp. 127–136.

Jackson, F. (1986). What mary didn't know. *The Journal of Philosophy 83*(5), 291–295.

Kaplan, D. (1979, January). On the logic of demonstratives. *Journal of Philosophical Logic 8*(1), 81–98.

Kripke, S. A. (1981). *Naming and necessity*. Wiley-Blackwell.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly 64*(4), 354–361.

Loar, B. (2003). Qualia, properties, modality. *Philosophical Issues 13*(1), 113–129.

Loar, B. (2007). Phenomenal states (second version). See Alter and Walter (2007).

Mulhall, S. (2005). *Routledge philosophy guidebook to Heidegger and Being and time*. Taylor & Francis US.

Nida-Rümelin, M. (1998, March). On belief about experiences. an epistemological distinction applied to the knowledge argument against physicalism. *Philosophy and Phenomenological Research 58*(1), 51–73.

Nida-Rümelin, M. (2002). Qualia: The knowledge argument. http://www.illc.uva.nl/~seop/entries/qualia-knowledge/.

Nida-Rümelin, M. (2010). Qualia: The knowledge argument. http://www.illc.uva.nl/~seop/entries/qualia-knowledge/.

Putman, H. (1975). The meaning of “meaning”. *Language, Mind and Knowledge, Burns & Maceachern Ltd, Canada*.

Stoljar, D. (2005). Physicalism and phenomenal concepts. *Mind & Language* 20, 469–494.

Sundström, P. (2008, July). Is the mystery an illusion? Papineau on the problem of consciousness. *Synthese 163*(2), 133–143.

Tye, M. (2009). *Consciousness revisited: materialism without phenomenal concepts*. Representation and mind. MIT Press.

Wittgenstein, L. (1973). *Philosophical Investigations* (3 ed.). Prentice Hall.

# Ein Dilemma für modale Argumente gegen den Materialismus

Sebastian J. Müller

In den vergangenen 40 Jahren haben Philosophen wie Saul Kripke (1980), George Bealer (1994) und David Chalmers (1996; 2010) versucht, auf Basis von Einsicht darein, was metaphysisch möglich ist, zu zeigen, dass der Materialismus falsch ist. Die Debatte um diese Argumente ist ausufernd, aber dennoch hat sich kaum ein Materialist von einem solchen Argument überzeugen lassen. Ich werde argumentieren, dass es gute Gründe hierfür gibt, da modale Argumente nur dadurch überzeugend wirken, dass sie zwei Konzeptionen von metaphysischer Modalität miteinander vermengen, von denen die eine deren ontologische Relevanz, die andere ihre epistemische Zugänglichkeit sichert. Wenn man die Argumente jedoch genauer betrachtet, so lässt sich feststellen, dass nicht beides gleichzeitig zu haben ist, so dass Vertreter modaler Argumente vor einem Dilemma stehen. Entweder sie konzipieren metaphysische Modalität so, dass diese ausreichende ontologische Relevanz hat - dann können wir nicht wissen, was metaphysisch möglich ist. Oder sie sichern den epistemischen Zugang, wodurch wir jedoch nicht mehr von modalen Prämissen auf die Falschheit des Materialismus schließen dürfen. Im Folgenden werde ich dies anhand der modalen Argumente von Kripke und Chalmers zeigen, die Paradebeispiele für die beiden Hörner des Dilemmas darstellen.

## 1. Modale Argumente gegen den Materialismus

Kripke (1980), Bealer (1994) und Chalmers (1996; 2010) haben in den vergangenen 40 Jahren modale Argumente gegen den Materialismus vorgebracht. Bevor ich diese Argumente im Einzelnen untersuche, werde ich hier zunächst eine schematische, verallgemeinernde Darstellung modaler Argumente gegen den Materialismus geben.

(1) Ein physisches Duplikat eines tatsächlichen Menschen, dessen phänomenale Zustände verschieden von diesem sind (ein DPZ), ist metaphysisch möglich.

(2) Wenn ein DPZ metaphysisch möglich ist, ist der Materialismus falsch.

(3) Der Materialismus ist falsch.

Die meisten modalen Argumente gegen den Materialismus haben in etwa diese Gestalt. Sie enthalten (i) eine Prämisse die besagt, dass etwas Bestimmtes metaphysisch möglich ist, (ii) eine Prämisse, die den Bereich des Modalen mit dem des Tatsächlichen verbindet, und (iii) eine Konklusion die Tatsachen betrifft, die wir nicht bereits wissen mussten, um eine der Prämissen zu begründen und die sich auch nicht aus logisch-begrifflichem Wissen ableiten lässt. Untersuchen wir nun Kripkes und Chalmers Argumente darauf hin, ob sie diesem Schema entsprechen.

## 2. Zwei Argumente, zwei Schwachstellen. Kripke und Chalmers gegen den Materialismus

### *2.1 Kripke. Essentialismus und das Problem der Einsicht in Modalität*

Kripkes Argument lässt sich wie folgt rekonstruieren (Vgl. Kripke 1980, 134ff.).

(1) Es erscheint metaphysisch möglich, dass Schmerz ≠ Das Feuern von C-Fasern (Kripke 1980, 146).

(2) Wenn es metaphysisch möglich erscheint, dass Schmerz ≠ Das Feuern von C-Fasern., und sich dieser Anschein nicht durch eine Fehlertheorie wegerklären lässt, dann ist es metaphysisch möglich, dass Schmerz ≠ Das Feuern von C-Fasern. (Vgl. Kripke 1980, 143)

(3) Es ist nicht durch eine Fehlertheorie wegerklärbar, dass es metaphysisch möglich erscheint, dass Schmerz ≠ Das Feuern von C-Fasern (Kripke 1980, 151).

(4) Also (aus 1,2,3): Es ist metaphysisch möglich, dass Schmerz ≠ Das Feuern von C-Fasern (Kripke 1980, 152).

(5) Wenn es metaphysisch möglich ist, dass Schmerz ≠ Das Feuern von C-Fasern, ist der Materialismus falsch (Vgl. Kripke 1980, 149).

(6) Also: Der Materialismus ist falsch.

5 und 6 sind Vereinfachungen, da sich Kripkes[1] Argument in seiner ursprünglichen Form nur gegen die Typ-Identitätstheorie wendet. Jedoch ist dies nicht der entscheidende Punkt und das Argument lässt sich problemlos erweitern, so dass es den Materialismus im Allgemeinen trifft. Entscheidend hier ist Prämisse 2*, die dringend einer Rechtfertigung bedarf.

Weshalb dürfen wir vom Anschein metaphysischer Möglichkeit von Schmerz ≠ Das Feuern von C-Fasern auf die metaphysische Möglichkeit von Schmerz ≠ Das Feuern von C-Fasern schließen, sofern keine Wegerklärung möglich ist? Kripke scheint wie folgt zu argumentieren: Alle Fälle, in denen es möglich scheint, dass p, ohne dass p möglich ist, sind derart, dass wir die Möglichkeit von p mit der Möglichkeit einer zu p qualitativ identischen Situation verwechseln (Kripke 1980, 142f).

In den meisten so gelagerten Fällen können wir nur mithilfe empirischen Wissens erkennen, dass wir diesen Fehler begangen haben, so dass wir meist empirisches Wissen brauchen, um sicherzustellen, dass etwas möglich ist. Im Falle von Schmerzen und dem Feuern von C-Fasern ist dies, laut Kripke, jedoch anders. Da jede Situation, die mit einer Schmerz-Situation qualitativ identisch ist, eine Schmerz-Situation ist, kann hier keine Fehlertheorie greifen, und wir können a priori einsehen, dass Prämisse 4* wahr ist. Der Materialismus ist also falsch.

Man kann dieses Argument an einigen Stellen kritisieren - z.B. indem man behauptet, auch hier ließe sich eine Fehlertheorie entwerfen (Vgl. Hill 1997).

Mein Kritikpunkt an Kripkes modalem Argument gegen den Materialismus ist grundlegender. Kripke akzeptiert klarerweise, dass metaphysische Modalität ihre Wurzeln in wesentlichen Eigenschaften von Dingen und/oder möglichen Welten hat (Kripke 1980, 39ff.). Sie soll grundlegend verschieden von begrifflicher Modalität sein. Dann ist jedoch Prämisse 2 ohne Rechtfertigung. Um vom Anschein von Möglichkeit auf Möglichkeit zu schließen, kann es unter Annahme eines solchen realistischen Bildes nicht genügen, zu zeigen, dass psychologische Anscheine, empirische Tatsachen etc. nicht verbieten, von 1 auf 4 zu schließen. Stattdessen bräuchte Kripke, damit dieser Schluss legitim wäre, eine Begründung dafür, weshalb ein Anschein von Möglichkeit überhaupt in zuverlässiger Weise mit dem Bestehen von Möglichkeit verbunden ist. Hierzu wäre jedoch ein Vermögen nötig, von dem sich zeigen lässt, dass es sensitiv gegenüber dem Vorliegen wesentlicher Eigenschaften von

[1] Ich lege hier die Soames-Lesart von Kripke zugrunde. Es gibt auch Stellen, an denen Kripke sich so äußert, dass er Zweidimensionalisten wie Chalmers näher steht, dann würde ihn die Kritik treffen, die ich gegen Chalmers richte (Vgl. Soames 2006). Tahko radikalisiert die Unterscheidung zwischen den Philosophen, die metaphysische Modalität linguistisch erklären und denen, die sie essenzialistisch erklären, in klarsichtiger Weise noch weiter als Soames, in dessen Ontologie von metaphysischer Modalität er ebenfalls anti-essenzialistische Züge ausmacht (Vgl. Tahko forthc.).

Dingen ist. Kripke hat jedoch nicht gezeigt, dass ein Anschein von Möglichkeit nur dann vorliegt, wenn Möglichkeit vorliegt - oder dass zumindest eine zwar fallible, aber starke Verbindung zwischen beiden besteht. Er zeigt nur, dass, wenn wir dies bereits annehmen, eine bestimmte Art von Wegerklärung des Anscheins von Möglichkeit im Falle phänomenaler Zustände nicht gelingen kann. Damit ist ein leichtes Problem modalen Wissens für einen Spezialfall eventuell beantwortet - das grundlegende Problem, ob wir überhaupt verlässlich darin sind, wahre Meinungen über metaphysische Modalität zu bilden, bleibt völlig ungelöst. Daher ist Kripkes Argument nicht erfolgreich: Die Prämisse, die den Anschein von Möglichkeit mit dem Vorliegen von Möglichkeit verbinden soll, ist nicht annähernd stark genug, um dies wirklich zu leisten.[2] Dabei geht es nicht nur darum, dass Kripke 2 nicht ausreichend begründet hat. Natürlich kann nicht jede Prämisse immer weiter begründet werden, und ein solcher Einwand wäre nichts weiter als eine Anwendung von Agrippas Trilemma. In diesem Fall geht jedoch mehr vor. Prämisse 2 ist dringend begründungsbedürftig. Möglichkeitsanscheine sind prima facie nicht wie Wahrnehmungsanscheine, die wir akzeptieren dürfen, solange sich kein Anzeichen für eine Täuschung auftut. Sie sind nicht durch unsere alltägliche Praxis oder die der Wissenschaften gestützt. Die meisten Menschen, einschließlich der meisten Wissenschaftler und Philosophen, haben keine Vormeinungen zugunsten von 2.[3] Wer sich auf diese Prämisse stützen will und Anscheine oder Vorstellbarkeit als Zugang zu metaphysischer Modalität nutzen möchte, der muss zeigen, wie genau diese Anscheine und die unabhängig davon bestehenden metaphysischen Möglichkeiten und Notwendigkeiten zusammenhängen. Solange dies nicht geleistet ist, sollten wir 2* nicht akzeptieren

### *2.2 Chalmers. Zweidimensionalismus und das Problem der ontologischen Relevanz des Modalen*

Chalmers Argument findet sich, im Gegensatz zu Kripkes, vollständig und in eindeutiger Form in seinen Texten.

i. P & ¬Q is conceivable.

ii. If P & ¬Q is conceivable, then P & ¬Q is 1-possible.

iii. If P & ¬Q is 1-possible, then P & ¬Q is 2-possible or Russellian monism is true.

iv. If P & ¬Q is 2-possible, materialism is false.

v. Materialism is false or Russellian monism is true. (Chalmers 2010, 152.)

Vieles an diesem Argument ist erklärungsbedürftig. Ich werde hier aufgrund des knappen Raumes einige Vereinfachungen vornehmen, von denen ich glaube, dass sie nicht sinnenstellend sind. „P“ ist die Gesamtheit der Wahrheiten über das Mikrophysische, „Q“ derer über das Phänomenale. „Conceivable“ bedeutet letztlich nicht viel mehr als „frei von logisch-begrifflichen Widersprüchen“ - zumindest genügt diese Bestimmung hier. 1-possibility ist weitgehend mit epistemischer Möglichkeit identisch, wobei hiermit weite epistemische Möglichkeit gemeint ist - etwas ist epistemisch genau dann möglich, wenn es nicht a priori ausgeschlossen werden kann. Es ist also genau dann 1-möglich, dass p, wenn es

[2] Analog lässt sich argumentieren, wenn Vorstellbarkeit oder eine Möglichkeits-Intuition die Grundlage des Arguments bilden. Ich kann dies hier aus Platzgründen nicht tun. Vgl. Roca-Royes 2011 zu einem verwandten Punkt.

[3] Auch innerhalb der analytischen Philosophie gibt es keinen Konsens zugunsten von 2*, sondern eine Vielzahl von Ansätzen zum Erwerb modalen Wissens, wobei die Anzahl von Theorien, die auf Anscheine oder Vorstellbarkeit setzt, sinkt.

nicht a priori erkennbar ist, dass ¬p. 2-possibility ist hier am interessantesten für mich. Es ist 2-möglich, dass p, gdw. es eine logisch mögliche Welt, betrachtet als kontrafaktische, gibt, in der p der Fall ist. Wenn es 1-möglich ist, dass p, jedoch nicht 2-möglich, dann muss es eine empirische Tatsache sein, die p 2-unmöglich macht. Damit ist 2-Möglichkeit vollständig von epistemischer Möglichkeit - die sich aus begrifflichen Zusammenhängen ergibt - und empirischen Tatsachen abhängig. Dies ermöglicht es Chalmers, Kripkes Problem zu umgehen. Nur begriffliche Zusammenhänge und empirische Tatsachen sind relevant dafür, ob etwas 2-möglich ist. Zusammen mit Kripkes Fehlertheorie, die ausschließt, dass empirische Tatsachen für phänomenale Zustände den Schluss von 1-Möglichkeit auf 2-Möglichkeit verhindern, ergibt sich ein Argument, dass bis Prämisse 3 schlüssig ist. Zwar kann man auch hier wieder vielfältige Kritik üben (Vgl. Roca-Royes 2011), doch ich glaube, Chalmers übersteht diese - wofür ich hier leider nicht argumentieren kann.

Das zentrale Problem für Chalmers Argument ist der Übergang von 2-Möglichkeit auf die Falschheit des Materialismus. Im Falle von klassischer metaphysischer Möglichkeit ist dies kein Problem. Wenn es wirklich so ist, dass ein Gegenstand ohne einen anderen existieren könnte, sind die beiden klarerweise nicht identisch. Doch auf diese Art von Möglichkeit darf Chalmers sich aufgrund seiner Konstruktion von 2-Möglichkeit aus begrifflicher Möglichkeit und empirischen Tatsachen nicht berufen. Stattdessen ist es lediglich so, dass begriffliche Zusammenhänge und empirische Tatsachen zusammen nicht ausschließen, dass P und Q identisch sind. Auf die Negation dieser These müssen sich Materialisten jedoch nicht festlegen. Diese müssen nur behaupten, dass P und Q de facto identisch sind, und nicht, dass sich dies aus begrifflichen Zusammenhängen und empirischen Tatsachen logisch herleiten lässt. Daher muss Chalmers behaupten, dass begriffliche oder epistemische Möglichkeit ausreichend ontologische Schlagkraft besitzt, um seinen Schluss zu rechtfertigen. Chalmers sieht dieses Problem und formuliert es explizit:

> One natural worry is this: if this modality is grounded in the rational domain, then how can it drive ontological conclusions? Why does the mere logical possibility of a zombie world entail the falsity of materialism, for example? (Chalmers 2010: 191).

Obwohl diese Formulierung eindeutig zeigt, dass Chalmers das Problem sieht, widmet er ihm nur sehr wenig Aufmerksamkeit. Hierzu schreibt er nur

> In response, it is obvious that modal notions from the rational domain have a bearing on ontology. For example, a priori entailment from unmarried men to bachelors gives us reason to accept that bachelors are not an ontological extra. (Chalmers 2010: 191).

Ich werde nicht in Frage stellen, dass die These, dass begriffliche und epistemische Modalität gewisse Implikationen für den Bereich der Tatsachenontologie haben, richtig ist. Diese genügt jedoch nicht, um seinen Schluss in Prämisse iv zu rechtfertigen. Vergleichen wir die Fälle. Wir wissen a priori, dass jeder, der ein Junggeselle ist, ein unverheirateter Mann ist. Dies ist begrifflich notwendig. Also dürfen wir schließen, dass es nicht neben den unverheirateten Männern noch die Junggesellen gibt. Gibt uns dies Wissen über einzelne Gegenstände? Nein. Wenn wir einen beliebigen Gegenstand betrachten und uns jemand fragt, ob er Junggeselle ist, können wir durch das Wissen über die begriffliche Notwendigkeit von „Junggesellen sind unverheiratete Männer“ keine Antwort geben. Was wir daraus ableiten können, ist nur das Wissen, dass wenn jemand ein Junggeselle ist, er ein unverheirateter Mann ist. Wir erhalten aus Wissen über begriffliche Modalität also nur konditionales Wissen (Vgl. Nimtz 2011). Im Falle von Chalmers antimaterialistischem Argument hingegen ist das Ziel nicht nur solches konditionale Wissen. Hier wollen wir über einen Gegenstand A wissen, ob er ein physisch-funktionaler Zustand ist. Chalmers modales Argument gegen den Materialismus soll uns also direkt ontologisches Wissen liefern, wobei es nicht analytisch ist, dass dieser Gegenstand nicht physisch-funktional ist. Es geht also um eine synthetische

Erkenntnis, und Einsicht in begriffliche Modalität kann diese prima facie nicht liefern. Chalmers bräuchte ein wirkliches Argument, um zu zeigen, dass seine 2-Modalität leisten, was metaphysische Modalität leistet, und dieses fehlt ihm. Er schreibt weiter:

> Furthermore, materialism is itself a modal thesis, or at least a modally constrained thesis, so the analysis of modality quite reasonably drives conclusions about materialism. (Chalmers 2010: 191).

Doch diese These ist keineswegs unproblematisch. Richtig ist, dass der Materialismus als metaphysisch modale These verstanden werden kann bzw. beinahe muss. Nur wenige Philosophen weichen hiervon ab. Doch die Frage ist, ob die Modalität, die für Materialisten relevant ist, überhaupt die Art von Modalität ist, die Chalmers „2-Modalität" nennt. Während metaphysische Modalität realistisch verstanden werden muss und sich zumindest den meisten Interpretationen zufolge aus wesentlichen Eigenschaften ergibt, ist 2-Modalität komplett durch logisch-begriffliche Modalität und empirische Tatsachen konstituiert. In einer Diskussion darüber, ob diese 2-Modalität dieselbe ontologische Schlagkraft wie metaphysische Modalität hat, ist nun die Berufung auf die modale Natur des Materialismus nicht angemessen – da es gerade darum geht, welche Arten von Modalität für die Wahrheit des Materialismus relevant sind. Typische A posteriori-Materialisten werden sich auf den Standpunkt stellen, logisch-begriffliche Möglichkeiten seien völlig irrelevant dafür, ob ihre Form des Materialismus wahr ist. Wenn Chalmers 2-Modalität in so starker Abweichung von klassischer metaphysischer Modalität konstruiert, wie er es in seinem antimaterialistischen Argument tut, ändert sich die Beweislast entschieden. Wenn Prämisse iv überzeugen soll, muss Chalmers zeigen, dass logisch-begriffliche Modalität über den Bereich analytischer Wahrheiten hinaus informativ sein kann. Das tut er jedoch an keiner Stelle. Prima facie scheint dies jedoch äußerst unplausibel – was dadurch verborgen bleibt, dass Chalmers an vielen Stellen einen Spagat zwischen Realismus und Antirealismus versucht.

## 3. Das Dilemma für modale Argumente gegen den Materialismus

Wenn wir das bisher Gesagte auf das Schema anwenden,

(1) Ein physisches Duplikat eines tatsächlichen Menschen, dessen phänomenale Zustände verschieden von diesem sind (ein DPZ), ist metaphysisch möglich.

(2) Wenn ein DPZ metaphysisch möglich ist, ist der Materialismus falsch.

(3) Der Materialismus ist falsch.

ergibt sich folgendes:

Wenn wir Kripkes Bild zugrunde legen, wird 2 unproblematisch. Hier geht es dann wirklich um die modalen Eigenschaften von Dingen, so dass der Schluss auf Tatsachen erlaubt ist. Jedoch wird es dann fraglich, wie wir 1 rechtfertigen sollen. Hierzu müssten wir sicherstellen, dass nicht die wesentlichen Eigenschaften von Menschen, physischen und phänomenalen Zuständen dafür sorgen, dass ein DPZ unmöglich wird. Hierfür brauchen wir ein Vermögen, dass für wesentliche Eigenschaften sensitiv ist - welches Kripke nicht bietet.

Chalmers umgeht dieses Problem, indem er metaphysische Möglichkeit durch 2-Möglichkeit ersetzt - Prämisse 1 wird dadurch erkennbar wahr, 2 hingegen scheint falsch zu werden.

Beide Argumente führen nicht zu einer erfolgreichen Widerlegung des Materialismus.

Dies trifft nicht nur auf modale Argumente gegen den Materialismus, sondern auf alle modalen Argumente, die in die folgende Form passen, überhaupt, zu:

(1) Es ist metaphysisch möglich, dass p.

(2) Wenn es metaphysisch möglich ist, dass p, dann q.

(3) Also: q.

Immer stehen wir vor dem Problem, dass entweder 1 unwissbar oder 2 falsch wird - je nachdem, ob wir Modalität essenzialistisch betrachten oder sie an begriffliche oder epistemische Modalität und empirische Tatsachen binden.

## 4. Lösungsansätze

Dieses Dilemma könnte auf zwei Weisen gelöst werden. Erstens könnte man zeigen, wie wir Wissen über wesentliche Eigenschaften (oder hier, über ihr Nicht-Vorliegen) gewinnen können (Vgl. Williamson 2007; Peacocke 1998;). Zweitens könnte man zeigen, dass Wissen über epistemische oder begriffliche Modalität anspruchsvolle ontologische Schlüsse erlaubt. Ich kann diese Ansätze hier nicht erschöpfend diskutieren, sondern nur darauf hinweisen, dass beide voller Probleme stecken und dass es Aufgabe der Verfechter antimaterialistischer Argumente ist, zu zeigen, wie dies funktionieren soll, was noch dadurch erschwert wird, dass beispielsweise empiristische Ansätze in der Erkenntnistheorie metaphysischer Modalität kaum geeignet scheinen, um Argumente gegen den Materialismus zu untermauern. Die Herausforderung, zu zeigen, wie wir a priori Tatsachenwissen erlangen können, bleibt ebenso gewaltig wie ungelöst. Für den Fall der modalen Argumente gegen den Materialismus hat sich hier gezeigt, dass die bisherigen Strategien zur Bewältigung dieser Herausforderung gescheitert sind.

**Sebastian J. Müller**
Universität Bielefeld
sebastian2607@googlemail.com

## Literatur

Bealer, G. 1994: „Mental Properties“, *Journal of Philosophy* 91, 185-208.

Chalmers, D. 1996: *The Conscious Mind. In Search of a Fundamental Theory*. Oxford: Oxford University Press.

— 2010: „The two-dimensional argument against materialism“, in D. Chalmers 2010a, 141-191.

— 2010a: *The Character of Consciousness*. Oxford: Oxford University Press.

Hill, C. 1997: „Imaginability, Conceivability, Possibility and the Mind-Body Problem“, *Philosophical Studies* 87, 61–85.

Kripke, S.A. 1980: *Naming and Necessity*. Oxford: Blackwell.

Nimtz, C. 2011: „A Priori Wissen als Philosophisches Problem“, *Jahrbuch für Philosophie* 3, 1154-1174.

Peacocke, C. 1998: *Being Known*. Oxford: Oxford University Press.

Roca-Royes, S. 2011: „Conceivability and De Re Modal Knowledge“, *Noûs* 45, 22-49.

— forthc.: „Modal Knowledge and Counterfactual Knowledge“, *Logique et Analyse.*

Soames, S. 2006: „The Philosophical Significance of the Kripkean Necessary A Posteriori“, *Philosophical Issues* 16, 288-309.

Tahko, T. forthc.: „Soames's Deflationism about Modality“, *Erkenntnis*.

Williamson, T. 2007: *The Philosophy of Philosophy*, Malden, Mass.: Wiley-Blackwell.

# How We Know Our Senses

Eva Schmidt

I propose a new criterion by which, I hold, subjects recognize and distinguish their sensory modalities. I argue that, rather than appealing to one of the standard criteria (sense organ, proximal stimulus, phenomenal character, or representational content (Grice 1962, Macpherson 2011a)) or to O'Dea's (2011) proprioceptive content, we need to introduce the criterion of *location in the functional architecture of the subject's personal-level* mind in order to make sense of an ordinary subject's ability to tell immediately which sensory modalities are employed in her occurrent perceptual experience. More specifically, a subject's personal-level mind is functionally organized into different faculties, and, seeing as it is *her* mind, she has a natural cognitive access to this structure; in the specific case of perceptual experience, perceptual input from the world is present to the subject as organized into the different sensory modalities, vision, hearing, touch, taste, and smell. I motivate and explicate my new criterion for distinguishing the senses, in particular its psychological aspects. Moreover, I show how it can handle problems raised by empirical findings, such as additional human senses (e.g. the vomeronasal sense) and cross-modal experiences (e.g. the experience of a speaker's voice emanating from his mouth).

## 1. How to Distinguish the Senses: The Traditional Criteria

In their investigations, philosophers of perception tend to focus on visual perceptual experience. By comparison, the non-visual sensory modalities (hearing, touch, taste, and smell) are usually, albeit undeservedly, neglected. One important question concerning the senses that has recently attracted more attention is the question of how to individuate the senses. Locus classicus of the corresponding debate is Grice's (1962) paper 'Some Remarks about the Senses', where he discusses the following four criteria by which our sensory modalities may be distinguished.[1]

(A) The sense organ that is involved in a particular sensory modality.

(B) The proximal stimuli relevant to the sensory modality.

(C) The specific phenomenal character associated with the sense.

(D) The properties typically represented by the sense.

For hearing, for instance, the criteria in question would be (A) the ear, plus the nerves and brain areas involved in processing input coming from the ear, (B) sound waves, (C) a particular auditory phenomenal character, and (D) sounds and their properties.

In what follows, I will focus on one specific version of the individuation question and argue that it cannot be answered satisfactorily by appeal to any of these criteria. The question I want to focus on is: How can (and do) normal perceivers individuate their sensory modalities in the act of perceiving? For instance, how can Gertrude know that she sees the roundness of the coin, rather than feeling it? This is an epistemological question (concerned with our self-knowledge of our sensory modalities). Note that this kind of knowledge seems so natural that it is (at least initially, see below) difficult to think of a scenario in which a subject might be wrong or unjustified in her beliefs about the sensory modality of her occurrent perceptual

[1] Another very helpful discussion of these criteria can be found in Macpherson (2011).

experiences. In this respect, it is similar to a subject's knowledge of the phenomenal character of her experiences – it is hard to come up with an example in which a subject is mistaken in her beliefs about the phenomenal character of one of her experiences.[2]

A very closely related psychological question will also become relevant in what follows. This is the question of how ordinary perceivers are able to form their immediate introspective judgments about the sensory modalities of their occurrent perceptual experiences. It is a question concerning the mechanism by which perceivers can form such introspective statements, which is neutral on the epistemological issue of whether these judgments constitute knowledge.

I will leave to one side a further question in the vicinity, viz. the question of the criteria by which scientists should best distinguish the different senses. I will not make claims about how scientists might best taxonomize sensory modalities. This kind of project is quite different from the one I will pursue here. It is not primarily concerned with a perceiver's self-knowledge or her immediate introspective judgments concerning her occurrent perceptual experiences. It plausibly has to take into account additional human senses (such as equilibrioception) and senses of other animals (such as bat echolocation).[3]

Here is why the aforementioned criteria cannot provide an answer to my question. (A) cannot work because of counterexamples such as the following: When cold water enters my ear, I perceive that the water is cold. The simple sense organ criterion would wrongly classify this as an auditory experience of hearing the cold. Ordinary perceivers do not make this mistake – they know that they are detecting something cold with their sense of touch. On the other hand, it is futile to appeal to the nerves and brain areas relevant to processing information coming in through the ear, for the self-knowledge of normal subjects about their sensory modalities is clearly independent of their knowledge of nerves and brains.

The same is true for criterion (B). Gertrude will be able to know that she feels rather than sees the roundness of the coin even if she is ignorant of the fact that, say, light waves are the proximal stimuli relevant to her sense of vision.

The phenomenal criterion (C), on the other hand, is quite promising as far as its accessibility to ordinary perceivers is concerned. One might suggest that Gertrude knows that her experience of the roundness of the coin is a tactile experience because of its tactile phenomenal character, which she simply could not confuse with any other phenomenal character. However, the problem with this option is that this supposed modality-specific phenomenal character is hard, if not impossible, to pin down.

The following objection is due to Grice (1962). If we try to explicate the phenomenal difference between Gertrude's visual experience and her tactile experience of the roundness of the coin, we end up describing a difference in the external objects and their properties that Gertrude's experiences present her with. For instance, we might end up saying that Gertrude sees the colour and shininess of the coin, but she feels its warmth and heaviness. This is a kind of transparency argument: All that seems to be left of the difference in phenomenal character between a visual and a tactile experience of the coin is a representational difference, which takes us to criterion (D), the representational criterion.

It seems plausible enough that, between two different sensory modalities, there is always a difference with respect to the total properties represented by them. In the coin example, colour is represented by sight, but not by touch; on the other hand, warmth is represented by touch, but not by sight. Yet, focussing on my question concerning a normal perceiver's knowledge of her sensory modalities, it seems wrong to say that she has to go through the list

[2] I will address potential examples of mis-individuation of perceptual experiences in the objections section.

[3] The point of the traditional four criteria is (mainly) to provide an account of what makes a certain sensory modality the sensory modality it is – they fit best with a *metaphysical* project of individuating the senses.

of properties represented by a specific perceptual experience before she can determine which kind of sensory experience she is undergoing. Quite the contrary, a subject's awareness of which sense she is employing in perceiving certain external properties seems to be more basic than her awareness of all the different kinds of properties she is perceiving in that sensory modality.

To put it somewhat differently, according to the representational criterion, it looks as though a perceiver has to infer which sensory modality she employs in undergoing a perceptual experience from her knowledge of what properties out there she perceives. Such an account cannot do justice to the immediacy and non-inferentiality of a perceiver's knowledge of the sensory modalities of her occurrent perceptual experiences.

These four criteria, then, are unsatisfactory. Another, more promising criterion has been proposed by O'Dea (2011). Perceivers know the sensory modalities of their occurrent perceptual experiences because each sense represents itself, as a body part to be used to explore the environment, along with features of the environment. O'Dea claims that subjects know which sense they are employing because perceiving is partly proprioceptive, i.e., a perceptual experience represents not merely external objects and properties, but also which sense organ is used in undergoing the experience. In particular, each sense represents itself as a tool to explore the environment.

One way to put his proprioceptive criterion is to say that perceivers know which sensory modality they are employing via

(E) The sense organ proprioceptively represented by the perceptual experience.

This view faces at least three problems. The first is that it is simply implausible that a perceptual experience represents not only things out in the subject's environment, but also, on a par with this, that a certain sense organ is being used. When Gertrude sees *the coin*, she does not also see that *she is seeing the coin*.

The second problem is a consequence of the assumption that the sensory modality is represented on a par with features of the environment: Perceptual misrepresentation of environmental features is widespread, so we should expect that there is also widespread misrepresentation of the sensory modality employed. But not so – as I mentioned before, it is hard to come up with examples of subjects' mistaken judgments of the sensory modalities of their occurrent perceptual experiences. This is to say that, on O'Dea's proprioceptive proposal, this kind of misrepresentation is a very rare occurrence.

The third problem is that perceivers have immediate knowledge of which sensory modalities they are employing. They do not have to figure out which sense they are using in a perceptual experience by first sorting through their perceptual contents to find out which sensory modality is represented proprioceptively. For instance, Gertrude knows that she sees the coin (rather than feeling it) immediately, without first having to sift through the content of her visual experience, all the while trying to figure out whether there is a proprioceptive representation of her eye (rather than her skin) included in it.

## 2. My Proposal

First off, let me provide a brief diagnosis of why individuation criteria (A) – (E), that have been proposed to provide an account of how ordinary perceivers can individuate their sensory modalities in the act of perceiving, have failed. Their problem is that they cannot respect the following features of our judgments about our sensory modalities:

(1) These judgments are immediate.

(2) These judgments are non-inferential.

(3) These judgments are (or at least appear to be, intuitively) incorrigible.

A better individuation criterion should respect these features of our judgments about the sensory modalities of our occurrent perceptual experiences. In this section, I will propose a new individuation criterion and argue that it can do justice to the aforementioned features of our judgments about our sensory modalities.

Here is my proposal. A perceiver can tell that she has a perceptual experience in a certain sensory modality because, in undergoing a certain perceptual experience, it is not just the perceptual content that is immediately available to her, but also the perceptual state (including its sensory modality) whose content it is. We might say that perceptual experience is a package deal: The personal-level availability of a certain perceptual content goes along with the availability of the perceptual state (be it a visual, a tactile, an auditory, a gustatory, or an olfactory state) that it is a content of.

Her perceptual states, including their sensory modality, are accessible to the perceiver as part of her overall access to the functional organization of her personal-level mind. To the subject, mental content is present as structured into different faculties, e.g. the perceptual faculties (or 'channels'). Functionally speaking, the senses are the mental faculties that convey different kinds of information about the subject's immediate environment with the purpose of action guidance and belief formation. Cognitively to access a certain perceptual content via a particular sensory channel is cognitively to access the functionally individuated sensory channel through which the perceptual content enters the personal-level mind. So, on my proposal, the subject knows the sensory modalities of her occurrent perceptual experiences thanks to

(F) The availability, to her, of the perceptual content's '*perceptual channel*' *in the* personal-level *functional architecture of her mind*.

To put this in terms of the example of feeling vs. seeing the roundness of a coin, Gertrude knows that she sees the roundness of the coin (rather than feeling it) because this visual content is available to her as the content of a visual experience, not of a tactile experience. The visual experience is present to her in terms of her visual faculty's particular functional significance in the overall functional architecture of her personal-level mind.

Note that my claim is not that this access to a perceptual state in terms of its functional significance translates to the perceiver's awareness of a certain kind of phenomenal character attached to the perceptual experience that is characteristic of the sensory modality in question. My claim is that subjects have an immediate access to how their minds are organized (at the personal level, of course), and to the channels through which mental content enters their personal-level minds. I hold that this is a kind of access that is not even mediated by a phenomenal character.

To return to the example, Gertrude knows that she is seeing rather than feeling the roundness of the coin not because of elements of the content of her visual experience, nor because of its specific phenomenal character, but because she has direct access to the functional organization of her personal-level mind. She has access to the fact that the information about the coin is coming in through the visual channel rather than the tactile channel.

The more general question in the vicinity is how a subject can ever tell that she believes something rather than desiring it, that she imagines something rather than remembering it, or that she perceives something rather than entertaining it as a thought. My suggestion is that, as a matter of functional structure, the personal-level mind is organized into different faculties, such as memory, belief, or perception. Perception, in its turn, is organized into the different perceptual channels of sight, hearing, touch, smell, and taste.

This structure echoes the functional structure of the sub-personal mind, where, for instance, it is organized into the respective perceptual systems. The subject has direct cognitive access

to the personal-level functional structure – this should come as no surprise, since it is her own mind. This, I want to suggest, is how she can tell that she is not remembering that the coin is round, nor wishing that the coin were round, nor imagining that it is round, but that she is perceiving, via her sense of vision, that the coin is round. To do so, she does not have to take any detours, for instance through a phenomenal character or through a content representing the sensory modality. She has simply cognitively to access the functional organization, and in particular the sensory channel through which a certain perceptual content enters her personal-level mind.

Let me add one clarificatory note. One may wonder whether the subject, in order to grasp the sensory modality of a particular perceptual experience, has to possess concepts pertaining to the functional role of states in that sensory modality. The idea would be that in order to gain knowledge of the sense modality of a certain perceptual experience, the subject has to exercise her concepts concerning the functional significance of the sense in question (e.g. that it is a sensory modality that deals with distal stimuli at a distance, that it takes light waves as its input, that it typically leads to this or that kind of belief, that it involves processing in the visual input module, etc.).

Let me emphasize that this is not the view that I am advocating. It would lead us back to the criteria that I have criticized above. Rather, my view is that the personal-level mind has a certain functional organization, and that this organization presents itself to the subject just as naturally as the mental content that enters the mind through the described perceptual 'channels'. To the ordinary subject, perceptual content simply comes organized into the different sensory modalities.[4]

My proposal is motivated, first, by the fact that none of the other criteria provides a satisfactory account of perceivers' self-knowledge of the sensory modalities of their occurrent perceptual experiences.

Second, my proposal does justice to the features of our judgments concerning our sensory modalities enumerated above. On my view, it is unproblematic to admit that our judgments of the sensory modalities of our occurrent perceptual experiences are immediate, non-inferential, and incorrigible. Our knowledge of our sensory modalities is immediate, for it is part of our direct introspective accessibility of the functional organization of our own minds. It is non-inferential because, on my account, we do not infer to judgments about our sensory modalities from, e.g., the phenomenal character or aspects of the content of our perceptual experiences. It is (possibly, problems for this see below) incorrigible because our direct awareness of the functional structure of our own personal-level minds does not leave a lot of room for error.

As a third point in favour, I might add that this kind of immediate access to the functional organization of one's personal-level mind, and thereby to the mental states that bear certain contents, makes sense from an evolutionary perspective as well. To see this, imagine the perceiver being confused over whether she *remembers* seeing a tiger in front of her, or whether she is *currently seeing* a tiger in front of her.

My account has both psychological and epistemological dimensions. The psychological dimension is that there actually is a functional organization of our personal-level minds as described, to which we have direct cognitive access, and from which our confidence in our knowledge of our sensory modalities is derived. The epistemological dimension concerns the question of whether this access constitutes knowledge to match this confidence. In what follows, I will mostly focus on the psychological issue.

---

[4] But which concepts does the subject employ in judging that she is undergoing an experience in this or that sensory modality? I am not sure as to the best answer to this question, but generally speaking, the options that are open to me here are similar to the ones exploited by defenders of phenomenal concepts in the qualia debate. For instance, our concepts of our own sensory modalities might plausibly be recognitional or demonstrative concepts.

## 3. Objections

I will discuss two problems for my account that result from recent research in cognitive science. The first problem is raised by senses that ordinary perceivers have, but know nothing about, such as the vomeronasal sense or equilibrioception. I claim that we can know our senses via our immediate cognitive access to the functional structure of our minds. How does this fit with scientists finding senses in humans that we know nothing about introspectively?

One clear example of this is the vomeronasal sense. In the human nose, there is a sense organ that detects pheromones and thereby helps control sexual behaviour. We have no immediate cognitive access to this sense – we are not conscious of the fact that our sexual behaviour is influenced by a pheromone detector in the nose, nor can we become aware of this fact by merely introspecting our sensory modalities. It seems, then, that the vomeronasal sense is a counterexample to my claim that we have direct cognitive access to our perceptual modalities.

Reply: I concede that not every channel through which we gather information about the environment registers as a distinct personal-level perceptual channel. In the case of the vomeronasal sense, we often just find ourselves being attracted to certain people, without being able to notice that there is something like a sensory modality involved informing us that someone is sexually attractive.

Take a different case, equilibrioception. Our sense of balance is a sense that we take note of only when we lose our balance, for instance by spinning until we get dizzy. Otherwise, this sensory modality does not seem to register as an independent personal-level sensory channel. Again, I concede that this is a plausible example of a sensory modality that does not make itself known as a distinct sensory channel in the functional architecture of our personal-level minds.

It is not problematic for me to admit this, for I am not in the business of making metaphysical claims about what makes something a particular sensory modality. Rather, I am in the business of explaining how we make our everyday judgments about the sensory modalities of our occurrent perceptual experiences (a psychological question), and of how we can have the corresponding everyday knowledge of our sensory modalities (an epistemological question). For these projects, there is no threat if there are other candidates for sensory modalities that we have no immediate introspective knowledge of. The most I am committed to, seeing as I am trying to argue that we can have immediate knowledge of our senses, is that the senses that we can cognitively access as personal-level sensory channels generally are as they appear. I hold that we are the authorities with respect to those sensory modalities that we can become aware of immediately (as I have proposed).

The second problem is that my proposal may seem to be threatened by the possibility of cross-modal experiences. For instance, scientific findings suggest that we perceptually experience voices as coming from the moving mouths of speakers because of a combined use of hearing and vision. This is why ventriloquists can apparently 'throw' their voices and why it is confusing to watch a movie in which the timing of the movements the speakers' mouths does not match their voices.

These findings apparently constitute counterexamples to my claim that there is a distinct personal-level channel for each of our sensory modalities, a channel about which we can form immediate and incorrigible introspective judgments. It appears that we are regularly mistaken about which sensory channel is relevant to a certain perceptual experience, for we would classify hearing a speaker's voice as an auditory experience, not as a cross-modal experience involving vision.

There are other cases in the same vein. The spicy flavour of chilies is detected, in part, by pain receptors on the tongue. Plausibly, we cannot tell this by introspectively accessing our experience, for tasting chilies appears to be an exclusively gustatory experience. Similarly, the

rich flavour of a tomato sauce appears to be a matter merely of our sense of taste, when in reality, it is partly due to sensors in the nose and thus partly based on our sense of smell.

Reply: Cases in which we do not notice the arguable cross-modality of our sensory experiences lead us back to the question of what individuates a perceiver's sensory modalities, metaphysically speaking. One option is to take the hard line with respect to this question and insist that what makes a perceptual experience the kind of perceptual experience it is is how it strikes me: I am the authority on the sensory modality of my occurrent perceptual experiences. If an experience of a speaker's voice strikes me as an auditory experience, then that's what it is.

My opponent's claim is that my experience of a speaker's voice must be cross-modal since it involves both my ears and my eyes and the physiological and nervous systems processing the input from ears and eyes, and since the proximal stimuli relevant to both vision and hearing are involved in this experience. But this reasoning presupposes that sense organ and proximal stimulus are relevant to the individuation of our senses, a claim I have rejected above. Given this, I can insist that the subject is the ultimate authority on her own sensory modalities, so that her senses must be individuated according to her judgments.

The other option is to take a softer line and to concede that sensory modalities are the sensory modalities they are in virtue of sense organ, proximal stimulus, and relevant physiological and neural structures. Consequently, the subject's judgment that she has an exclusively auditory experience is only apparently incorrigible. Even though it seems unconceivable to the uninformed perceiver that she could be mistaken in her judgments about her sensory modalities, cognitive science shows that subjects are prone to error in such judgments. For they fail to recognize the cross-modality of some of their perceptual experiences.

But even this result would not be problematic for the psychological side of my proposal: For one, we do sometimes notice when an experience is cross-modal. For instance, eating a spicy chili involves not just a gustatory experience of spiciness; it is also a painful experience. For another, the (arguable) fact that our judgments of our sensory modalities are not always reliable is compatible with my psychological claim that we form immediate judgments regarding our sensory modalities based on our introspective access to the functional organization of our personal-level minds. For it is possible that this access is not perfect. Alternatively, it is possible that not everything that goes on in our minds at sub-personal levels is perfectly reflected in the functional make-up of our personal-level minds.

To sum up this section, I have discussed two problems for my proposal raised by, first, sensory modalities that we have no immediate introspective knowledge of, and second, cross-modal perceptual experiences. Both of these phenomena are initially problematic for my view because, according to it, we have immediate and incorrigible access to the sensory modalities of our occurrent perceptual experiences in functional terms.

As to point one, I have conceded that there may be sensory modalities that we have no introspective access to, but argued that this is no problem for my view. With respect to the second problem, I have suggested a hard line one might take: The subject alone is the authority on her sensory modalities, so that supposedly cross-modal experiences turn out to be experiences in only one modality if this is how things strike the perceiver. But I have also described a soft line, according to which our judgments of the sense modalities of our occurrent perceptual experiences are only apparently incorrigible. Really, these judgments are sometimes mistaken, in particular in the case of cross-modal experiences.

I have to admit that I am undecided which of the two lines to take. On the one hand, it is preferable not to devise views in the philosophy of mind that conflict with findings in cognitive science – this consideration clearly favours the soft line. On the other hand, the soft line casts doubt on our self-knowledge of the sensory modalities of our occurrent perceptual experiences. The problem lies not so much in my proposal of how we can individuate our

senses, but rather in a conflict between the intuitive view that we can have immediate, incorrigible knowledge of our senses and recent findings of cognitive science.

## 4. Conclusion

In this paper, I have tried to find an answer to the question of how ordinary subjects can know about the sensory modalities of their own occurrent perceptual experiences. I have argued against four classical individuation criteria, viz. sense organ, proximal stimulus, phenomenal character, and representational content, and against one recent one, viz. proprioceptive representational content. Further, I have identified three intuitive features of our judgments of the sensory modalities of our perceptual experiences: their immediacy, non-inferentiality, and their incorrigibility.

My own proposal, which can respect these features, is that subjects have immediate introspective access to the functional organization of their personal-level minds, including the perceptual channels through which certain aspects of their environments are presented to them.

I have discussed two problems for the claim that we can form immediate and incorrigible judgments concerning our sensory modalities (introspectively inaccessible sensory modalities and cross-modal experiences). I conceded that the latter problem indeed casts doubt on the (intuitively plausible) incorrigibility of our immediate judgments about our senses. This is not a problem for my psychological claim about the mechanism which enables us to form judgments about the sense modalities of our occurrent experiences. But it is a problem for the plausible related epistemological claim that we have privileged access to (and incorrigible self-knowledge of) the functional organization of our personal-level minds, including the perceptual channels through which perceptual content enters our minds.

This could be taken as a starting point for further very interesting epistemological questions as to how we can know our senses, on my proposal: Are the proposed personal-level perceptual channels a direct reflection of the sub-personal functional organization of our minds? Or are they somewhat analogous to the user interface of a computer, which at best corresponds very vaguely to its actual functional makeup? Unfortunately, I will have to leave a discussion of these questions for another occasion.

**Eva Schmidt**

Universität des Saarlandes
eva.schmidt@mx.uni-saarland.de

## References

Grice, H.P. 1962: 'Some Remarks about the Senses', in R. Butler (ed.): *Analytical Philosophy: First Series*, Oxford: Basil Blackwell, 133-153.

Macpherson, F. 2011: 'Individuating the Senses', in — (ed.): *The Senses: Classical and Contemporary Readings*, New York: Oxford University Press, 3-43.

O'Dea, J. 2011: 'A Proprioceptive Account of the Sense Modalities', in F. Macpherson (ed.), 2011, 297-310.

# The *arche* of Cognition – Grounding Representations in Action

Arne M. Weber & Gottfried Vosgerau

Unfortunately, the term "embodied cognition" does not refer to a unified theory and questions like, 'what does "embodiment" or "to be embodied" mean?' are still to be answered in detail as well as its explanatory advantage for theoretical progress remains unclear. To achieve some progress in this regard, we clarify theoretical presuppositions of a certain understanding of "embodied" cognition, namely "grounded" cognition. We present *grounded action cognition* as a theoretical framework for understanding the interdependencies of motor control and action-related cognitive processes, like perceiving an action or thinking about an action. We distinguish between grounding *qua* acquisition and grounding *qua* constitution of cognitive abilities and exhibit three possible theoretical conceptions, Furthermore, we draw on recent empirical evidence to motivate our inclination towards a particular theory. According to this theory we get a clearer picture of the architecture of mind and expose the ground of cognition: there are certain representations involved in action cognition and action perception that are not modality-specific as usually proposed by advocates of grounded cognition. The explanatory advance of those multi-modal action-related representations of the body can not only provide a clearer picture of the architecture of the mind but also of its *arche*.

## 1. Classical Cognitive Science and Embodied Cognitive Science

The embodied cognition research program contrasts with classical theories in cognitive science in that it motivates an understanding of cognition as *embodied*. This contrast of the emerging program consists in a critic of two ideas in classical cognitive science—the modularity of mind thesis (Fodor 1983) and the language of thought hypothesis (Fodor 1975). Fodor and, for example, Newell & Simon (1976), proponents of the *physical symbol system*-hypothesis, considered the elements of thought as symbols being manipulated and computed during cognitive processing. Cognitive processes were treated as computational functions over representations with syntactic structure and appropriate semantics. The productivity and systematicity of cognition were thought of as to be entirely explainable by language-like and rule-governed structures of thought. While thinking was understood by Fodor as some kind of mental language operating on a higher level, motor control and perception are conceived of as functioning on a lower level. Further, within this classical framework of cognitive sciences, motor control, perception and cognition are viewed as strictly separated faculties operating in different domains or modules of the mind; thought consisted in operations generally distinct from operations found in the sensorimotor system. For example, thinking and acting have been assumed to operate in different domains or modules because thinking draws on conceptual representations, whereas motor control functions are thought to rely on relatively low-level or automatic processes. Thus, thinking was characterized as entirely distinct from the sensorimotor abilities, like those governing perception and motor control. Consequently, the content of these representations was understood as independent from bodily experience and its modal basis.

This traditional, modular perspective on the domains of action, perception and thought is now often criticized as the "sandwich-model of the mind" (cf. Hurley 1998) since it assumes that the mind is built out of three distinct, separable layers: the input layer (perception), the

output layer (motor control), and the layer in between (cognition) as the delicious filling of the sandwich. The embodied cognition research program directly addresses the shortcomings of the traditional views, where the body and its interactions with the world are considered as peripheral to the understanding of the nature of mind and cognition. This deficiency is pointed out by proponents of a new research program stating that cognition is "embodied" or "grounded", though it is often unclear what these terms mean.

Proponents of the embodied cognition research program typically aim not only at challenging and further undermining classical views of the mind, such as the computational and representational theories of mind presented by Newell & Simon and Fodor, but they also often deny the commonly accepted assumption in the philosophy of mind that the mind is identical to the brain or is somehow realized by it. We say "typically" and "often" because the research program of embodied cognition does not have a single or specific theoretical perspective (for an overview cf. Wilson 2002). It is rather characterized by a heterogeneous variety of accounts, which altogether stress the relevance of the body of a cognitive system and its interactions with the environment for its cognitive processes; according to this general view, cognition presupposes a perceiving and moving body (cf. Gallagher 2005; Thompson 2007). Due to the variety of viewpoints, the precise meaning of the terms "embodied cognition" or "embodiment" is hardly univocal (cf. Clark 2008a, Clark 2008b, Kiverstein & Clark 2009, Shapiro 2011). Indeed, perhaps these different accounts are only unified by sharing a commitment to criticizing or even replacing traditional approaches to cognition. Many do so by focusing on the contribution of the nonneural body itself to cognitive processes, such that mind itself is constituted by both brain and body (see Shapiro 2011: 158-200). Others do so by arguing that real biological systems do not require representations to achieve cognitive tasks, and that a fortiori they should not be modeled as representational systems, because they are computationally too costly.

## 2. Representationalism and Anti-Representationalism

In the very beginning of embodied cognition approaches, researchers were generally inspired by anti-representationalist theories like Gibson's ecological psychology and new models for robotics capable for guiding adaptive behavior and flexible interactions with the environment. Gibson developed an *ecological theory of perception* (1979) to explain visual perception as a result of interactions between the body of a cognitive system and the environment. On this account perception is understood as "direct", i.e. not involving additional computation or internal representations. For Gibson there is no need to postulate further cognitive processing in addition to perception and action because for him all necessary information for guiding action is already given in perception. For the influence of Gibson's account for further conceptions of *embodied cognition* is most important that the bodily constitution of an organism determines every perception and action. And more precisely, especially the perception of one's own body is the underlying and prevailing principle for perceiving the environment:

> [E]xteroception is accompanied by proprioception – that to perceive the world is to perceive oneself. (Gibson 1979: 141)

Nowadays, the most prominent anti-representationalist view is the dynamical systems theory, as presented by, e.g., Beer (2000, 2003), Thelen and Smith (1994) and Port and van Gelder (1995), and most recently by Chemero (2009). Dynamical systems theory is used to explain adaptive behavior and is also supported by research in robotics and artificial intelligence (cf. Pfeifer and Bongard 2007). Generally, a dynamical system is understood as a system that changes over time and is best described by the elements and products of a set of differential

equations.[1] Such a view is also supported by Brook's (1991) research in robotics and in artificial life because his findings can be applied to explain adaptive behavior without any strong notion of representation. Brooks insists: "Representation is the wrong unit of abstraction in building the bulkiest parts of intelligent systems" (Brooks 1991: 140), because "the world is its own best model" (Brooks 1990: 6). Those authors do not postulate the existence of an internal representational model of a cognitive system that guides its behavior and its actions. Moreover, they reject the idea that a cognitive system ought to be conceived of as relatively independent from the world in its ability to representationally reproduce the external structure of the environment to guide its behavior.

But many still suspect that in developing an adequate theory of cognition, especially as regards belief formation and action planning, we need to postulate internal mental representations. For example, at first sight a theory of dynamical interaction between an organism and its environment alone provides no satisfying explanation when it comes to anticipatory behavior. A dynamical systems theory may face serious difficulties in providing an account of how systems are able to deal with "representation-hungry" tasks, such as those involving abstract thought and high-level reasoning. Cognitive processes like problem solving, planning, language acquisition, thinking about absent, even non-existent states of affairs, counterfactual reasoning and learning in general which are all conceived of as cases of purely internal processing in this context seem most naturally explained by appeal to internal representations (cf. Clark and Toribio 1994). So, the essential challenge is to explain those aspects of behavior that involve internal factors beyond immediate interaction.

We further want to pay attention to the relation between perception, action and cognition as conceived in dynamical approaches. For instance, Thelen et al. state that "cognition *depends* on the kinds of experiences that come from having a body with particular perceptual and motor capabilities that are *inseparably linked*" (2001: 1; emphasis added by the authors). But what does "inseparably linked" or "depend" mean for the mutual contribution of the domains or "capabilities" of perception, action and cognition while cognition is – in some not yet specified sense – "embodied"? What do we learn about the nature or architecture of the mind, its functional makeup and its relation to bodily functions? Is cognition in this sense "embodied" that it is entirely dependent of sensorimotor processes?

An adequate theory of embodied cognition ought to allow for a further understanding of the nature of mental representations. An alternative and more moderate (qua representationalist) view in embodied cognitive science is a project called "grounded cognition" (Barsalou 1999, 2008). Therefore, we discern that the major outstanding issues center on the representationalist understanding of *grounded cognition* in contrast to anti-representationalist conceptions of *embodied cognition*. The idea of grounded cognition is that cognitive and perceptual mechanisms share the same representational states: the core sources of representation that "ground" cognition are thought of as simulations in the brain's modal systems such as the sensorimotor system. Contrary to the language of thought hypothesis, the theory especially challenges the view that core representations are amodal symbols and data structures processed independently of the brain's modal systems for perception, action, and introspection. Barsalou summarizes the project of grounded cognition as follows:

> Grounded cognition rejects traditional views that cognition is computation on amodal symbols in a modular system, independent of the brain's modal systems for perception,

[1] This is not to claim that every possible dynamical account like the *dynamical systems theory* is *per se* an anti-representationalist view. There are works rebuilding and integrating even an elaborated notion of representation in dynamic approaches; for example, Spivey's "attempt to raise awareness of the benefits of emphasizing continuous processing, and therefore continuous representation as well" (Spivey 2007: 3) suggests some kind of "symbolic dynamics" (2007: 262 ff.), thereby reconsidering symbolic, but not computational representations.

> action, and introspection. Instead, grounded cognition proposes that modal simulations, bodily states, and situated action underlie cognition. (Barsalou 2008: 1)

The traditional assumption that knowledge resides in the form of amodal symbols in a modular semantic system separated from modality-specific systems is generally criticized. Following the grounded cognition approach, higher-order abilities such as thinking or conception are "grounded" in low-level sensorimotor abilities. Consequently, cognition is here understood as consisting of representations including activation patterns from various sensory modalities, i.e. the "perceptual symbol system" (Barsalou 1999). Thereby, Barsalou's perceptual symbol systems theory is also a sophisticated challenge of the classical separation of perception, cognition and motor control.

The proponents of grounded cognition frequently argue that there is a lack of direct empirical evidence for amodal representations as proposed in classical cognitive science. Gallese & Lakoff (2005) and Pulvermüller (1999) suggest that completely modality-free categories are rare, because concepts in general are distributed over modality-specific domains and involve reactivation of states in sensorimotor systems. Besides, Mahon & Caramazza (2008), for example, critically remark that the empirical data alone cannot decide between theories postulating modal or amodal representations because the recent experimental results are compatible with a huge variety of theories at hand. Even a 'disembodied' view of conceptual representation can be inferred from the empirical evidence and the 'embodied' view may be without empirical support compared to traditional theories. Hence, the most interesting question is whether there are only modality-specific representations, or whether we need to assume multi-modal or even amodal representations in addition.

## 3. Constitution and Acquisition

At this point it is only interesting for us what it means that a certain cognitive ability is "embodied" or – more specific – "grounded"; we will investigate the kinds of possible representations involved elsewhere later on. It is necessary to focus on plausible cases of cognitive processing in order to develop a comprehensive and precise account of "grounded". One obvious place to start with is what we will call *action cognition*, i.e. thinking about actions and perceiving actions. It is an obvious place to start with because this kind of thinking involves action-related concepts which may be grounded in our sensorimotor abilities, to perform and perceive these actions. By focusing especially on action cognition and action perception we want to get a close view on the interdependencies of the three domains of action, perception and cognition. In the case of action cognition, the "embodiment"-relation in question is as close and as direct as possible. More complex abstraction mechanisms within conception have not to be taken into our account here. Thus, we will analyze and examine the idea of *grounded cognition* by focusing especially on the relations between motor control, action perception, and action cognition; let us analogically call this "grounded action cognition". So, the guiding questions are how and to what extent action perception and action cognition are "grounded" in basic sensorimotor abilities? What are the implications for an adequate description and explanation and for the general conception of the architecture of the mind?

First, we provide an analysis of the term "grounded" that goes beyond mere metaphor and fosters a more specific theoretical understanding. To define the ambiguous term "grounded", we suggest recognizing it in terms of the conditions of acquisition or constitution of a given ability (cf. Weber and Vosgerau 2012):

Interpreted in terms of acquisition conditions, "grounded" means that ability A is grounded in ability B if B is necessary *to acquire* A.

Understood in terms of constitution conditions, "grounded" means that ability A is grounded in ability B if B is necessary to *possess ability* A.

For example, Held and Hein (1963) showed that self-produced movement is necessary for the development of vision (in cats). In their experiment they divided kittens into two groups: the kittens in one group pulled a carriage around a room in a horizontal rotating carousel; kittens in the other group rode the carriage. Both groups of kittens had the same visual stimuli, but the riding kittens did not develop certain normal perceptual abilities (e.g. depth perception). These kittens were not able to integrate their visual experiences with movement in their own bodies, which caused visual deficits in the long run (see also Gibson 1969). In short, self-generated change in sensory stimuli is an acquisition condition for certain perceptual abilities. In some cases though, once we have acquired A on the basis of B, B can be lost without disturbance of A. In the case of the kittens this means that a kitten that has already acquired normal sight will not lose its ability to see if we hindered its active movement. Active movement is thus necessary to acquire certain perceptual abilities, but not necessary to maintain them. This latter contrast is precisely the contrast between acquisition and constitution. Understood in terms of constitution conditions, "grounded" means that ability A is grounded in ability B if B is necessary to possess ability A. Whenever B is lost, A is lost as well (or at least severely impaired). The ability to move the legs is, e.g., constitutive of the ability to walk (it is necessary but obviously not sufficient).

Whereas some conditions are only necessary for the acquisition of mental abilities but are not relevant anymore once the ability is acquired, other conditions have to be fulfilled each time the ability is put into practice. Only the latter conditions will qualify as constitutive for the ability in the narrow sense. Whereas acquisition conditions could be viewed as mere contingent factors of mental development, constitution factors are part of the ontology of the abilities and are thus necessary in a metaphysical sense.

Regarding the different possible cases of constitution – entire, partial, and no constitution – we distinguish between three theses of grounded action cognition, which we label as "strong", "moderate" and "weak" (for details see Weber and Vosgerau 2012):

> (1) Strong thesis: basic motor abilities are constitutive for *all* processes within the domains of action cognition and action perception.

In other words: action cognition (and action perception) would be nothing but a subclass/a kind of motor abilities. This also implies that a complete breakdown of motor abilities ought to automatically result in a breakdown of action cognition and action perception. In sum, the strong reading of the thesis is equivalent to recent "motor theories" of thoughts (e.g., Campbell 1999; Ito 2008; Schmahmann 2004), which conjecture that thoughts are (a kind of) motor processes. Moreover, it implies that the classical boundaries between the domains of motor control and higher cognition are meaningless, and consequently that the whole idea of modular processing (for these abilities) has to be given up. If this thesis is correct and is taken seriously, it will drastically reshape our conception of the architecture of the mind.

> (2) Moderate thesis: motor abilities are constitutive for *certain* processes within the domains of action cognition and action perception, but not of others.

A breakdown in motor abilities would impair both action cognition and action perception; but it would not, on this view, lead to a complete breakdown of action cognition and action perception. Accordingly, there should be cognitive abilities that are not easily classifiable as part of the motor control domain or the action cognition domain as they can be counted in both classes (the classes overlap to a certain degree). Therefore, this thesis implies that that the boundaries between these different domains are blurred; however, it does not imply that we have to give up the distinction between different domains completely. If this moderate thesis turns out to be empirically adequate, the classical modular picture of the mind is seriously undermined.

(3) Weak thesis: motor abilities are *not* constitutive for action cognition and action perception but are still among the acquisition conditions for these abilities.

This means that motor abilities are necessary to acquire the abilities to perceive and to think about actions. Once we have acquired these perceptual and conceptual abilities, the basic motor abilities can be lost without any damage to action cognition or action perception. If this thesis turns out empirically adequate, our classical modular picture of the mind is not affected, although important information about the interrelation between different modules might be gained.

## 4. Motor Control and Cognition

Now we want to examine the plausibility of the three suggested versions of grounded action cognition. For the strong thesis, there are already various conceptual problems:

(i) The view faces the threat of an infinite regress – motor processes are triggered by intentions and intentions are best classified as thoughts; if thoughts a kind of motor processes, thoughts would thus *prima facie* be triggered by thoughts (cf. Gallagher 2004: 4). This general worry becomes even more acute when a motor theory of thoughts is tied to the comparator model of motor control. The central idea of this model is that the intention triggering the movement is compared with the actual outcome of the movement (e.g. Frith 1992 and many others; for a recent critical discussion see Synofzik, Vosgerau & Newen 2008). Applying this picture to thoughts, like e.g. Campbell (1999) did, means that the intention to think a certain content *p* is compared with the content of the actually occurrent thought *p*. In order to be comparable at all (in a computational manner), both contents have to be the same. Thus, the threat of infinite regress looms already generated for a single thought content, independent of the question whether the intention is consciously available or not (cf. Vosgerau & Newen 2007).

(ii) Central prototypical features of thoughts are not shared by motor control processes and vice versa. For example, thoughts seem to be compositional and (fairly) systematic in ways that movements are not. To further illustrate this point: if two concepts are combined, a new concept will systematically result. For instance, the combination of RED and BALL, e.g., yields the new concept RED BALL; the combination of BLUE and CAR yields the new concept BLUE CAR; we can also combine RED and CAR and have the concept RED CAR. But under certain descriptions the combination of two actions does not systematically lead to a new kind of action. For example, a grasping hand movement combined with a forward reaching arm movement can be described as a grasping action, but the grasping hand movement on its own is a grasping action of arguably the same kind. Moreover, thoughts stand in systematic inferential relations to one another (cf. Aizawa 2002; Vosgerau and Synofzik 2010); it is hard to see how this could be true of actions.

(iii) Further, a feature of endogenously generated movements is that they are controlled and corrected "online", i.e. during the execution of the movement. There are pathological cases in which patients reach too far or too short when trying to grasp a certain object, or the force of their grip is inappropriate in respect to the object they want to grasp. These failures of adjustment within motor control are different symptoms of "dysmetria". So some proponents of the motor theory of thoughts propose a "dysmetria of thought" in analogy to these pathologies (cf. Schmahmann 1998, 2004). This would mean that patients suffering from such a disease would intend to think a specific thought but would end up with something very close to it but not exactly the intended thought. This assumption seems, at the very least, radical. For intuitively, motor actions are dynamic entities with a trajectory through space and time; thoughts seem comparably "static" in precisely the sense that they do not (at least necessarily) move through space and time in any concrete manner. Thus, while we have a

clear idea of how small adjustments of movement trajectories are made in actions, "online"-controlling or -correction is somewhat less intelligible for thoughts and thus seems unlikely as a common feature (cf. Vosgerau and Synofzik 2010). Movements can be run "offline" in imagination—a feature that is hardly to be found in thinking; it would consist in imagining a thought without thinking it.

(iv) If the distinction between motor control and action cognition is denied, there would be no sense in talking about "grounded" cognition at all, since then there is nothing left to be "grounded". This means that the truth of the strong thesis would preclude the formulation of any claim about grounded action cognition and is thus to be abandoned as a sensible version of a grounded cognition thesis.

Thus, thoughts and motor processes are clearly distinguishable in important aspects and the strong thesis can already be rejected on conceptual reasons. Therefore, thoughts cannot be a kind of motor ability.

Now, we want to take a closer look on the empirical evidence that can decide between the two suggested more plausible versions of *grounded action cognition*. Several studies suggest that impaired motor control mechanisms influence action cognition, such that action cognition is degraded but not lost altogether. Consider for instance, that deficits in word-description matching associated with judgments about actions, but not judgments about objects, correlate with cortical atrophy in motor regions, such as amyotrophic lateral sclerosis (ALS) (cf. Grossman et al. 2008). This suggests that degraded (although not completely absent) knowledge of action features—an obvious case for a process of action cognition—is due to impairments of the motor cortex of those patients. If this is correct, then ALS is not only a neurodegenerative condition affecting the motor system, but also a cognitive impairment of the comprehension of action related words. Similarly, Parkinson's disease patients present a significant impairment, without complete loss, in their capacity to name actions (again, as compared to their capacity to name objects). Moreover, these results support the idea that verb representations are grounded in neural networks to which brain areas involved in motor control contribute (cf. Rodríguez-Ferreiro et al. 2009). The fact that motor features contribute to linguistic concept application and comprehension is evidence for the moderate thesis (2) and evidence against the weak thesis (3): motor abilities are constitutive of processes within the domain of action cognition.

Diseases of the brain's motor systems, e.g. of the cerebellum (cerebellar ataxia) or the basal ganglia (Parkinson's disease), have traditionally been considered to be solely "motor disorders". But as seen above, behavioral experiments with patients suffering from these disorders show that motor deficits are accompanied by action perception and action cognition deficits. Other studies demonstrate that not only motor, but also perceptual abilities are basic for action cognition (cf. Adamovich et al. 2001; Konczak et al. 2009; Maschke et al. 2003). Some motor deficits are actually secondary to action perception deficits, like shortcomings in the perception of one's own actions. For instance, it seems that patients with cerebellar lesions are impaired in motor control because they lack updated internal sensory predictions of the visual consequences of their motor behavior: thus, the patients' impairments in motor control are due to an impaired perception of their own actions and not the other way round (cf. Synofzik, Lindner and Their 2008). This suggests that action perception dysfunctions do not just accompany motor dysfunctions, but rather motor deficits seem to go—surprisingly—hand in hand with selective deficits in action perception and action cognition. Are these genuinely "motor deficits", or are they deficits in action perception, or even deficits in action cognition? The lack of clear answers here suggests that "motor disorders" are not just motor disorders. Overall, the fact that the boundaries between these different domains become blurry is consistent with (2) and is a counterexample to (3).

We now want to focus on the influence of motor factors and thoughts on perceiving one's own and other's agency. For example, misattributions of agency of one's actions in schizophrenia

reveal that perception and attribution of action are grounded in internal motor control processes, such as the so-called "efference copy" (cf. Feinberg 1978; Frith 1992; Frith et al. 2000). These misattributions in certain psychiatric patients indicate, however, that the central mechanisms underlying this experience can go astray. According to the influential comparator model, agency attribution is based on the comparison of the actual sensory information with the consequences of one's action, as predicted on the basis of internal action related signals such as efference copies. If this internal prediction matches the actual sensory event, an action is registered as self-caused; in case of a mismatch, the difference is interpreted as externally produced (cf. Synofzik et al. 2010). The question here is firstly, to what extent the perception and attribution of agency are grounded in primary motor control processes such as efference copy, and secondly, to what extent attribution of agency results from primary or accompanying thought processing and thus needs to be regarded as a rather high level process.

Indeed, it can be shown that perception of action does not (only) rely on motor control processes, but is largely penetrable and often dominated by primarily conceptual processes, e.g. background beliefs and emotions (cf. Synofzik, Vosgerau, & Newen 2008; Vosgerau & Synofzik 2010). In particular, it is argued that the explanatory scope of one of the currently most popular accounts of agency, namely the so-called comparator model of agency which proposes a pure motor account of agency perception, is severely limited because it leaves out this conceptual, non-motor part that plays a crucial role.

Motor control factors and the functional integrity of the motor system contribute to the understanding of one's own and other's actions even at a conceptual level. In the case of deafferentiation, for example, subjects do not receive peripherally originating sensory feedback of movements they perform, either at a perceptual level nor at the subperceptual level required for motor control. These deficits in proprioception, and in the internal processing involved in action planning, result in inadequate inferences concerning action-related beliefs. This became apparent when Bosbach et al. (2005) showed that two deafferentiated subjects lacking cutaneous touch and somatic proprioception show a selective deficit in interpreting another person's anticipation of weight, when seeing individuals lifting boxes. Bosbach et al. give an interesting explanation:

> [T]he interpretation of others' actions seems to require both observation of the other and a form of motor knowledge. More precisely, action recognition seems to involve the direct mapping of a perceptual representation of a particular observed action onto a representation of the appropriate motor pattern for the same action in the observer. [...] Thus, peripheral sensation from one's own body may contribute to inferences about certain mental states of other people derived from observing their actions. (Bosbach et al. 2005: 1295)

These patients suffer from lack of control and seem to have a general impairment in generating representations of their own actions; it is by this very deficit that they also show specific differences in action categorization. If this is correct, then adequately functioning motor processes in the observer can be understood as a constitutive condition for his abilities of action perception and action cognition. The fact that impairments of both action cognition and action perception are brought about by deficits in motor control mechanisms provide additional reasons to prefer the moderate thesis (2).

## 5. The Body and the Architecture of the Mind

The presented results provide evidence for the claim that impairments in motor control mechanisms lead to impairments in (already acquired) action cognition abilities without causing a complete breakdown of the latter (as it would follow from the strong thesis (1)).

Because of this partial impairment we discern that motor abilities are constitutive for *certain* processes within the domain of action cognition. The fact that we don't find a complete breakdown can also be seen as evidence against the strong thesis (1). Further, the selective impairment is a counterexample for the weak thesis (3), since it implies that impairments in motor control should not lead to impairments of conceptual abilities (already acquired), which is not the case. Altogether, we find empirical support for the moderate thesis (2): some motor abilities constitute action cognition and action perception. There seem to be clear interdependencies between motor control mechanisms and action cognition, but nevertheless none of the impairments of motor control considered leads to a complete breakdown. Thus, in our picture of an architecture of mind we find the three domains of motor control, action perception and action cognition in partial independence and in partial dependence regarding the specific breakdown patterns.

Against this general background of an architecture of mind we also have to reconsider the thesis of Thelen et al. that "cognition depends on the kinds of experiences that come from having a body with particular perceptual and motor capabilities that are inseparably linked" (Thelen et al. 2001: 1). By taking the moderate grounded action cognition thesis into account we can further explain in what sense these domains are "linked" or grounded. But the realms of the domains – motor control, action perception, and action cognition – are not entirely congruent and can still be differentiated, i.e. contrary to Thelen et al. they are not inseparable. We also have to attenuate Gibson's slogan "that to perceive the world is to perceive oneself" (Gibson 1979:141). We have shown cases as in deafferentiation in which exteroception is possible without any kind of proprioceptive feedback; given that this impairment results only in a selective deficit in action perception. If Gibson's slogan would have been true there should be no exteroception while proprioception is missing. But there is no complete breakdown of exteroception in the absence of proprioception – therefore, exteroception is only "accompanied", but not entirely constituted by proprioception. The pathological cases presented cannot be sufficiently explained by Gibson's anti-representationalist approach, because it does not take account of the condition of acquisition. We are inclined to say that the perceptual abilities may be acquired in company with proprioception, but only some perceptual abilities constitutively depend on perceiving one's own body. Those abilities require a special kind of internal mental representation, e.g. body representations – and more specific: action-related body representations.

For certain representations it is not decidable whether they belong to the motor control domain, the action perception domain, or the action cognition domain. So, we have to consider the possibility of representations which belong to several domains at the same time. In addition to the empirical fact that perceiving the actions of others in deafferentiation depends on internal – and in this special case missing – representations, it has been shown that some motor deficits are at least secondary to perceiving one's own actions. So, there is an obvious "overlap" between motor processes and action perception. Moreover, it turns out that some "motor disorders" are not just motor disorders, but in fact disorders related to representations that are multimodal or maybe even amodal. Therefore, thesis (2) also leads to the view that not only modality-specific representations exist, contrary to the usual assumption made by advocates of grounded cognition. Even if the boundaries between the different domains become blurry, there nevertheless is, in principle, no reason that a discrimination between the different domains cannot be made anymore. Indeed, all discussed empirical studies do meaningfully use this distinction in describing the phenomena. Thus, although the picture of different domains can be maintained, the classical modular picture of the mind should be abandoned in favor of an empirically corroborated theory of grounded action cognition.

## 6. Body Representation and the *arche* of Cognition

Based on Barsalou's general idea of grounded cognition we set out a specified grounded action cognition-perspective and referred to recent empirical evidence to motivate our view. Our formulation of a moderate grounded action cognition-thesis provides not only a critique but also a prolific prospect of current perspectives labeled "embodied cognition" or notably "grounded cognition", but may also give supplementary insights into the nature of mental representations and their relation to the body of a cognitive system. In our view, motor abilities are constitutive for *certain* processes within the domain of action cognition, but not all action cognition abilities acquired before are constitutively dependent on motor abilities. We want to further assume that certain body representations are essentially involved in action cognition and action perception. We have pointed out that some "motor disorders" are not just motor disorders, but in fact disorders related to multimodal representations. Thus, we assume that some of these representations that underlie cognition, namely action-related body representations, are not only modality-specific as usually claimed by proponents of grounded cognition. To argue for the importance of action-related body representations in what we call moderate grounded action cognition we want to refer to de Vignemont (2011):

> On a minimal definition of the notion of representation, a body representation is an internal structure that has the function to track the state of the body and encode it, that can misrepresent the body and that can be decoupled from the body. (de Vignemont 2011)

Overall, the existence of misrepresentations is a good indication that there are at least some representations involved (cf. Dretske 1986). This definition given by de Vignemont is useful for an explanation of all cited cases but at first we have to say a little bit more about such kind of bodily representations that can be "decoupled from the body". According to Prinz, to say of cognition in general, that it is embodied, means either that it "depends on the possession and use of a body, not just a brain" or that it depends on "mental representations or processes that relate to the body" (Prinz 2009: 420). De Vignemont and Alsmith (2012) keep track of this difference pointed out by Prinz:

> [W]e will call any view that gives a clear explanatory role to the body a 'strongly embodied' view (or 'strong embodiment'); by contrast, we will call any view that gives a clear explanatory role to representations of the body, whilst not also giving a clear explanatory role to the body itself, a 'weakly embodied' view (or 'weak embodiment'). (de Vignemont and Alsmith 2012: 3)

Wilson makes a similar point by arguing that the cognitive system as a representational system is able to "decouple" from the body itself while not putting actions into practice (or to run "off-line"; see Wilson 2002: 633). Given that the body itself is assigned a clear explanatory role in the general embodied cognition approach, it is sufficient for our current purpose to refer to the 'weak embodiment' concerning the representation of the body to provide an explanation of the given cognitive phenomena. Generally, two body representations can be distinguished based on functional criteria: the body image for recognition and the body schema for action (Gallagher 2005; Paillard 1999). The evidence provided to support the distinction between these two kinds of body representations relies mainly on neuropsychological dissociations between deafferentation (disruption of body schema) and neglect (disruption of body image): Deafferentiated subjects who have lost all their tactile and proprioceptive input from the neck down cannot use their whole body schema for motor control. But it is possible for them to control their movements by visual guidance while the intact body image has to fulfill a new and further task. In contrast, in the case of a unilateral neglect the patient ignores one side of his body (i.e. shows a disrupted body image), while the motor abilities for walking and the usage of both hands are not

impaired (implying an intact body schema). Thus, both the body schema and the body image contain perceptual information about the body and can be used to guide movements, i.e. are action-related representations (the body schema dealing with proprioception and automatic guidance, the body image dealing more with exteroception and visual guidance).

When we focus on action cognition we have to take a special class of representations into account which is particularly suited to embodied cognitive science, i.e. action-related representations (cf. Clark 1997; Mandik 2005). In the case of deafferentiation the proprioceptive bodily representation is not only a major source of information for the maintenance of posture and the governance of movement but is also crucial for the question how cognition arises. While an anti-representationalist theory is not sufficient to explain this fact, we argue that body representations (body schema and body image) act as an interface between perception and action. Moreover, these body representations classified as action-related representations play a fundamental role as a ground for cognitive abilities.

We have argued that action cognition is not grounded in modal specific (i.e. sensory and motor) representations, but essentially in multimodal representations of the body (i.e. body schema and body image), which result from the overlapping architecture of the mind (in particular: the overlap between action perception and motor control). Further, those representations are not only action-related bodily representations but they are also "grounded" as outlined in our framework of moderate grounded action cognition. We thereby want to motivate the claim that action-related representations of the body can be understood as the *arche* – or call it the beginning, ground, principle or origin – of cognition.

**Arne M. Weber & Gottfried Vosgerau**

Institut für Philosophie, Heinrich-Heine-Universität, Düsseldorf
arne.weber@phil.uni-duesseldorf.de
vosgerau@phil.uni-duesseldorf.de

## References

Adamovich, S.V., M.B. Berkinblit, W. Hening, J. Sage, and H. Poizner 2001: "The interaction of visual and proprioceptive inputs in pointing to actual and remembered targets in Parkinson's disease", *Neuroscience* 104 (4), 1027–1041.

Aizawa, K. 2002: *The systematicity arguments*. Norwell (MA): Kluwer.

Barsalou, L.W. 1999: "Perceptual symbol systems", *Behavioral and Brain Science* 22, 577–660.

— 2008: "Grounded cognition", *Annual Review of Psychology* 59, 617–645.

Beer, R.D. 2000: "Dynamical approaches to cognitive science", *Trends in Cognitive Sciences* 4 (3), 91–99.

— 2003: "The dynamics of active categorical perception in an evolved model agent", *Adaptive Behavior* 11 (4), 209–243.

Bosbach, S., J. Cole, W. Prinz, and G. Knoblich 2005: "Inferring another's expectation from action: The role of peripheral sensation", *Nature Neuroscience* 8, 1295–1297.

Brooks, R.A. 1990: "Elephants Don't Play Chess", *Robotics and Autonomous Systems* (6), 3–15.

— 1991: "Intelligence without representation", *Artificial Intelligence* 47 (1–3), 139–159.

Campbell, J. 1999: "Schizophrenia, the space of reasons and thinking as a motor process", *The Monist* 82 (4), 609–625.

Chemero, A. 2009: *Radical embodied cognitive science*. Cambridge (MA): MIT.

Clark, A. 1997: *Being there: Putting brain, body, and world together again*. Cambridge: MIT.

— 2008a: "Pressing the flesh: A tension in the study of the embodied, embedded mind?", *Philosophy and Phenomenological Research* 76 (1), 37–59.

— 2008b: "Embodiment & explanation", in P. Calvo and T. Gomila (eds.) *Handbook of cognitive science. An embodied approach*, San Diego: Elsevier, 41–58.

Clark, A., and J. Toribio 1994: "Doing without representing", *Synthese* 101 (3), 401–31.

de Vignemont, F. 2011: "Bodily Awareness", in Edward N. Zalta (ed.), The Stanford Encyclopedia of Philosophy (Fall 2011 Edition), URL = <http://plato.stanford.edu/archives/fall2011/entries/bodily-awareness/>. Retrieved 17 Dec 2012.

de Vignemont, F. and A. Alsmith 2012: "Embodying the mind and representing the body", in F. de Vignemont and A. Alsmith (eds.): *The body represented/Embodied representation*. Review of Philosophy and Psychology 3 (1), 1-13.

Dretske, F. 1986: "Misrepresentation", in R.J. Bogdan (ed.), *Belief*, Oxford: Oxford University Press.

Fodor, J.A. 1975: *The language of thought*. Cambridge (MA): Harvard University Press.

— 1983: *Modularity of mind: An essay on faculty psychology*. Cambridge: MIT Press.

Frith, C.D. 1992: *The cognitive neuro-psychology of schizophrenia*. Hove: Erlbaum.

Frith, C.D., S. Blakemore, and D. Wolpert 2000: "Explaining the symptoms of schizophrenia: Abnormalities in the awareness of action", *Brain Research Reviews* 31 (1–3), 357–363.

Gallagher, S. 2004: "Neurocognitive models of schizophrenia: A phenomenological critique", *Psychopathology* 37 (1), 8–19.

Gallese, V., and G. Lakoff 2005: "The Brain's Concepts: The Role of the Sensory-Motor System in Reason and Language", *Cognitive Neuropsychology* 22, 455-479.

Gibson, E.J. 1969: *Principles of perceptual learning and development*. New York: Appleton-Century-Crofts.

Gibson, J.J. 1979: *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Grossman, M., E. Zurif, C. Lee, P. Prather, J. Kalmanson, M.B. Stern, and H.I. Hurtig 2002: "Information processing speed and sentence comprehension in Parkinson's disease", *Neuropsychology* 16 (2), 174–181.

Grossman, M., C. Anderson, A. Khan, B. Avants, L. Elman, and L. McCluskey 2008: "Impaired action knowledge in amyotrophic lateral sclerosis", *Neurology* 71 (18), 1369–1401.

Held, R., and A. Hein 1963: "Movement-produced stimulation in the development of visually guided behavior", *Journal of Comparative and Physiological Psychology* 56 (6), 872–876.

Hurley, S.L. 1998: *Consciousness in action*. London: Harvard University Press.

Ito, M. 2008: "Control of mental activities by internal models in the cerebellum", Nature Reviews Neuroscience 9 (4), 304–313.

Kiverstein, J., and A. Clark (eds.) 2009: "The enacted mind and the extended mind", *Topoi: an International Review of Philosophy* 28 (1).

Konczak, J., D.M. Corcos, F. Horak, H. Poizner, M. Shapiro, P. Tuite, J. Volkmann, and M. Maschke 2009: "Proprioception and motor control in Parkinson's disease", *Journal of Motor Behavior* 41 (6), 543–552.

Mahon, B.Z., and A. Caramazza 2008: "A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content", *Journal of Physiology* 102 (1–3), 59–70.

Mandik, P. 2005: "Action-oriented representation", in A. Brook and K. Akins (eds.) *Cognition and the brain: The philosophy and neuroscience movement*, New York: Cambridge University Press, 284–305.

Maschke, M., C.M. Gomez, P.J. Tuite, and J. Konczak. 2003: "Dysfunction of the basal ganglia, but not the cerebellum, impairs kinaesthesia", Brain 126 (10), 2312–2322.

Newell, A., and H.A. Simon 1976: "Computer science as empirical enquiry: Symbols and search", *Communications of the Association for Computing Machinery* 19 (3), 113–126.

Paillard, J. 1999: "Body schema and body image: A double dissociation in deafferented patients", in: G. N. Gantchev, S. Mori, and J.Massion (eds.) Motor control, today and tomorrow , 197–214.

Pfeifer, R., and J. Bongard. 2007: *How the body shapes the way we think: A new view of intelligence*. Cambridge: MIT.

Port, R., and T. van Gelder 1995: *Mind as Motion*. Cambridge, MA: MIT Press.

Prinz, J. 2009: "Is consciousness embodied?", in P. Robbins and M. Aydede (eds.) *The Cambridge handbook of situated cognition*, New York: Cambridge University Press, 419–437.

Pulvermüller, F. 1999: "Words in the brain's language", *Behavioral and Brain Sciences* 22, 253–336.

Rodríguez-Ferreiro, J., Menéndez, M., Ribacoba, R., Cuetos, F. 2009: "Action naming is impaired in Parkinson disease patients", *Neuropsychologia* 47, 3271-3274.

Schmahmann, J.D. 1998: "Dysmetria of thought: Clinical consequences of cerebellar dysfunction on cognition and affect", *Trends in Cognitive Sciences* 2 (9), 362–371.

— 2004: "Disorders of the cerebellum: Ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome", *Journal of Neuropsychiatry & Clinical Neurosciences* 16 (3), 367–378.

Shapiro, L. 2011: *Embodied cognition (New problems of philosophy)*. New York: Routledge.

Synofzik, M., A. Lindner, and P. Their 2008: "The cerebellum updates predictions about the visual consequences of one's behavior", *Current Biology* 18 (11), 814–818.

Synofzik, M., G. Vosgerau, and A. Newen. 2008: "Beyond the comparator model: A multifactorial two-step account of agency", *Consciousness and Cognition* 17 (1), 219–239.

Synofzik, M., P. Thier, D.T. Leube, P. Schlotterbeck, and A. Lindner 2010: "Misattributions of agency in schizophrenia are based on imprecise predictions about the sensory consequences of one's actions", *Brain* 133 (1), 262–271.

Thelen, E., and L.B. Smith. 1994: *A dynamic systems approach to the development of cognition and action*. Cambridge: MIT.

Thelen, E., G. Schöner, C. Schleier, and L.B. Smith. 2001: "The dynamics of embodiment: A field guide of infant perseverative reaching", *Behavioral and Brain Science* 24 (1), 1–86.

Thompson, E. 2007: *Mind and life*. Cambridge: Harvard University Press.

Vosgerau, G., and A. Newen. 2007: "Thoughts, motor actions, and the self", *Mind & Language* 22 (1), 22–43.

Vosgerau, G., and M. Synofzik 2010: "A cognitive theory of thoughts", *American Philosophical Quarterly* 47 (3), 205–222.

Weber, A.M., and G. Vosgerau 2012: "Grounding Action Representations", in F. de Vignemont, and A. Alsmith (eds.) *The body represented/Embodied representation*. Review of Philosophy and Psychology 3 (1), 53-69.

Wilson, M. 2002: "Six views of embodied cognition", *Psychonomic Bulletin & Review* 9 (4), 625–636.

Wilson, R.A., and L. Foglia 2011: "Embodied cognition", in Edward N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy* (Fall 2011 Edition). Retrieved 5 Jan 2012.

Wolpert, D.M., and J.R. Flanagan 2001: "Motor prediction", *Current Biology* 11 (18), R729–R732.

Wolpert, D.M., and Z. Ghahramani 2000: "Computational principles of movement neuroscience", *Nature Neuroscience* 3, 1212–1217.

Wolpert, D.M., Z. Ghahramani, and M.I. Jordan 1995: "An internal model for sensorimotor integration", *Science* 269 (5232), 1880–1882.

Wolpert, D., R.C. Miall, and M. Kawato. 1998: "Internal models in the cerebellum", *Trends in Cognitive Sciences* 2 (9), 338–347.

# Integrating Evaluation and Affectivity into the Intentionality of Emotions

Wendy Wilutzky

What characterizes all emotions is their pronounced affective and inherently evaluative nature and any adequate theory of emotions must account for how these features characterize the intentionality of emotions. As a case in point I will discuss the account put forward by Michelle Montague (2009). Montague dissects the intentionality of emotions into a cognitive, an evaluative and an affective content and attempts to explain how these relate to one another. Central to her account is the notion of *framing*, by which she means to denote not only a *thin* content but a kind of *thick* cognitive content in which a state of affairs is represented as bearing a certain relation to a subject. One and the same state of affairs may be framed in different ways, e.g. as a success for someone or as a failure for someone else. The framing itself, however, is not yet evaluative or has any connection to affective content. Instead, the framing first needs to be associated with an evaluative content, which may then bring about a third kind of content, namely the affective phenomenality of emotions. Montague's account poses two problems. First, the separation of the framing process from the evaluative content of an emotion seems questionable, since the framing of a state of affairs is itself already an evaluation of a situation. Secondly, the affective, evaluative and cognitively framed contents of an emotion appear as distinct objects in emotion experience according to Montague. This is not only phenomenologically implausible but furthermore neglects the way in which the affective content of an emotion may inform the other aspects of an emotion's intentionality. These two points of criticism will be explicated by contrasting Montague's account with those of Bennett Helm and Peter Goldie in these respects. The overall conclusion to be drawn from these considerations is that the evaluative and affective contents of emotions are not distinct components that need only be added to an otherwise 'cold' or purely cognitive intentionality. Instead, the evaluative and affective contents of emotion are intertwined and also figure in the cognitive content.

## 1. Introductory Remarks

In the past, emotions have often falsely been characterized as a merely bodily reaction to a distinct mental content, i.e. as some kind of bodily arousal. Since the emergence of cognitivist theories of emotions, however, it has been commonly acknowledged that emotions are not only bodily responses to a cognitive content, but that emotions themselves are best to be regarded as cognitive phenomena. The reason for this reconceptualization of emotions was the insight that emotions have intentionality, that is, that they are always directed at or about the world: one is afraid of something, happy about something, angry at someone etc. Emotions, since they have intentionality, clearly belong to the class of mental phenomena where they form a kind of mental state of their own, as their intentionality is not reducible to that of other mental states, such as beliefs and desires (this point has been extensively and rather successfully argued for by e.g. Goldie 2000, Slaby 2008). What sets emotions apart from other mental states is their pronounced affective and inherently evaluative nature and the way in which these characterize an emotion's intentionality.

A few years ago, Michelle Montague (2009) put forth an account of emotions intended to address just these issues. Montague not only showed that emotions are not reducible to beliefs and desires, but also insisted that besides a cognitive content, an affective and an evaluative content are equally to be included in an account of the intentional structure of emotions. Although in general Montague's efforts and aims are favorable and in many

respects promising, she, unfortunately, misconstrues the role which the evaluative and affective aspects play in bringing about an emotion's intentionality. As will be shown, the main problems of Montague's account are (1) that the cognitive contents of emotions cannot be construed independently from the evaluative contents, and (2) that the affective content of an emotion may not be understood as the result of a previously established cognitive and corresponding evaluative content, but that the affective content instead can contribute to an emotion's cognitive and evaluative properties. The critical discussion of Montague's account is meant to illustrate the immense difficulty of explaining how the cognitive, evaluative and 'affective' aspects of emotions are intricately intertwined and that an account of emotions which fails to acknowledge the substantial contribution of the evaluative and affective aspects to the intentionality of emotions is untenable.

## 2. Affective, Evaluative, and Cognitive Content of Emotions

As this paper's main concern is the interdependence of cognitive, evaluative and affective contents in emotions, the first point of order is a clarification on what is meant by these terms. In the emotion literature affect is generally understood as the (foremost bodily) feeling which is present in emotions, moods and other so-called affective phenomena. When in an affective state we are influenced, moved or 'affected' to a greater extent by an event than by regular perception or thought. We experience some kind of bodily arousal in form of bodily feelings; e.g. when afraid, we feel our muscles tense, the palms of our hands getting sweaty, our hearts beating faster, maybe feel the urge to run away etc. While in the empirical emotion research the term affect usually refers to these bodily changes themselves (e.g. the galvanic skin response or rate of heart beat is measured), in the philosophy of emotions 'affect' denotes the feeling and experience of these bodily changes. Feeling these bodily changes does not necessarily entail that one is aware of the particular physiological reactions that one is experiencing, and insofar feelings are not necessarily about the occurring physiological changes. As Peter Goldie (2000) has convincingly argued, these feelings and their affective content can be directed towards objects and events in the world, so that, e.g., the feeling of one's churning stomach when disgusted at something is not about the current state of one's digestive system but is in part the way in which one thinks of that object or experiences it as disgusting. An emotion's affective phenomenology is inextricably linked with the emotion's intentionality, and thus, if emotions are directed towards events and objects in the world and not only at one's bodily constitution, affective content too has content that is about or directed at something beyond the body.

Emotions are also evaluative, that is, in emotion a state of affairs is experienced as something of value or disvalue, as good or bad for oneself. To again make this point exemplarily: being happy about a certain event involves some kind of recognition that the event in question is of positive value for oneself or conducive to one's goals; when afraid of something, that something is evaluated as a potential threat or danger to oneself or one's goals etc. The central role which evaluation plays in emotion has led some philosophers of emotion to equate emotions with evaluative judgments (cf. e.g. Solomon 1993 or Nussbaum 2001). Such theories, however, have typically treated emotions as a kind of belief or a judgment that is made on the basis of beliefs, which, as we will see shortly, leads to an inadequate characterization of emotions' intentionality. Furthermore, such so-called cognitivist or reductive accounts of emotion have been criticized for 'over-intellectualizing' emotions (cf. Goldie 2002), disregarding emotions' possible 'inferential encapsulation' (cf. de Sousa 1987), and neglecting the role of the pronounced bodily or affective nature of emotions. But although it thus appears that emotions are best not characterized as full-blown, rational judgments or beliefs about value, there is nonetheless some evaluative contents that they bear: some emotions (such as joy, hope, love) feel good, whereas other emotions (such as fear, anger,

regret) feel bad. This 'goodness' or 'badness', which Jan Slaby (2008) has aptly described as the 'hedonic valence' of emotions, is somehow related to the value of the event, which the emotion is directed at, has for you.

Despite the aforementioned warnings about an overly cognitive (a.k.a. cognitivist) construal of emotions, emotions do have cognitive content. (Note that I am not claiming that emotions require cognitions, but that they themselves have cognitive content. Hence, on this view, emotions are cognitive phenomena themselves, and not non-cognitive states which result out of previously established cognitive states.) Any evaluation of a situation clearly requires some cognitive assessment of the situation and possibly an understanding of implications a situation may have. Certain emotions rely heavily on cognitive content (e.g. indignation, regret, fear of losing money on the stock-market) since they require the capacity for counterfactual reasoning, whereas other instances may involve minimal cognitive content (e.g. fear of an approaching predator), where no inferences must be drawn.

## 3. Michelle Montague's Account of Emotions' Intentionality

The above characterization of affective, evaluative and cognitive contents is consistent with Montague's construal of these terms. Montague (2009) neatly dissects emotions' intentionality into these three components and then proceeds to explain how each of these comes about and figures in the intentionality of emotions.

Montague sets out by arguing that an emotion's rich intentionality cannot be understood only in terms of thin content, but can only be properly construed in terms of thick content. Often, Montague points out, the content of an intentional attitude is thought to be that which is designated by the that-clause in the familiar intentional attitude statements of the kind "He believes *that it is raining*". This is what Montague refers to as thin content, which is nothing but the denotation of a state of the world, and thus wholly distinct from the mode of presentation and from thinking or feeling. In contrast, in thick content "everything that one experiences in having the experience" is captured, so that the mode of presentation, e.g. doubting that p versus believing that p, makes a difference to the content (ibid 173). Thick content therefore comprises not only thin content but also phenomenological content, which Montague dissects into several components, among which we find cognitive, affective and evaluative contents (which are of interest here), as well as sensory and Fregean content. In Montague's words, affective content is the "feeling associated with an emotion" that "is discernible by the (affective) coloring that the intentional attitude verb indicates" (ibid 174). The evaluative content of an emotion is the representation of a situation's evaluative properties, such as good, just, or bad.

What is crucial in bringing about the affective and the evaluative contents of emotion is a cognitive content that is framed in a certain way. Framing, according to Montague, is the mode of presentation of a state of affairs or the way in which a state of affairs is thought of by a subject. The need for this notion of framing is illustrated with the example of Jane, who is the only successful applicant for a grant, which is tantamount to all her competitors failing in the application process. Although Jane may be aware that her success implies everyone else's failure, she may be happy about the former but not about the latter. This is because Jane frames the same state of affairs in two different ways – once as a success for herself and once as a failure for everyone else – and these differently framed contents lead to disparate emotions with different associated evaluative and affective contents. (The depicted example resembles a Fregean puzzle but also is decisively different from the typical cases which are applied to beliefs: When learning, to play on a classical example, that Cicero is Tully one immediately draws the inference that Tully was a famous Roman orator and thereby make one's beliefs concerning Cicero and Tully coherent. Anything else would be a failure of

rationality. In emotion, however, different inferential sensitivities prevail and it is no breach in rationality to experience disparate emotions concerning one and the same state of affairs since the emotions concern different framings (ibid 178).) Montague argues that it is due to the different framings that one and the same state of affairs may lead to diverging evaluative content and thereby in turn lead to different emotional responses and affective contents (ibid 179). More specifically, Montague maintains that a state of affairs is first non-evaluatively framed and only subsequently a "value property is associated with" that framed content, thereby determining one's emotional response (ibid 179). The emotional response then in turn yields a certain affective content.

Montague intends to exceed those theories of emotions which reduce emotions' intentionality to beliefs and desires that trigger some bodily arousal. She therefore shows that the cognitive content of emotions is different to that of beliefs and desires, in particular that it cannot be captured in terms of that-clauses, but must be understood in a sufficiently fine-grained manner, namely in terms of a framed content. For, it is only by reference to the framed content that the evaluative and affective content of an emotion can be appropriately specified. Furthermore, by her appeal to recognize emotions' thick content, the affective content is treated as part of an emotion's intentionality instead of a negligible byproduct of a cognitive process. All in all, these suggestions are rather favorable and the objectives are promising. However, as appealing and comprehensible as Montague's neat dissection of emotions' intentionality into different kinds of contents may seem at first glance, the way in which the individual contents - cognitive, evaluative and affective - are treated as distinct from one another runs into difficulties, as shall become evident in the following.

## 4. Critique of Montague's Account

To be clear, this is not a resistance to the decomposition of the intentionality of emotions into individual components, which may be analyzed independently, in general. Such an analysis may be very fruitful to a deeper understanding of certain aspects. However, when dissecting a complex phenomenon such as the intentionality of emotions into individual components, one also owed an account of how the single parts interact and one must take heed to explain how they come together to form the holistic phenomenon which it was one's aim to explain in the first place. This second part of the task is strongly neglected by Montague in her analysis. The intentional structure of emotions, which results from Montague's clear-cut model, has its shortcomings, particularly regarding the evaluative and affective aspects of emotions' intentionality. As we will see, Montague's adherence to the dissociation of different aspects of an emotion's intentionality, however orderly and thus appealing it may be, cannot do justice to the reality of emotional experiences.

### *4.1 First Problem: Severing the Evaluative from the Cognitive Content*

Firstly, the separation of the framing process from the evaluative content of an emotion is questionable. According to Montague, a framed content is itself not yet evaluative or affective, but a non-evaluative cognitive content that correlates with a value property. For instance, framing a state of affairs as a success is generally associated with positive value so that Jane feels joy when she frames her acceptance in terms of success. The evaluation of the situation is thus arrived at by first 'neutrally' representing a state of affairs, which is only thereafter labeled with a certain value.[1] However, this severance of framing and evaluation is

[1] What is equally worrying is the question of whether on Montague's explanation the framing of the state of affairs itself is arrived at only after a full and unframed representation of the situation. Montague does not elaborate on this issue, but as a perception of the entire situation is described as the starting point from which different framings are established, it does not seem too far fetched to be considered as

unwarranted, for representing something as a success in emotion already is an evaluation of the state of affairs. (This does not mean that one cannot represent the conditions of someone's success in a non-emotional way, e.g by coolly analyzing the situation's implications without any personal investment. But when one perceives something as a success in emotion, it is not first represented non-evaluatively.)

In this respect Bennett Helm's (2009) account of the evaluative nature of emotions' intentionality advances over Montague's. For Helm, the framing or, in his terminology, the representation of a state of affairs relative to a subject's focus already gives the evaluative content of an emotion, due to its inherent relatedness to the subject's interests and concerns: Jane only feels joy over her successful application because she regards it as something desirable and something that is of import to her. She cares about winning the grant and this caring commits Jane not only to feel joy when winning it, but also to an entire range of emotions, such as hoping to win the grant, fearing that she will not get it and being disappointed if she should not get it. All these emotions arise from one and the same focus, namely that winning the grant has import to her. Any failure of Jane's to respond emotionally in such ways under the respective circumstances would entail either a failure in rationality or that winning is in fact not of import to her. Likewise, if, for whatever reason, Jane was hoping not to win she would be disappointed upon hearing that she won, even though according to Montague Jane's framing of the situation (i.e. her being successful) would have to be the same as in the original scenario. Montague fails to provide an explanation of why one and the same framing would be associated with such disparate evaluative contents. In fact, Montague's proposal has the even greater difficulty of explaining why a non-evaluative, cognitive framing of a situation is at all associated with an evaluative content and thereby leads to an emotion, whereas other framings of situations could go by without any emotional response at all. This is what Helm has termed the "problem of import": why are certain cognitive states infused with emotionality, whereas others remain "cold" and lead to no emotional response? Helm's answer is that certain evaluative aspects (such as import and concern) must be in place before a situation is cognitively assessed, as any post-hoc addition of evaluative contents would only return to the problem of import (cf. Helm 2009, 250). In other words: you have to value something first in order to have any kind of emotional response at all. Situations are not assessed from an "emotional nowhere" and then eventually lead to an emotion. Rather, there is always a background of import and concern out of which emotions arise and which scaffold the cognitive content. A subject's focus on those things that are of import to her direct her attention and guide her perception of a situation, thereby making Montague's proposal of a fully non-evaluative, cognitive framing as the first step in constructing an emotion's intentionality nonviable.

The same point can be made by appeal to an established distinction of the intentional objects of emotions: the target and the formal object of emotion. The target of an emotion is that which an emotion is directed at in a particular situation (e.g. Jane's winning the grant), whereas the formal object is the relation in which the experiencing subject stands to the target and which makes the emotion intelligible (e.g. winning the grant is a success).[2] It is widely held that emotion types are characterized by different kinds of formal objects, so that the formal object of fear for instance is the evaluation of a target as dangerous to one and in sadness a target object is assessed as irretrievably lost. Seemingly unbeknown to Montague, the notion of a formal object resembles her concept of framing, as is, inter alia, suggested by

a possibility. If this were indeed Montague's claim, then her account would by no means square well with studies on how attention and perception mechanisms are guided by emotions (cf. Lewis 2005).

[2] I will leave it open here whether the formal object requires the reference to a belief (e.g. the belief that a certain target is dangerous, as Anthony Kenny proposed) or whether it is a property of the situation (as Ronald de Sousa has suggested), and furthermore whether formal objects are in fact characteristic of emotion types (i.e. that the formal object of fear is to evaluate something as dangerous, that of anger is to evaluate a target as offensive etc.).

the examples of framing that she uses (e.g. winning the grant being framed *as a success*, *the sadness* of one' cat's death). However, the relation of the formal object to the target object (i.e. whether a dog charging at you is perceived as dangerous or whether your cat's death is considered sad or whether Jane regards winning the grant as a success) can only be established with respect to the subject's focus on that object, that is, what import or value the target has for the subject (cf. Helm 2009, 5). Thus, the target of Jane's joy (winning the grant) can only be framed as a success, and thereby establish the formal object, because Jane's focus on the target object is of such a kind that she regards it as something that she values and deems desirable. In order for emotions to occur, certain evaluative aspects must be in place from the very beginning and shape any further cognitive assessment of a situation. Situations are not assessed from an "emotional nowhere" and then eventually lead to an emotion. Rather, there is always a background of import and concern out of which emotions arise and which scaffold the cognitive content. Montague's proposal of a non-evaluative or "neutral" framing prior to evaluative content is simply irreconcilable with these considerations.

To use a cognitively less loaded and affectively more salient example, imagine you are watching a small child that you are fond of being approached by a large, barking dog. Because the child has import to you, it is an object worth of your attention and, if so required, also your action. You are constantly vigilant and prepared to act on its behalf, so that watching the dog coming close to the toddler is bound to grab your attention and, if so required, also your action. You are afraid for the child, and this fear is not merely a response to the barking dog being associated with a danger, possibly through some inborn instinct, but arises from your concern for the child. The dog is perceived as a threat or danger *to* the child, not as a potential threat or danger that just happens to be in the child's vicinity. In other words, the dog is perceived as a danger not because he has a general property of 'dangerousness' which can be cognitively and non-evaluatively assessed or framed, but because it stands in a certain relation to an object that has import to you.

What seems to lie at the root of Montague's problematic proposal that a situation can be non-evaluatively framed as joyful, sad etc. is her assumption that perceptions of value may be true or false, or, put differently, that value can be correctly or incorrectly represented. "[W]hether the state of affairs in question has been legitimately or accurately framed", in Montague's opinion, depends on the correct representation of "the evaluative features of the state of affairs in question" (Montague 2009, 181). For Montague values seem to be out there in the world and need only be detected. Hence, framing is merely a representation of what is given in the world itself and not an active evaluation of the situation performed by the subject. That this assumption of the objectivity of value is untenable when wanting to explain the evaluative intentionality of emotions should be evident by now, given Helm's considerations on the importance of a subject's concerns in determining a situation's value for a subject. To repeat, the value of a situation for a subject can only be determined with respect to her other evaluations and concerns.

### *4.2. Second Problem: Affective, Evaluative and Cognitive Contents as distinct from one another*

Secondly, Montague treats the cognitive, evaluative and affective contents of emotion as distinct components and, what is more, even as distinct objects in the experience of emotions. Considering the example of being sad over a cat's death, Montague claims that the non-evaluative framing of the cat's death as sad is "experienced as something as objective" and is an object which the subject stands in an intentional relation to (ibid 183). This intentional relation in which the subject stands to the "objective sadness" of the situation gives rise to a second intentional object, the affective content. Montague wishes to surpass those emotion theories which regard affectivity as a mere accompanying, non-intentional feeling to emotion and argues that the affective content itself is world-directed, namely in virtue of being an

experiential representation of the framed situation. However, Montague also explicitly states that "the sadness of the cat's death, experienced as something objective" (ibid 183) is distinct from the affective phenomenology of the experience, i.e. the feeling of sadness, and even maintains that they appear as distinct objects in the experience of emotion. This simultaneous endowment of affectivity with intentionality and segregation of the affective content from the cognitive content is problematic.

On the one hand, as Peter Goldie (2000) has pointed out, it is phenomenologically implausible to assume two such independent objects in emotion: one does not first judge a situation to be e.g. dangerous and in addition feel fear; rather, one "thinks of the danger with feeling" from the very start (Goldie 2000, 40). While emotion can be made intelligible by reference to beliefs, desires and feelings, when actually experiencing an emotion or acting out of emotion, beliefs and desires do not usually figure in the experience's content (cf. ibid 147). Furthermore, Goldie has shown that both cognitive and feeling elements may be present in consciousness without them being experienced as distinct objects: There is an ambiguity in saying that A feels an emotion E. It can either simply mean that A has feelings which are E-related, or it can mean that A is aware of feelings which A recognizes as being E-related. In the latter case A is reflectively conscious of her feelings and thus the feelings are indeed an individual object in the experience of emotion as A directs her attention towards them. On the former interpretation A is "unreflectively emotionally engaged with the world" so that she is not aware of her feelings in an object-like manner, but nonetheless A's thoughts and actions are structured by these feelings so that it would be wrong to claim that A is unconscious of her feelings (ibid 64-66). As it stands, Montague's proposal cannot accommodate the familiar phenomenon of unreflective feeling described by Goldie. What is needed is an explanation of how the affective content structures the cognitive content, so that both come together as one object in experience.

On the other hand, the separation of the affective from the cognitive content, and in particular Montague's description of the affective content as the result of the interactions between evaluative and cognitive contents, neglects the ways in which affectivity itself can contribute to both the cognitive and the evaluative aspects of emotion. The rendition of the affective phenomenology of emotions as an experiential representation of the evaluative content, which in turn is arrived at via the cognitive framing, implies that the affective content is merely a reiteration of a content that has already been previously established in a non-affective way (i.e. what was formerly "a representation of her joyful win" now is "an experience of her joyful win", ibid 190). Although Montague concedes that affective content may have a reinforcing effect on the cognitive content, so that the feeling of sadness may strengthen the framing of one's cat's death as sad (cf. ibid 184), no further way in which the affective content may help establish the intentionality of emotions is discussed. Indeed, any such endeavors would most likely stand at odds with Montague's previously outlined account, since the affective content is effectively the result of (or something triggered by) previously established cognitive and evaluative contents. That the affective contents of emotions is more than a byproduct of an otherwise non-affective, cognitive process becomes evident in Helm's portrayal of the interconnectedness of evaluative and affective contents: the loss of an object of affection is felt as painful, and it is in virtue of this feeling that the loss is evaluated as bad. Conversely, it is because of the value that the lost object has that its loss pains you in the first place (Helm 2009, 250). Affect and evaluation are thus interdependent and affective content therefore can certainly not be regarded as the end of a causal chain of events. Instead, affective content can contribute to cognitive and other functions in several respects. For one thing, the affective contents "reverberate through [a subject's] entire mental economy, affecting not only her desires, her expressive behaviour and the way in which she acts [...] also her imaginations and memories" (Goldie 2002, 245). The way in which we think of those objects towards which we feel emotion changes drastically due to the affective content that is experienced, and these new ways of thinking can lead us to generate new goals and desires -

which evidently are cognitive efforts. Furthermore, affective phenomenology has a motivational pull to it, i.e. part of the feelings experienced in emotion are motivational feelings, e.g. to run away in fear, attack your offender or jump for joy (Helm 2009, 249). These motivational feelings are not felt in addition to the emotion but rather are essential to how the emotion feels, i.e. to her affective content. Relatedly, Goldie (2000) who convincingly argues that actions performed out of emotion cannot be explained by reference to the agent's beliefs and desires alone and the fact that these actions just happens to be accompanied by certain feelings. Rather, an action performed out of emotion (e.g. fearfully running away or passionately making love) is fundamentally different from that same action when not performed out of emotion, in that it is driven and motivated by the agent's feelings (Goldie 2000, 40). Montague not only fails to address any of these roles of the affective content, but, given her characterization of affective content as the experiential representation of a previously established cognitive content, makes them difficult to square with her remaining account.

## 5. Conclusions

The discussion of Montague's account was meant to illustrate just how difficult it is to adequately describe how the evaluative character of emotions' intentionality comes about and how the evaluative and affective contents are intricately intertwined with one another and also with the cognitive contents of emotion. Summarizing, the conclusions to be drawn here are (from the first point) that the way a state of affairs is represented is determined by the value it has for a subject and (from the second point) that the affective content's contribution to the cognitive and evaluative aspects of the intentionality of emotions must be recognized. Although Montague's notion of thick content was meant to explain the richness of emotions' intentionality by including cognitive, evaluative and affective contents, the above considerations hopefully show that it is not enough to include these as intentional objects, but that the truly demanding task lies in describing the relation of these to one another. However tempting it may be to succumb to the idea of deconstructing emotions' intentionality into single components in order to gain a deeper understanding of the intentional structure underlying emotions, without an adequate description of how these components come together to produce the complex phenomenon that is the intentionality emotions, more insight might be lost than actually gained.

**Wendy Wilutzky**
Universität Osnabrück
wewilutz@uni-osnabrueck.de

## References

de Sousa, R. 1987: *The Rationality of Emotions*, Cambridge: MIT Press.

Goldie, P. 2000: *The Emotions: A Philosophical Exploration*. Oxford: Oxford University Press.

Goldie, P. 2002: 'Emotions, Feelings and Intentionality', *Phenomenology and the Cognitive Sciences* 1, 235–254.

Helm, B. W. 2009: 'Emotions as Evaluative Feelings', *Emotion Review* 1(3), 248–55.

Lewis, M. D. 2005: 'Bridging emotion theory and neurobiology through dynamic systems modelling', *Behavioral and Brain Sciences* 28, 169–245.

Montague, M. 2009: 'The Logic, Intentionality and Phenomenology of Emotions', *Philosophical Studies* 145(2), 171–192.

Nussbaum, M. 2001: *Upheavals of Thought – The Intelligence of Emotions*, Cambridge: Cambridge University Press.

Slaby, J. 2008: *Gefühl und Weltbezug. Die menschliche Affektivität im Kontext einer neo-existentialistischen Konzeption von Personalität*, Paderborn: mentis.

Solomon, R. 1993: *The Passions: Emotions and the Meaning of Life*, Hackett Publishing.

# Nichtwillentliche Aktivität

André Wunder

Wenn es so etwas wie nichtwillentliche Aktivität gibt, was wäre dann ein Kriterium für eine solche Aktivität? Es werden zwei Kriterien vorgestellt, das von Kenny und Alvarez/Hyman und das von Dretske. Beide Kriterien können nicht überzeugen. Ausblickend wird jeweils ein weiterer Ansatz zur Bestimmung nichtwillentlicher Aktivität bei den betrachteten Autoren aufgezeigt.

## 1. Die Unterscheidungen zwischen *willentlich* und *nichtwillentlich* und zwischen *aktiv* und *passiv*

Nach John Hyman wird von vielen Philosophen die Aktiv-Passiv-Unterscheidung mit der Willentlich-Nichtwillentlich-Unterscheidung vermengt, so dass aktiv und willentlich auf der einen und passiv und unwillentlich auf der anderen Seite gleich gesetzt werden. Nichtwillentliche Aktivität und willentliche Passivität werden deshalb weitestgehend ignoriert.

> In reality the voluntary/non-voluntary distinction and the active/passive distinction cut across each other, since activity can be either voluntary or not voluntary, and the same is true of passivity. But philosophers have commonly ignored or failed to notice two of these possibilities. [...] They thought about voluntary activity, but they have ignored voluntary passivity, or even denied that it exists. On the other hand, activity and voluntary activity have commonly been equated, as if activity were always voluntary. So they have ignored activity that is not voluntary. (Hyman 2011: 296)

| | **Willentlich** | **Nichtwillentlich** |
|---|---|---|
| **Aktiv** | Willentliche Aktivität | Nichtwillentliche Aktivität |
| **Passiv** | Willentliche Passivität | Nichtwillentliche Passivität |

Ob dieser Vorwurf insgesamt berechtigt ist, soll uns hier nicht weiter interessieren. Die Annahme nichtwillentlicher Aktivität jedenfalls scheint berechtigt und ergibt insbesondere in Bezug auf weniger hochentwickelte Tiere Sinn. Darum ist Hyman zuzustimmen, wenn er Definitionen, die Aktivität mit willentlicher Aktivität gleichsetzen, als verfehlt bezeichnet:

> If we accept – as we certainly should – that action is not limited to animals capable of acting voluntarily, it is obvious that these definitions cannot be right. (Hyman 2011: 298)

## 2. Nicht-menschliche Akteure

Die Vorstellung von nicht-menschlichen Akteuren teilt Hyman mit anderen Philosophen, z.B. Anthony Kenny[1], Harry Frankfurt und Fred Dretske.

Kenny:

> Voluntary actions are a subclass of a very much wider genus. Agency is a universal phenomenon; and though it may be human agency which intrests us most, it is absurdly provincial to restrict the application of the concept to human beings or even to living beings. [...] Animal agency is undeniable: but animals are not the only non-human agents. The grass pushing its way between the crazy paving, the Venus' flytrap closing on its prey, the action of acqua regia on gold of hydrochloric acid on litmus paper – these are examples of action by non-conscious agents. (Kenny 1975: 46)

Frankfurt:

> There is a tendency among philosophers to dicuss the nature of action as though agency presupposes characteristics which cannot plausibly be attributed to members of species other than our own. [...] There are numerous agents besides ourselves, who may be active as well as passive with respect to the movement of their bodies. Consider the difference between what goes on when a spider moves its legs in making its way along the ground, and what goes on when its legs move in similar patterns and with similar effect because they are manipulated by a boy who managed to tie strings to them. In the first case the movements are not simply purposive, as the spider's digestive processes doubtless are. They are also attributable to the spider, who makes them. In the second case the same movements occur but they are not made by the spider, to whom they merely happen. (Frankfurt 1978: 162)

Dretske:

> Voluntary behavior, though, is only one species of behavior. What we are here concerned is a much more general notion, one that applies to animals, plants and perhaps even machines in very much the same way it applies to people. It applies to people, furthermore, when there are no purposes or intentions, those that allegedly qualify a system as an agent and its purposeful activity as voluntary. (Dretske 1988: 3)
>
> If the lowly cockroach doesn't have a mind, doesn't have purposes or intentions, and therefore doesn't exhibit what we think of as intentional behavior, this doesn't mean the poor creature doesn't do anything. (Dretske 1988: 4)
>
> The fact is that insects, worms, snails, crickets, leeches and even paramecia behave in a quite intressting way. They aren't stones, whose fate is completely at the mercy of external forces. If we ask why the activities (to use as neutral a word as possible) of even the simplest living creature are regarded as behavior by those who study them, the answer seems obvious. It is not because such movements are thought to be voluntary. It is not because it is thought that leeches and sponges have reasons – beliefs, desires, purposes or intentions – for doing the things they do. No, these activities are deemed behavior [...]: because the movement, these changes of state are internally produced. (Dretske 1988: 7)

[1] Auf Kenny weist Hyman explizit hin.

## 3. Kriterien für *agency*

Dass es nicht-menschliche Akteure gibt, darin sind sich die genannten Philosophen einig. Wer oder Was jedoch noch als Akteur in Frage kommt, darüber herrscht Uneinigkeit. Für Kenny und Hyman (Vgl. Alvarez/Hyman 1998: 243ff) sind es neben Tieren und Pflanzen sogar anorganische Substanzen, wie z.B. Salzsäure. Für Frankfurt hingegen sind im Club der Akteure nur Lebewesen zugelassen. Während Dretske neben Tieren und Pflanzen zumindest auch Maschinen als Akteure in Erwägung zieht. Steine sind aber für Dretske keine Akteure: Steine verhalten sich nicht, sie tun nichts.

Auf der einen Seite haben wir also Kenny und Alvarez/Hyman für die agency ein universelles Phänomen ist und auf der anderen Seite haben wir Frankfurt und Dretske die agency weitestgehend auf Lebendes beschränken. Diese Uneinigkeit ist in der Anwendung unterschiedlicher Kriterien für die Zuschreibung von agency zu vermuten.

### *3.1 Kenny und Alvarez/Hyman*

Was ist nun bei Kenny und Alvarez/Hyman das Kriterium für *agency*? Konzentrieren wir uns auf *agency* im unbelebten Bereich, da hier der Dissens zu Frankfurt und Dretske besteht. Kenny räumt ein, dass es in diesem Bereich oft sehr schwierig ist echte *agency* zu identifizieren und zu bestimmen, was hier ein Akteur ist und was nicht. Der wissenschaftliche Fortschritt mag auch gelegentlich unsere bisherige Bestimmung korrigieren, aber dass wir falsch lagen und nicht etwa etwas Unsinniges gesagt haben, zeigt zumindest, so Kenny, dass die Begriffe *agency* und Akteur auch im unbelebten Bereich eine berechtigte Anwendung finden (Vgl. Kenny 1975: 46). Kennys Begriff von *agency* ist eng mit dem Begriff *natural power* verbunden.

> Wherever we can talk of substances in nature, wherever we can talk of natural kinds, we can talk also of natural agency and natural powers. The concepts of agency and power are obviously connected: the natural actions of an agent are exercises of its natural powers. (Kenny 1975: 46)

Dasselbe gilt für Alvarez/Hyman: „inanimate things have causal powers whose exercise is agency." (Alvarez/Hyman 1998: 244) *Natural powers* sind so etwas wie Dispositionen, denn wenn bestimmte Bedingungen erfüllt sind, dann wird die *natural power* unfehlbar ausgeübt.

> Natural powers needed certain preconditions for their exercise: fire will burn wood only if wood is sufficiently dry. But if these conditions are met, then the power will infallibly exercised. (Kenny 1989: 66)

Das Feuer wäre in diesem Beispiel ein Akteur und das Holz wahrscheinlich ein Patient. Aber warum? Weil das Feuer sich nicht verändert, das Holz hingegen schon, es verbrennt? Hat Feuer eine *natural power*, weil es eine Veränderung des Holzes bewirken kann? Auch die Beispiele von Alvarez/Hyman legen eine solche Deutung nahe.

> oxygen rusts iron (Alvarez/Hyman 1998: 221)

Sauerstoff macht etwas mit Eisen, er ist ein Akteur, denn der Sauerstoff bleibt unverändert, während sich das Eisen ändert, es rostet. Sauerstoff hat eine *natural power*, denn es bewirkt unter bestimmten Bedingungen eine Veränderung des Eisens.

> a volume of acid [...] dissolves a lump of zinc (Alvarez/Hyman 1998: 245)

Die Säure macht etwas mit dem Zinkklumpen, die Säure ist ein Akteur, denn die Säure bleibt unverändert, während sich der Zinkklumpen ändert, er löst sich auf. Die Säure hat eine *natural power*, denn sie bewirkt unter bestimmten Bedingungen eine Veränderung des Zinkklumpens. Vielleicht könnte man das Kriterium allgemein so formulieren:

> Kommen zwei Substanzen A und B unter bestimmten Bedingungen C zusammen und verändert sich B daraufhin sichtbar, dann ist A ein Akteur und B ein Patient. A hat die *natural power* eine Veränderung von B zu bewirken. Die Ausübung dieser *natural power* von A unter C ist *agency*.

Wie überzeugend ist dieses Kriterium für *agency*? Vor allem wie verhält es sich zur aktiv/passiv-Unterscheidung? Schauen wir uns die Beispiele noch einmal genauer an. Bei allen Beispielen findet eine chemische Reaktion statt, die sich bei den letzten beiden Beispielen so darstellen lässt:

$$4\,Fe + 3\,O_2 + 6\,H_2O \longrightarrow 4\,Fe(OH)_3\ [= \text{Eisenrost}]$$

$$Zn + 2\,HCl \longrightarrow ZnCl_2 + H_2$$

Ist man jetzt immer noch geneigt zu sagen, dass der Sauerstoff und die Säure aktiv sind, während das Eisen und das Zink passiv seien? Würde man nicht viel eher sagen, dass Sauerstoff, Eisen und Wasser bzw. Zink und Salzsäure miteinander reagieren und dabei Reaktionsprodukte (Eisenrost bzw. Zinkchlorid und Wasserstoff) entstehen? Unter der Annahme, dass die tatsächlichen Vorgänge durch die chemischen Reaktionsformeln angemessener wiedergegeben werden als durch eine alltagssprachliche Beschreibung unserer Beobachtung, neige ich sehr zu letzterer Auffassung. Es ist eben, m.E., auch nicht so, wie Kenny meint, dass wir bei genauerer Untersuchung der Phänomene festgestellt hätten, dass wir falsch liegen, dass, sagen wir, bei der Eisenkorrosion Eisen den aktiven Part hat und Sauerstoff passiv ist. Vielmehr erscheint die Aktiv-Passiv-Unterscheidung bei miteinander reagierenden Substanzen unangemessen. Insgesamt scheint die Auffassung von Kenny und Alvarez/Hyman[2] auf einer an Aristoteles orientierten animistischen Konzeption von Natur zu basieren, die nicht mehr haltbar ist.

### *3.2 Dretske*

Welches Kriterium hat nun Dretske für *agency*?[3] Zunächst ist festzuhalten, dass Dretske nicht von *agency*, sondern von Verhalten spricht. Verhalten ist das, was ein Organismus tut, im Gegensatz zu dem, was mit ihm geschieht, was er erleidet. Dretskes Auffassung von Verhalten liegt also die vertraute Unterscheidung zwischen aktiv und passiv zu Grunde. Und wie die oben angeführten Zitate zeigen, ist willentliches Verhalten für ihn nur eine Subspezies von Verhalten.

Der Unterschied zwischen dem Verhalten eines Organismus (die Ratte bewegt ihren Schwanz) und dem, was einem Organismus geschieht (ich bewege den Schwanz der Ratte) ist für Dretske durch *die Lokalisierung der Bewegungsursache* charakterisiert. Beim Verhalten liegt die Ursache im Organismus (internal cause), bei dem, was dem Organismus geschieht, außerhalb (external cause)[4].

---

[2] Alvarez/Hyman betonen ausdrücklich, dass dies nicht der Fall sei (Vgl. Alvarez/Hyman 1998: 245).

[3] Ich beschränke mich hier auf eine Darstellung von Dretskes Kriterium, da dessen Auffassung in *Explaining Behavior* sehr detailliert ausgearbeitet ist.

[4] Ein möglicher Schwachpunkt von Dretskes Auffassung, auf den ich hier nicht weiter eingehen will, könnte darin bestehen, dass die Bestimmung von Verhalten davon abhängt, dass man das System vom Nicht-System (Umwelt) abgrenzen kann. Als Systemgrenze könnte zunächst einfach die Membran gelten, die den Organismus umgibt. Dretske weist aber in einer Fußnote daraufhin, dass „internal“ nicht einfach nur „innen drinnen (inside)“ oder „unter der Haut (beneath the skin)“ heißen soll, ohne jedoch ein anderes Kriterium zu liefern, wodurch sich Systemzugehörigkeit bestimmen ließe (Vgl. Dretske 1988: 3).

> [...] the suggestion is that behavior is endogenously produced movement, movement that has its causal origin within the system whose parts are moving. (Dretske 1988: 2)

Dretskes Auffassung scheint auch van Inwagen zu befürworten und er grenzt sie deutlich gegen die Auffassung von Kenny und Alvarez/Hyman ab:

> The concept of causal power or capacity would seem to be the concept of invariable disposition to react to certain determinate changes in the environment in certain determinte ways, whereas the concept of an agent's power to act would seem not to be the concept of a power that is dispositional or reactive, but rather the concept of a power to *originate* changes in the environemt. (van Inwagen 1983: 11)

Die Idee einer Produktion von Bewegung oder Veränderung allein aus dem System heraus ist jedoch problematisch. Hyman demonstriert dies am Beispiel von John Locke[5]. Dessen Vorstellung zufolge produziert eine Billardkugel, die eine andere anstößt, keine Bewegung, sondern sie überträgt diese lediglich, denn sie verliert genauso viel Bewegung, wie sie weitergibt.

> For when bodies interact, motion is communicated, but it is not produced; and action, Locke insists, is the production of motion, or some other kind of changes. (Hyman 2011: 301)

Wenn wir nun etwas allgemeiner formulieren, können wir sagen, dass die erste Billardkugel keine Veränderung produziert, weil sie nur genauso viel Energie weitergibt, wie sie ihrerseits verliert. Jetzt wird das Problem deutlich, denn eine echte Produktion von Bewegung müsste dem Energieerhaltungssatz der Physik widersprechen.

> This is the crux of the matter. Locke denies that the first ball produces motion in the second ball because it ‚loses in itself so much, as other received', in other words, because the interaction between the balls conserves the total quantity of motion. But it follows that he can only acknowledge that the production of motion – in other words, action – has occured if the total quantity of motion is not conserved, but increased. An action must therefore be a breach or an exeption to the law of nature. In other words, it must be a miracle, an interference in the natural course of events by a being with strictly supernatural abilitiy to inject motion into the natural world, rather than transferring it to something else. (Hyman 2011: 302)

Zumindest Dretske erkennt, dass „internally produced" nicht heißen kann, dass Organismen einfach Kausalketten starten können und somit Energie erzeugen anstatt sie lediglich umzuwandeln.

> Even if every event has, for any given time, some unique cause, internal (and external) causes themselves have causes. Hence, by tracing the causal sequence far enough back in time, one can, sooner or later, find external causes for every change or bodily movement. (Dretske 1988: 22)

Daraus folgt aber, dass ich nun kein Kriterium mehr habe um festzustellen, ob einem Organismus gerade etwas geschieht oder ob er etwas tut. Angenommen ich beobachte wie sich ein Hase beim Auftauchen eines Raubvogels an den Boden drückt. Je nachdem, ob ich nun als Ursache der Drückbewegung das Auftauchen des Raubvogels oder aber z.B. bestimmte neuronale Aktivitäten ansehe, klassifiziere ich das Beobachtete entweder als etwas, das dem Hasen geschieht oder aber als Hasenverhalten, etwas, das der Hase tut. Dem stimmt Dretske in vollem Umfang zu:

---

[5] Hyman geht es hierbei vor allem um eine philosophiegeschichtliche Erklärung, warum Aktivität fast immer mit willentlicher Aktivität gleichgesetzt wurde.

> Unless there is a principled way of saying which causal factor is to be taken the cause of movement or orientation, the present system of classification provides no principled way of saying whether the cat is doing anything. It gives us no telling of what is, and what isn't, behavior. (Dretske 1988: 23f)

Das ist irgendwie ernüchternd. Und es hilft dann auch wenig, wenn Dretske uns versichert, dass das alles gar keine Rolle spielt, zumindest für ihn nicht (Vgl. Dretske 1988: 24). Für ihn ist allein wichtig, dass wenn etwas als Verhalten von S klassifiziert wird, dass dann die bestimmten Bewegungen von S zugleich auch als das Resultat von Ereignissen, die in S stattfanden, klassifiziert werden (Vgl. ebd.: 24f). Ob aber etwas als Verhalten von S klassifiziert wird oder ob nicht, das ist nach Dretske völlig von den Interessen und Zwecken des Betrachters abhängig. Ein derartig subjektives Kriterium ist natürlich wenig hilfreich.

## 4. Fazit und Ausblick

Hymans Behauptung, dass man einerseits zwischen aktiv und passiv und anderseits zwischen willentlich und nichtwillentlich unterscheiden müsse, ist nach wie vor plausibel. Auch die Annahme von nichtwillentlicher Aktivität oder nichtwillentlichen Verhalten z.B. bei niederen Lebewesen, wie Bakterien, im Gegensatz zu dem, was diesen lediglich geschieht, ist m.E. einleuchtend. Allerdings hat die Diskussion gezeigt, dass es ziemlich schwierig ist anzugeben, auf welchen Kriterien die Unterscheidung zwischen aktiv und passiv beruhen soll, ohne doch wieder auf die willentlich/nichtwillentlich Unterscheidung zurückzugreifen. Keines der beiden hier untersuchten Kriterien konnte diesbezüglich überzeugen.

Um nicht mit diesem rein negativen Ergebnis zu schließen, möchte ich zwei weitere Ansätze zur Bestimmung nichtwillentlicher Aktivität andeuten, die sich bei den hier untersuchten Autoren finden lassen. Beginnen wir mit Dretske, bei dem wir hinsichtlich der Klassifikation von Reflexen als Verhalten folgende Überlegung finden:

> [...] we often classify reflexes as behavior. We do so because the reaction to a stimulus, although perfectly reliable, is quite unlike the body's Newtonian response to a shove (where acceleration is proportional to net impressed force). The reflexive behavior exhibits a change in form, direction or magnitude. [...] As Sherington (1906, p. 5) observed, the stimulus acts like a ‚releasing force' on the organism in the sense that the energy expended in the response far exceeds the energy provided by the eliciting stimulus. (Dretske 1988: 26)

Der Grund dafür, dass Reflexe doch Verhalten sind, obwohl die Ursache klarerweise extern ist, soll darin liegen, dass die Energie des Reizes bei weitem nicht ausreicht um die ausgelöste Bewegung allein zu verursachen. Im Fall der oben erwähnten Billardkugeln stammt die Energie zur Bewegung der zweiten Kugel hingegen vollständig von der kinetischen Energie der ersten Kugel. Es besteht hier also ein interessanter Unterschied zwischen den beiden Fällen. Um diesen Unterschied genauer zu erfassen, können wir mit Norbert Bischof zwischen Arbeitskausalität und Steuerkausalität unterscheiden:

> Arbeitskausalität = eine naturgesetzlich garantierte Beziehung, bei der die Ursache selbst die für die Wirkung erforderliche Energie liefert. (Bischof 2009: 147)

> Steuerkausalität = eine Beziehung, bei der die Energiebilanzen von Ursache und Wirkung getrennt bleiben und die Einflussnahme lediglich durch die Struktur des Systems gewährleistet wird. (Ebd.)

Wäre es möglich, dass es einen engen Zusammenhang zwischen nichtwillentlicher Aktivität (Verhalten) und Steuerkausalität gibt? Etwa: Ohne Steuerkausalität, kein Verhalten, keine Aktivität. Ein solcher Zusammenhang wäre zwar noch kein Kriterium, aber schon mal ein Anfang. Zunächst scheint aber der Begriff der Steuerkausalität erläuterungsbedürftig. Der

Begriff „Steuerkausalität“ stammt von Hassenstein und wird von ihm unter Erläuterung des physikalischen Prinzips des Steuerns eingeführt.

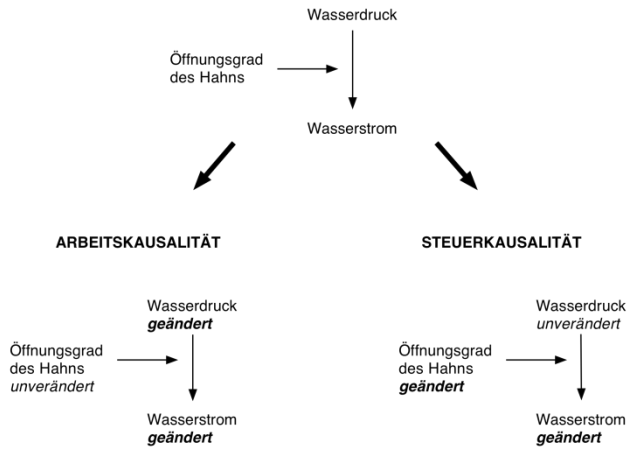
> Nach dem Kausalprinzip ist jedes Ereignis durch eine Konstellation von Umständen eindeutig determiniert, wobei zu dieser Konstellation nicht nur energieliefernde oder –entziehende Umstände, sondern auch solche gehören können, die dem Vorgang weder Energie liefern noch entziehen. *Steuern heißt Ändern bzw. Kontrollieren dieser nicht durch Energieübertragung wirksamen Kausalbedingungen für den gesteuerten Vorgang*. (Hassenstein 1960: 349)

Dies können wir uns am Beispiel eines Wasserhahns klar machen:

> Die Stärke des Wasserstroms, der aus dem Hahn ausfließt, wird im wesentlichen durch zwei Bedingungen bestimmt, den Wasserdruck und den Öffnungsgrad des Hahns. Der Wasserdruck ist Ausdruck der potentiellen Energie, welche die Quelle kinetischen Energie des Wasserstroms ist. Der Öffnungsgrad wirkt dagegen auf die Stärke des Wasserstroms ein, ohne diesem Energie zu liefern oder zu entziehen. Natürlich ist das *Ändern* des Öffnungsgrades ein Prozeß, bei dem Energie umgesetzt wird. Doch geht auch von dieser Energie nichts an den Wasserstrom über. (Hassenstein 1960: 349)

Eine Veränderung der Stärke des Wasserstromes lässt sich also prinzipiell auf zweierlei Weise erreichen, erstens durch eine Veränderung des Wasserdrucks oder zweitens durch eine Veränderung des Öffnungsgrades des Hahns (siehe Abbildung unten). Nur die zweite Veränderung ist ein Steuern und nur den zweiten Kausalzusammenhang bezeichnet Hassenstein als Steuerkausalität.

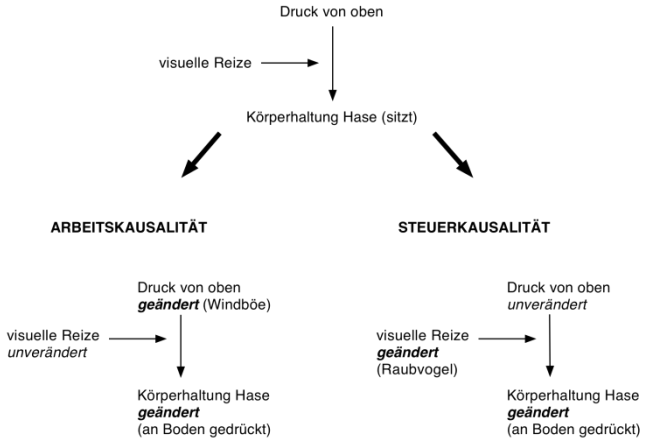
> Der eigenartige Kausalzusammenhang zwischen einem steuernden und einem gesteuerten Vorgang, der die beiden ohne Energieaustausch miteinander verbindet, soll im folgenden kurz als ‚Steuerkausalität’ bezeichnet werden. (Hassenstein 1960: 350)




Wir können dieses Schema auf unser Beispiel mit dem Hasen, der sich in Gegenwart eines Raubvogels an den Boden drückt, übertragen[6]. Nun können wir erklären, warum es sich bei der Drückbewegung des Hasen in Gegenwart eines Raubvogels um ein Verhalten (eine Aktivität) handelt, während dieselbe Drückbewegung, verursacht durch eine starke Windböe, kein Verhalten (keine Aktivität) ist, und zwar obwohl in beide Fällen die Ursache extern ist. Denn die Photonen, die der Raubvogel reflektiert, liefern nicht die Energie für die Drückbewegung, diese stammt aus dem Metabolismus des Hasen. Die Photonen ändern nur

[6] Selbstverständlich kann eine solche Übertragung nur partiell gelingen. Es bestehen natürlich auch wichtige Unterschiede zwischen den beiden Fällen.

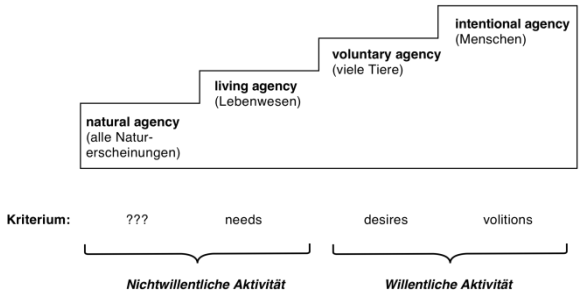
gewisse Randbedingungen (neuronale Zustände des Hasen), so dass die vom Metabolismus erzeugte und teilweise gespeicherte Energie die Bewegung erzeugt. Im Fall der Windböe hingegen stammt die Energie für die Drückbewegung ausschließlich aus der Böe selbst.




Eine ähnliche Erläuterung wäre auch für Frankfurts Spinnen- und Dretskes Rattenbeispiel (siehe Zitate oben) vorstellbar und in keinem dieser Fälle müsste man die von Hyman kritisierte merkwürdige Verletzung des Energieerhaltungssatzes annehmen.

Könnte das Vorfinden von Steuerkausalität also ein Kriterium für die Zuschreibung nichtwillentlicher Aktivität sein? Wenn dem so wäre, dann müssten wir auch dem Wasserhahn eine nichtwillentliche Aktivität zuschreiben, denn das Öffnen des Hahns ist ein Fall von Steuerkausalität. Eine solche Zuschreibung scheint mir aber falsch, denn intuitiv ist hier derjenige aktiv, der den Hahn öffnet, und nicht der Hahn selbst[7]. Wir können also festhalten, dass nichtwillentliche Aktivität vermutlich etwas mit Steuerkausalität zu tun hat. Das Vorfinden von Steuerkausalität allein ist aber wahrscheinlich nicht hinreichend für die Zuschreibung nichtwillentlicher Aktivität.

Wenden wir uns nun noch einmal der Position von Kenny zu. Auch hier haben wir m.E. noch nicht alle Ressourcen zur Bestimmung nichtwillentlicher Aktivität genügend berücksichtigt. Bisher haben wir nur Kennys Auffassung zur nichtwillentlichen Aktivität anorganischer Substanzen kennengelernt und verworfen. Nun vertritt Kenny aber insgesamt so etwas wie eine Hierarchie der *agency's* (siehe Abbildung), die aus folgenden Rangstufen besteht: Die höchste Stufe bildet die *intentional agency*, gefolgt von der *voluntary agency*, dann kommt die *living agency* und zu unterst die bereits kennengelernte *natural agency* (Vgl. Kenny 1989: 38).




[7] Wenn wir ein Thermostat betrachten, wird diese Intuition schon etwas schwächer. Der Thermostat steuert im oben genannten Sinn einen Energiestrom, der eine bestimmte Raumtemperatur verursacht und ein Abweichen von dieser Raumtemperatur wiederum bewirkt die Änderung des Thermostaten.

Die beiden obersten Stufen gehören in den Bereich willentlicher Aktivität, der uns hier nicht interessiert. Für uns sind nur die unteren beiden Stufen wichtig, denn beide gehören in den Bereich nichtwillentlicher Aktivität. Die unterste haben wir bereits betrachtet und wenden uns daher der *living agency* zu. Diese besitzt im Gegensatz zur *natural agency* ein mehr oder weniger explizites Kriterium. Wir können nämlich zwischen dem, was einem Organismus geschieht, und dem, was er tut, unterscheiden, indem wir die Bedürfnisse (*needs*) des Organismus berücksichtigen.

> Living Things, unlike non-living natural agents, have needs. Plants need a particular kind of soil if they are to thrive, flowers need water if they are not to die. (Kenny 1989: 34)

> Indeed, when it comes to living things, the active/passive distinction takes a new dimension. For here we can distinguish between what an organism does and what happens to it by reference to the organism's *needs*. Thus plants fulfil their needs by growing roots, orienting their leaves, emitting chemical compounds to deter predators, etc. (Glock 2012: 901)

Nun stellt sich aber die Frage, wodurch die Bedürfnisse eines Organismus bestimmt sind. Stellen wir uns vor, wir beobachten zwei Veränderungen einer Pflanze, von denen wir die eine intuitiv als aktiv und die andere als passiv klassifizieren würden. Einige Insekten befallen eine Pflanze, daraufhin finden zwei Veränderungen statt: 1) es werden chemische Substanzen von der Pflanze abgesondert, die die Freßfeinde der Insekten anlocken und 2) es sterben einige Blätter der Pflanze ab. Die erste Veränderung scheint aktiv die zweite passiv. Aber warum? Auf der Ebene des Molekulargeschehens kann man vermutlich keinen wesentlichen Unterschied erkennen. Es tauchen jedenfalls bei der ersten Veränderung nicht etwa Bedürfnisse in Form von kausal wirksamen Komponenten auf, die bei der zweiten Veränderung fehlen. Worauf basiert dann unsere Unterscheidung? Wir könnten sagen, dass die erste aber nicht die zweite Veränderung zur Selbsterhaltung beiträgt oder wir sagen, dass nur die erste aber nicht die zweite Veränderung einen Selektionsvorteil bietet[8]. Bei beiden Antworten spielen aber m.E. strukturelle Alternativen zur betrachteten Pflanze eine wesentliche Rolle. D.h. Varianten der betrachteten Pflanze, die eine ähnliche aber dennoch verschiedene Struktur haben, so wie die Individuen innerhalb einer biologischen Art variieren[9]. Im ersten Fall kontrastieren wir die Pflanze mit einer strukturellen Alternative, die keine Chemikalien absondert. Während wir im zweiten Fall die Pflanze mit einer Alternative kontrastieren deren Blätter trotz Insektenbefall nicht absterben. D.h. im ersten Fall ist betrachtete Pflanze besser[10] als ihre strukturelle Alternative und gilt deshalb als aktiv, während sie im zweiten Fall schlechter als ihre Alternative ist und daher als passiv gilt.

Was passiert aber, wenn wir herausfinden, dass die Absonderung der Chemikalien alternativlos ist, dass jede Pflanze, die von diesen Insekten befallen wird, entsprechende Chemikalien absondert? Oder wenn wir feststellen, dass alle anderen Pflanzen (strukturellen Alternativen) bei einem gleichartigen Insektenbefall (Spezies, Intensität, Dauer etc.) viel

---

[8] Für Kenny haben *needs* ebenfalls etwas mit Überleben bzw. Fitness zu tun. „Need is a very important concept comparatively little studied by philosophers; and I can not pretend to have an adequate account of it. As a first approximation we may say that an agent A, has at a given time *t* a need for something X if A lacks x at *t*, and A cannot continue to survive (or survive as a good specimen of ist species) unless A comes to possess X." (Kenny 1975: 48) Ich würde die zweite Alternative (Selektionsvorteil) vorziehen. Erstens ist der Begriff der Selbsterhaltung ziemlich vage und zweitens müssten Veränderungen, die die Reproduktion betreffen, aufgrund ihres oft sehr hohen Energieaufwandes unter der Perspektive der Selbsterhaltung als passiv gelten.

[9] Eine genaue Bestimmung von „strukturelle Alternative" ist mir derzeit nicht möglich. Der Begriff der biologischen Art ist hierbei wenig hilfreich, weil dessen Bestimmung ähnlich problematisch ist.

[10] „Besser" heißt hier entweder „höhere Wahrscheinlichkeit zur Selbsterhaltung" oder „hat bei sonst gleichen Eigenschaften einen höheren Selektionswert".

mehr Blätter verlieren und zudem Wurzeln und Zweige absterben? Würden wir dann nicht zu einer ganz anderen Einschätzung kommen? Nämlich, dass die erste Veränderung passiv ist und die zweite aktiv.

Wenn dies stimmen sollte, dann wäre die Aktiv-Passiv-Unterscheidung zumindest im Bereich von nichtwillentlichen Veränderungen *relativ*. Wir könnten nicht einfach erkennen, ob eine Veränderung passiv oder aktiv ist, solange wir nicht wissen mit welchen Alternativen wir den Träger der Veränderung vergleichen müssten. Nun könnte man behaupten, dass wir diese Unterscheidung trotzdem anwenden, zeigt, dass der Unterscheidung etwas anderes zugrunde liegen müsse. Gerade in Bezug auf Pflanzen oder Einzeller scheinen wir aber m.E. gar nicht so sicher in der Anwendung der Aktiv-Passiv-Unterscheidung zu sein, so dass dieser Einwand zumindest an Stärke verliert. Zudem stellt sich die Frage, ob das Kriterium, das wir tatsächlich anwenden – welches dies auch immer sein mag – berechtigt ist. Mit Hilfe der vorangegangenen Überlegungen könnten wir immerhin ein Kriterium zur Bestimmung von nichtwillentlicher Aktivität entwickeln, das in etwa so lauten könnte:

> Die Veränderung eines Systems S gilt als aktiv, wenn die Veränderungen der strukturellen Alternativen von S, die zu einem früheren Zeitpunkt existierten, einen geringeren Selektionswert gehabt haben als die von S.

Dies scheint mir zumindest ein attraktiver Ansatz, der weiter ausgearbeitet werden sollte. Es wäre m.E. vor allem interessant herauszufinden, ob man in diesen an der Evolutionstheorie orientierten Ansatz das Konzept der Steuerkausalität sinnvoll integrieren kann.

**André Wunder**

Universität Zürich
andre.wunder@philos.uzh.ch

## Literatur

Alvarez, Maria und John Hyman. 1998: „Agents and their Actions“, *Philosophy* 73, 219-245.

Bischof, Norbert. 2009: *Psychologie. Ein Grundkurs für Anspruchsvolle*. Stuttgart: Kohlhammer.

Dretske, Fred. 1988: *Explaining Behavior. Reasons in a World of Causes*. Cambridge (MA): MIT Press.

Frankfurt, Harry. 1978: „The Problem of Action“, *American Philosophical Quarterly* 15, 157-162.

Glock, Hans-Johann. 2012: „Animals: Agency, Reasons and Reasoning“, in J. Nida-Rümelin and E. Özmen (Hrg.): *Welt der Gründe. Proceedings XXII. Deutscher Kongress für Philosophie*. Hamburg: Meiner. 900-913.

Hassenstein, Bernhard. 1960: „Die bisherige Rolle der Kybernetik in der biologischen Forschung“, Naturwissenschaftliche Rundschau 13, 349-355.

Hyman, John. 2011: „Wittgenstein on Action and the Will“, *Grazer Philosophische Studien* 82, 285-311.

Kenny, Anthony J. P. 1975. *Will, Freedom and Power*. Oxford: Basil Blackwell.

— 1989. *The Metaphysics of Mind*. Oxford: Clarendon.

van Inwagen, Peter. 1983: *An Essay on Free Will*. Oxford: Clarendon Press.

# 5. Erkenntnistheorie

# Explanatorisches Verstehen: Ein Definitionsvorschlag

Christoph Baumberger

In diesem Aufsatz entwickle ich eine Definition von explanatorischem Verstehen, indem ich dieses mit Wissen vergleiche. Erstens zeige ich, inwiefern explanatorisches Verstehen das Erfassen einer anspruchsvolleren Erklärung und eine anspruchsvollere Rechtfertigung verlangt als explanatorisches Wissen. Zweitens argumentiere ich dafür, dass die Erklärung den Tatsachen gerecht werden muss, explanatorisches Verstehen aber im Gegensatz zu Wissen nicht immer faktiv ist. Drittens verteidige ich die Auffassung, dass explanatorisches Verstehen, anders als Wissen, mit epistemischem Glück kompatibel ist. Als Ergebnis schlage ich vor, dass S (in einem gewissen Ausmaß) versteht, warum p, gdw. S jede der folgenden Bedingungen (in einem gewissen Ausmaß) erfüllt: (a) S legt sich auf eine Erklärung E fest, die aus dem Explanandum p und einem Explanans besteht, das zeigt, wie p von q abhängt, (b) S erfasst E, (c) S ist fähig, E zu rechtfertigen, und (d) E wird den Tatsachen gerecht. Die Bedingungen (b) und (c) verlangen bestimmte Fähigkeiten: das Erfassen einer Erklärung die Fähigkeit, von ihr Gebrauch zu machen, und das Rechtfertigen einer Erklärung die Fähigkeit, zeigen zu können, dass sie die beste verfügbare und eine hinreichend gute Erklärung ist. Ich zeige, wie die Definition für objektuales Verstehen adaptiert werden kann, und argumentiere gegen eine Reduktion des explanatorischen Verstehens auf propositionales Verstehen.

## 1. Einleitung

Die Erkenntnistheorie wird in der Regel als Theorie des Wissens charakterisiert, wobei Wissen mit propositionalem Wissen identifiziert wird. In den letzten Jahren ist dieser Fokus auf Wissen in Frage gestellt und zunehmend die Auffassung vertreten worden, dass vielmehr Verstehen unser primäres kognitives Ziel ist. Es ist dafür argumentiert worden, dass eine solche Auffassung es ermöglicht, den Wissenschaften gerecht zu werden (Elgin 1996; 2007), intellektuelle Tugenden zu identifizieren (Zagzebski 2001; Riggs 2003), das Wertproblem für Wissen zu vermeiden (Kvanvig 2003; Pritchard 2010) und die Moral gegen den Egoisten zu verteidigen (Hills 2010). Während diese Rollen für den Verstehensbegriff ausführlich diskutiert wurden, gibt es bisher in der Erkenntnistheorie kaum Versuche, ihn zu definieren.

Man kann zumindest drei Verstehenstypen unterscheiden. Auch wenn die Umgangssprache nicht immer ein verlässlicher Führer ist, wird in der Regel je nachdem, ob die Zuschreibung einen dass-Satz, einen indirekten Fragesatz oder eine Nominalphrase verwendet, zwischen propositionalem, interrogativem und objektualem Verstehen unterschieden (Grimm 2011: 84–88). In diesem Aufsatz schlage ich eine Definition von explanatorischem Verstehen vor, das oft als die wichtigste Form interrogativen Verstehens angesehen wird. Explanatorisches Verstehen ist in erster Näherung ein Verstehen, warum etwas der Fall ist, anhand einer Erklärung, welche die Warum-Frage beantwortet. Ein solches Verstehen muss nicht über einen indirekten Fragesatz zugeschrieben werden; man kann dazu auch eine Nominalphrase verwenden, zum Beispiel, wenn man sagt, jemand verstehe die Ursache von etwas.

Ich entwickle meine Definition, indem ich explanatorisches Verstehen mit Wissen vergleiche. In Abschnitt 2 zeige ich, dass man wissen kann, dass und selbst warum etwas der Fall ist, ohne zu verstehen, warum es der Fall ist, explanatorisches Verstehen also weder mit propositionalem noch mit explanatorischem Wissen äquivalent ist. Für meine Definition resultiert eine Erfassensbedingung und eine Rechtfertigungsbedingung. In Abschnitt 3 zeige

ich, dass man verstehen kann, warum etwas der Fall ist, ohne zu wissen, warum es der Fall ist, explanatorisches Verstehen also nicht einmal eine Art von Wissen ist. Die Überlegungen führen zu einer Bedingung der Tatsachentreue und einer Festlegungsbedingung; ich argumentiere zudem gegen eine zusätzliche Anti-Glück-Bedingung. In Abschnitt 4 präsentiere ich meine Definition und zeige, wie sie für objektuales Verstehen adaptiert werden kann. Abschließend (Abschnitt 5) argumentiere ich gegen eine Reduktion des explanatorischen Verstehens auf propositionales Verstehen.

## 2. Wissen ohne explanatorisches Verstehen

Es ist offensichtlich möglich zu wissen, dass etwas der Fall ist, ohne zu verstehen, warum es der Fall ist. Ich kann zum Beispiel wissen, dass die globale Mitteltemperatur seit Mitte des 20. Jahrhunderts deutlich angestiegen ist, ohne zu verstehen, warum dem so ist. Fälle von Hörensagen zeigen, dass es sogar möglich ist zu wissen, warum etwas der Fall ist, ohne zu verstehen, warum es der Fall ist (Pritchard 2010: 81; Hills 2010: 192). Nehmen wir an, ein Klimawissenschaftler erklärt mir, dass die globale Mitteltemperatur angestiegen ist, weil die Treibhausgaskonzentrationen in der Atmosphäre zugenommen haben. Wenn er Recht hat und ich gute Gründe für seine Verlässlichkeit habe, weiß ich, warum die globale Mitteltemperatur angestiegen ist. Aber ich verstehe nicht, warum dem so ist, solange ich keinerlei Auffassung davon habe, wie erhöhte Treibhausgaskonzentrationen die globale Erwärmung verursachen können, und der einzige Grund, den ich für die Erklärung geben kann, darin besteht, dass ich sie von einem verlässlichen Experten habe.

Dieses Beispiel legt zwei Unterschiede zwischen explanatorischem Verstehen und explanatorischem Wissen nahe. Erstens, während explanatorisches Wissen eine korrekte Überzeugung darüber beinhaltet, was die Ursachen oder Gründe für etwas sind, verlangt explanatorisches Verstehen zusätzlich eine mehr oder weniger korrekte Auffassung ihrer Beziehung zu dem, wofür sie Ursachen oder Gründe sind. Wenn q die Ursache oder der Grund für p ist, dann gilt: Wenn ich weiß, warum p der Fall ist, dann habe ich die korrekte Überzeugung, dass p der Fall ist, weil q der Fall ist. Wenn ich verstehe, warum p der Fall ist, dann bin ich nicht nur überzeugt, dass q die Ursache oder der Grund für p ist, sondern habe ich auch eine mehr oder weniger korrekte Auffassung davon, wie p von q verursacht oder begründet wird und erfasse damit eine anspruchsvollere Erklärung als für explanatorisches Wissen nötig ist. Dies führt zur Erfassensbedingung (Abschnitt 2.1). Zweitens verlangt explanatorisches Verstehen eine anspruchsvollere Rechtfertigung als explanatorisches Wissen. Wenn ich aufgrund von Hörensagen weiß, dass p der Fall ist, weil q der Fall ist, mag der einzige Grund, den ich für meine Überzeugung angeben kann, darin liegen, dass ein verlässlicher Experte mir das gesagt hat. Wenn ich verstehe, warum p der Fall ist, habe ich reflexiv zugängliche Gründe für die Erklärung und bin daher in der Lage, sie zu rechtfertigen, indem ich diese Gründe angebe. Dies führt zur Rechtfertigungsbedingung (Abschnitt 2.2).

### *2.1 Erfassensbedingung*

In welchem Sinn muss die Erklärung, die ein Verstehenssubjekt erfasst, anspruchsvoller sein als jene, von der ein Wissenssubjekt überzeugt ist? Um diese Frage zu beantworten, betrachte ich erst kausale Erklärungen und weite meine Überlegungen dann auf nicht-kausale Erklärungen aus. Schließlich zeige ich, worin das Erfassen einer Erklärung besteht.

#### *2.1.1 Kausale Erklärungen*

Eine Erklärung der Form, dass p der Fall ist, weil q der Fall ist, zeigt, von welchen Faktoren p abhängt. Eine Erklärung, die anspruchsvoll genug ist für explanatorisches Verstehen, muss zudem zeigen, wie p von den spezifizierten Faktoren abhängt. Eine solche Erklärung der

globalen Erwärmung zeigt (zumindest in qualitativen Begriffen), wie diese von steigenden Treibhausgaskonzentrationen abhängt. Abhängigkeiten dieser Art werden typischerweise durch Generalisierungen erfasst.[1] In einfachen kausalen Fällen umfasst eine hinreichend anspruchsvolle Erklärung damit ein Explanandum p und ein Explanans, das aus einer Anfangs- oder Randbedingung q besteht, die angibt, von welchen Faktoren p abhängt, und einer Generalisierung G, die angibt, wie p von den in q spezifizierten Faktoren abhängt.

Eine Generalisierung hat nur dann Erklärungskraft, wenn sie kontrafaktische Konditionale stützt, die beschreiben, wie p sich ändern würde, wenn einige der in q erwähnten Faktoren in verschiedener Hinsicht anders wären. In unserem Beispiel beschreiben sie, wie die globale Mitteltemperatur sich mit veränderten Treibhausgaskonzentrationen ändern würde. Solche Generalisierungen erlauben damit die Beantwortung dessen, was James Woodward „What-if-things-had-been-different questions" nennt (kurz: „Was-wäre-wenn-Fragen"). Das sind Fragen dazu, welchen Unterschied es für das Explanandum machen würde, wenn einige der im Explanans erwähnten Faktoren in verschiedener Hinsicht anders wären (Woodward 2003: 11). Es ist jedoch wohlbekannt, dass nicht jede Generalisierung, die kontrafaktische Konditionale stützt, eine Kausalbeziehung beschreibt; sie kann beispielsweise auch bloß zwei Wirkungen einer gemeinsamen Ursache zueinander in Beziehung setzen. Um eine Kausalbeziehung zu beschreiben muss eine Generalisierung nach dem vielversprechenden Vorschlag von Woodward zudem invariant sein unter einer Menge von Interventionen an den in q spezifizierten Faktoren (vgl. Woodward 2003: 14–16; Kap. 3).

Nun kann man aber im Besitz einer kausalen Erklärung sein, die eine invariante Generalisierung enthält, welche kontrafaktische Konditionale stützt, die darüber Auskunft geben, was unter Interventionen geschehen würde, und dennoch nicht verstehen (sondern nur wissen), warum das Explanandum-Ereignis eintritt. Das ist dann der Fall, wenn man keinerlei Auffassung davon hat, wie die spezifizierte Ursache die fragliche Wirkung hervorbringen kann. Das war eine der Leitideen meines Einstiegsbeispiels. Ich kann wissen, dass die globale Erwärmung durch steigende Treibhausgaskonzentrationen verursacht wird und selbst wissen, wie Treibhausgaskonzentrationen und die globale Erwärmung korreliert sind, ohne irgendeine Auffassung des zugrundeliegenden kausalen Mechanismus des Treibhauseffekts zu haben. Eine solche Auffassung mag nicht notwendig sein für eine Erklärung der globalen Erwärmung; aber sie scheint notwendig für ein explanatorisches Verständnis derselben. Die Generalisierung G muss deshalb Informationen über den zugrundeliegenden Kausalmechanismus enthalten, der p und q verknüpft.

### *2.1.2 Nicht-kausale Erklärungen*

Nicht jede Erklärung ist eine Kausalerklärung. Offensichtliche Beispiele nicht-kausaler Erklärungen einzelner Tatsachen oder Ereignisse finden sich in den Bereichen der Moral und der Ästhetik; zum Beispiel, wenn wir erklären, warum eine Ungleichverteilung von sozialen oder ökonomischen Gütern gerecht ist, indem wir zeigen, dass die am schlechtesten gestellten Mitglieder der Gesellschaft von ihr profitieren; oder wenn wir erklären, warum ein Gemälde schön ist, indem wir darauf hinweisen, dass seine Komposition harmonisch ist. Aber auch die Physik kennt nicht-kausale Erklärungen; zum Beispiel wenn erklärt wird, warum ein Gas die Temperatur t hat, indem darauf hinwiesen wird, dass die es konstituierenden Moleküle die mittlere kinetische Energie m haben. Eine solche Erklärung kann keine Kausalerklärung sein, da die Temperatur t nicht durch die mittlere kinetische Energie m verursacht werden kann, wenn Temperatur = mittlere kinetische Energie (Ruben 1990: 219). Natürlich gibt es auch nicht-kausale Erklärungen allgemeiner Verhältnisse wie sie von Regularitäten, Gesetzen, Prinzipien oder Theoremen ausgedrückt werden. Klare Fälle sind explanatorische Beweise in

[1] Alternativen dazu sind z.B. Diagramme und gerichtete Graphen (vgl. Woodward 2003: 42).

der Mathematik; zum Beispiel, wenn wir erklären, warum Gödels Theorem gilt, indem wir zeigen, wie seine Wahrheit von den Annahmen abhängt, von denen es abgeleitet ist.

Es gibt zwei Strategien, um nicht-kausale Erklärungen einbeziehen zu können (vgl. Grimm 201*). Eine besteht darin, den Begriff der Ursache auszuweiten. Diese Strategie ist jedoch mit einem Dilemma konfrontiert: Entweder ist der Begriff der Ursache zu eng, um alle Arten von Erklärungen abdecken zu können, oder er weicht zu stark von unserem modernen Begriff der Wirkursache ab. Woodward unterliegt dem ersten Horn. Ihm zufolge gilt jede Erklärung, die kontrafaktische Abhängigkeiten aufzeigt, die damit zu tun haben, was unter Interventionen geschehen würde, als Kausalerklärung (Woodward 2003: 221). Ein solcher Begriff der Kausalerklärung mag zwar bestimmte Erklärungen (wie z.B. Gleichgewichtserklärungen) umfassen, die normalerweise als nicht-kausal gelten. Er trifft aber beispielsweise nicht auf explanatorische Beweise zu, da der kausale Begriff der Intervention in der Mathematik keine Anwendung hat. John Greco unterliegt dagegen dem zweiten Horn. Er verallgemeinert Aristoteles' Lehre der vier Ursachen zu einer Lehre unterschiedlicher Typen explanatorischer Faktoren und bezeichnet jede Art von Beziehung, die einer Erklärung zugrunde liegt, als kausale Beziehung: „Understanding involves ‚grasping', ‚appreciating', or knowing causal relations taken in the broad sense, i.e., the sort of relations that ground explanation." (Greco 2010: 9) Damit stellen sich die scheinbar nicht-kausalen Erklärungen, die ich erwähnt habe, kraft Stipulation als Kausalerklärungen heraus. Ein so weiter Ursachenbegriff, nach dem selbst explanatorische Beweise in der Mathematik als Kausalerklärungen gelten, weicht aber zu stark von unserem modernen Begriff ab, der eng an die Ausübung kausaler Kräfte gebunden ist, um noch ein Ursachenbegriff zu sein.

Damit bleibt nur die zweite Strategie, die auf den allgemeineren Begriff der Abhängigkeit zurückgreift und Kausalbeziehungen als nur eine Form von Abhängigkeitsbeziehungen betrachtet, die Erklärungen zugrunde liegen. Jaegwon Kim drückt die Idee wie folgt aus:

> [M]y claim will be that dependence relations of various kinds serve as objective correlates of explanations. [...] We speak of the „causal dependence" of one event or state on another; that is one type of dependence, obviously one of central importance. Another dependence relation, orthogonal to causal dependence and equally central to our scheme of things, is mereological dependence [...]: the properties of a whole, or the fact that a whole instantiates a certain property, may depend on the properties and relations by its parts. (Kim 1994: 183–184) [2]

Weitere Abhängigkeitsbeziehungen, die Erklärungen zugrunde liegen, sind Supervenienzbeziehungen wie in meinen ersten beiden Beispielen, Identitätsbeziehungen wie in meinem dritten Beispiel, logische und mathematische Beziehungen wie in meinem vierten Beispiel, aber auch begriffliche und teleologische Beziehungen.

Nicht-kausale Erklärungen weichen in zwei Hinsichten von Kausalerklärungen ab, wie ich sie oben charakterisiert habe. Erstens, ich habe behauptet, dass Abhängigkeitsbeziehungen, die Kausalerklärungen zugrunde liegen, die Beantwortung von Was-wäre-wenn-Fragen ermöglichen müssen, indem sie kontrafaktische Konditionale stützen, die darüber Auskunft geben, was der Fall wäre, wenn bestimmte erklärende Faktoren in verschiedener Hinsicht anders wären. Im Fall typischer nicht-kausaler Erklärungen können solche kontrafaktischen Konditionale jedoch nicht so verstanden werden, dass sie darüber Auskunft geben, was unter Interventionen geschehen würde. Zweitens, ich habe die Auffassung vertreten, dass Kausalerklärungen, die explanatorisches Verstehen und nicht bloß explanatorisches Wissen ermöglichen, Informationen über den zugrundeliegenden kausalen Mechanismus liefern müssen. Für nicht-kausale Erklärungen gilt offensichtlich keine entsprechende Bedingung.

[2] Vgl. Ruben 1990: 210; Strevens 2008: 178–179. Auch Greco verfolgt inzwischen diese zweite Strategie; vgl. Greco 2012: 122–123.

### *2.1.3 Erfassen einer Erklärung*

Bisher habe ich diskutiert, in welchem Sinn eine Erklärung, die explanatorisches Verstehen ermöglicht, anspruchsvoller sein muss als eine, die explanatorisches Wissen erlaubt. Für explanatorisches Verstehen reicht es nicht, dass man von einer solchen Erklärung überzeugt ist; man muss sie auch erfassen. Da das Erfassen einer Erklärung eine graduelle Angelegenheit ist, ist explanatorisches Verstehen selbst graduell:

(1) Wenn S (in einem gewissen Ausmaß) versteht, warum p, dann erfasst S (in einem gewissen Ausmaß) eine Erklärung E, die aus dem Explanandum p und einem Explanans besteht, das zeigt, wie p von q abhängt.

Die Rede vom Erfassen einer Erklärung ist eine Metapher, die einer Ausbuchstabierung bedarf. Nach einem naheliegenden Vorschlag erfasst man eine Erklärung, wenn man von ihr korrekterweise überzeugt ist und ein sogenanntes „Aha-Gefühl“ hat. Verstehen ist in der Tat oft von einem solchen Gefühl begleitet. Aber das Haben eines solchen Gefühls ist nicht notwendig und nicht einmal hinreichend, wenn man die Erklärung kennt und daher weiß, wie p von q abhängt. Wie stark das Aha-Gefühl dabei auch ist, man versteht nicht, warum p der Fall ist, wenn man nicht in der Lage ist, von der Erklärung Gebrauch zu machen.

In der Lage sein, von einer Erklärung Gebrauch zu machen, beinhaltet die Fähigkeiten, sie auf einen bestimmten Fall anzuwenden und kontrafaktische Fälle mit ihrer Hilfe zu beurteilen. Das führt zum folgenden Vorschlag (vgl. Hills 2010: 194–195):

(2) S erfasst E (in einem gewissen Ausmaß) gdw. S ist (in einem gewissen Ausmaß) fähig, indem S von den relevanten Elementen von E Gebraucht macht,

- (i) zu schließen, dass p (oder dass wahrscheinlich p), gegeben, dass q,
- (ii) gegeben, dass p, p anhand von q zu erklären,
- (iii) für p* und q*, die ähnlich aber nicht identisch sind mit p und q, zu schließen, dass p* (oder dass wahrscheinlich p*), kontrafaktisch angenommen, dass q*,
- (iv) kontrafaktisch angenommen, dass p*, p* anhand von q* zu erklären.

Die Fähigkeiten (i) und (ii) konstituieren die Fähigkeit, die Erklärung E auf einen bestimmten Fall anzuwenden; die Fähigkeiten (iii) und (iv) konstituieren die Fähigkeit, kontrafaktische Fälle anhand von E zu beurteilen. Die Fähigkeit (iii) ist die Fähigkeit, Was-wäre-wenn-Fragen in Woodwards Sinn zu beantworten: sagen zu können, welchen Unterschied es für das Explanandum p machen würde, wenn bestimmte in der Anfangs- oder Randbedingung q erwähnte Faktoren in verschiedener Hinsicht anders wären. Die Fähigkeit, Was-wäre-wenn-Fragen mithilfe von E beantworten zu können, erweist sich damit als bloß ein Aspekt der allgemeineren Fähigkeit, von E Gebrauch machen zu können. Es ist wichtig zu sehen, dass die spezifizierten Fähigkeiten anspruchsvoller sind als die Fähigkeit, „also p“ (oder „also, p*“) sagen zu können, gegeben die Information, dass q (oder q*), und die Fähigkeit, „weil q“ (oder „weil q*“) sagen zu können, gegeben die Information, dass p (oder p*). Sie beinhalten vielmehr die Fähigkeit, von den relevanten Elementen von E Gebrauch zu machen, um das Argument zu durchlaufen, das E konstituiert. Welches diese relevanten Elemente sind und damit wie die Definition (2) auszubuchstabieren ist, hängt vom Erklärungstyp ab, zu dem E gehört. Im Fall einfacher (kausaler oder nicht-kausaler) Erklärungen einzelner Tatsachen oder Ereignisse ist die Generalisierung G das relevante Element.

Ich schlage vor, dass selbst ein minimales Verständnis, warum p der Fall ist, die Fähigkeiten (i) bis (iv) in einem kontextuell bestimmten Ausmaß erfordert. Auch Wissen involviert das Haben bestimmter Fähigkeiten: Selbst aufgrund von Hörensagen zu wissen, warum die globale Mitteltemperatur angestiegen ist, beinhaltet die Fähigkeit, die Erklärung zitieren zu können, die der Klimawissenschaftler einem gegeben hat. Da man jedoch wissen kann,

warum p der Fall ist, ohne die Fähigkeiten (i) bis (iv) zu haben, ist zu verstehen, warum p der Fall ist, nicht dasselbe wie zu wissen, warum p der Fall ist.

### *2.2 Rechtfertigungsbedingung*

Wie Pritchard feststellt, ist Verstehen ein internalistischer Begriff, „in the sense that it is hard to make sense of how an agent could possess understanding and yet lack good reflectively accessible grounds in support of that understanding" (Pritchard 2010: 76). Neben dem Erfassen einer Erklärung und damit der Fähigkeit, von ihr Gebrauch zu machen, verlangt explanatorisches Verstehen deshalb gute reflexiv zugängliche Gründe für die Erklärung und damit die Fähigkeit, zumindest gewisse konkurrierende Erklärungen auszuschließen. Angenommen, ich erfasse die korrekte Erklärung der globalen Erwärmung, habe aber einem Klimaskeptiker nichts entgegenzuhalten, der einwendet, dass sich das Klima schon immer aufgrund natürlicher Ursachen verändert hatte und der Anstieg der Mitteltemperatur deshalb durch Veränderungen natürlicher Faktoren wie Sonnenaktivität und Aerosolkonzentration aufgrund von Vulkanausbrüchen anstatt durch anthropogene Ursachen erklärt werden muss. Wenn ich nicht in der Lage bin, eine solche konkurrierende Erklärung auf der Grundlage der Rechtfertigung meiner eigenen Erklärung zurückzuweisen, dann habe ich nicht wirklich ein explanatorisches Verständnis der globalen Erwärmung. Eine Definition explanatorischen Verstehens verlangt deshalb eine internalistische Rechtfertigungsbedingung:

(3) Wenn S aufgrund der Erklärung E (in einem gewissen Ausmaß) versteht, warum p, dann ist S (in einem gewissen Ausmaß) fähig, E zu rechtfertigen.

Wenn man in einem gewissen Ausmaß fähig ist, eine Erklärung zu rechtfertigen, ist man in der Lage zu zeigen, dass sie (in einem gegebenen Kontext) hinreichend gut und besser als gewisse konkurrierende Erklärungen ist; wenn man diese Fähigkeit in einem hohen Ausmaß hat, ist man sogar in der Lage zu zeigen, dass die Erklärung die beste verfügbare Erklärung ist. Es liegt deshalb nahe, von der Debatte über den Schluss auf die beste Erklärung Hinweise dafür zu erwarten, welche spezifischen Fähigkeiten involviert sind in der Fähigkeit, eine Erklärung zu rechtfertigen. Aus dieser Debatte kann man die folgende Lehre ziehen:

(4) S ist (in einem gewissen Ausmaß) fähig, E zu rechtfertigen gdw. S ist (in einem gewissen Ausmaß) fähig zu zeigen, dass E

  (i) kohärent ist mit den Hintergrundüberzeugungen von S,

  (ii) den verfügbaren Belegen entspricht und

  (iii) explanatorische Desiderata (wie Reichweite, Einfachheit, Präzision und Mechanismus) optimiert.

Die Bedingung (i) und (ii) betreffen, was Lipton die *likeliness* einer Erklärung nennt: wie wahrscheinlich es ist, dass sie wahr ist, gegeben alle verfügbaren Belege. Die Bedingung (iii) betrifft, was Lipton die *loveliness* einer Erklärung nennt: wie groß ihre Erklärungskraft ist, wenn sie korrekt ist (vgl. Lipton 2004: 59). Die Belege, von denen in (ii) die Rede ist, müssen nicht Beobachtungen sein; es kann sich bei ihnen beispielsweise auch um Intuitionen handeln. Das ermöglicht den Einbezug nicht-empirischer Erklärungen, zum Beispiel in Ethik und Ästhetik. Welche explanatorischen Desiderata gemäß der Bedingung (iii) optimiert werden sollen, hängt vom Erklärungstyp ab, um den es geht. Mechanismus ist beispielsweise relevant bei Kausalerklärungen, nicht aber bei mathematischen Erklärungen.

Wiederum schlage ich vor, dass selbst ein minimales Verständnis, warum p der Fall ist, die spezifizierten Fähigkeiten in einem kontextuell bestimmten Ausmaß erfordert. Da es – zumindest gemäß externalistischen Theorien – möglich ist, zu wissen, warum p der Fall ist, ohne diese Fähigkeiten zu haben, ist explanatorisches Verstehen nicht dasselbe wie explanatorisches Wissen.

## 3. Explanatorisches Verstehen ohne Wissen

Zagzebski (2001), Elgin (2007) und Riggs (2009) vertreten die These, dass Verstehen im Gegensatz zu Wissen nicht faktiv ist; Kvanvig (2003), Pritchard (2010), Hills (2010) und Morris (2011) behaupten, dass sich Verstehen im Gegensatz zu Wissen mit (bestimmten Formen von) epistemischem Glück verträgt. Ich argumentiere im Folgenden dafür, dass beides zutrifft und man deshalb verstehen kann, warum etwas der Fall ist, ohne zu wissen, warum es der Fall ist; explanatorisches Verstehen ist damit nicht einmal eine Art von Wissen. Auch wenn Verstehen nicht faktiv ist, muss die Erklärung, anhand derer man versteht, den Tatsachen gerecht werden. Ich schlage deshalb eine Bedingung der Tatsachentreue vor (Abschnitt 3.1). Eine Folge aus der Nicht-Faktivität von Verstehen ist, dass man explanatorisches Verstehen haben kann, ohne von allen Aussagen der Erklärung überzeugt zu sein. Explanatorisches Verstehen erfordert aber, dass man sich auf seine Erklärung festlegt. Ich ergänze deshalb eine Festlegungsbedingung (Abschnitt 3.2). Weil Verstehen sich mit epistemischem Glück verträgt, verlangt eine Definition explanatorischen Verstehens keine zusätzliche Anti-Glück-Bedingung neben der Rechtfertigungsbedingung (Abschnitt 3.3).

### *3.1 Bedingung der Tatsachentreue*

Jemand kann die Erfassensbedingung und die Rechtfertigungsbedingung erfüllen und dennoch nicht verstehen, warum p der Fall ist, weil seine Erklärung für p schlicht falsch ist. Um zu verstehen, warum p der Fall ist, muss die Erklärung für p den Tatsachen gerecht werden, was wiederum eine graduelle Angelegenheit ist:

(5) Wenn S aufgrund der Erklärung E (in einem gewissen Ausmaß) versteht, warum p, dann wird E (in einem gewissen Ausmaß) den Tatsachen gerecht.

Verstehen wird deshalb häufig für faktiv gehalten (Kvanvig 2003: 190–191; Grimm 2006: 518; Pritchard 2010: 75–76; Hills 2010: 190; Khalifa 2011: 95). Im Folgenden argumentiere ich dagegen, die Bedingung der Tatsachentreue (5) als Faktivitätsbedingung zu konstruieren. Da Wissen ist im Gegensatz zu Verstehen sicherlich faktiv ist, ist es deshalb möglich, zu verstehen, warum etwas der Fall ist, ohne zu wissen, warum es der Fall ist.

Wie würde eine solche Faktivitätsbedingung lauten? Nach Hills ist explanatorisches Verstehen genau dann faktiv, wenn man nicht verstehen kann, warum p der Fall ist, wenn „p“ falsch ist (Hills 2010: 190). Die Faktivitätsbedingung wäre demnach:

(6) Wenn S (in einem gewissen Ausmaß) versteht, warum p, dann ist „p“ wahr.

Explanatorisches Verstehen mag zwar faktiv sein in diesem Sinn: Man kann nicht verstehen, warum die globale Mitteltemperatur angestiegen ist, wenn sie nicht angestiegen ist. (6) ist aber sicher zu schwach, um die Faktivität von explanatorischem Verstehen sicherzustellen. Betrachten wir die Faktivität von Wissen, welche der Wahrheitsbedingung folgt. Man weiß, warum p der Fall ist, wenn man weiß, dass p der Fall ist, weil q der Fall ist; das impliziert, dass „p ist der Fall, weil q der Fall ist“ wahr ist. Explanatorisches Verstehen scheint deshalb genau dann faktiv zu sein, wenn man nicht verstehen kann, warum p der Fall ist, wenn man q als Ursache oder Grund für p hält, aber „p“ oder „q“ falsch sind oder q nicht die Ursache oder der Grund ist, warum p der Fall ist. Die Faktivitätsbedingung wäre demnach:

(7) Wenn S (in einem gewissen Ausmaß) versteht, warum p, und q für die Ursache oder den Grund für p hält, dann ist „p, weil q“ wahr.

Nach Pedro Schmechtig stellt dies nur die Veridizität von Verstehen sicher. Um faktiv zu sein, müsse explanatorisches Verstehen die Wahrheit von „p, weil q“ nicht bloß implizieren, sondern in einem stärkeren Sinn präsupponieren, so dass sie auch unter Negation der Verbphrase erhalten bleibt. Damit explanatorisches Verstehen in diesem Sinn faktiv ist,

müssten „S versteht, warum p“ ebenso wie „S versteht nicht, warum p“ beide implizieren, dass „p, weil q“ wahr ist (Schmechtig 2011: 26–29). Wenn man diese Unterscheidung akzeptiert, betreffen meine Ausführungen die Veridizität, nicht die Faktivität. Ich werde nichts über Faktivität in diesem anspruchsvolleren Sinn sagen, weil sie zwar einen weiteren Grund für die Unterscheidung zwischen Verstehen und Wissen liefern kann, aber nicht zu einer weiteren Bedingung für explanatorisches Verstehen führt. Wir sind also wieder bei (7).

An dieser Stelle reicht die Feststellung, dass (7) vernachlässigt, dass die Erklärung, anhand der S versteht, warum p der Fall ist, mehrere Aussagen umfasst. Selbst in einfachen Fällen enthält sie eine Anfangs- oder Randbedingung, welche die Faktoren angibt, von denen das Explanandum-Ereignis oder Faktum abhängt, und eine Generalisierung, die angibt, wie dieses von den spezifizierten Faktoren abhängt. In komplexeren Fällen mit mehreren interagierenden Ursachen oder Gründen sind mehrere Anfangs- oder Randbedingungen und allenfalls mehrere Generalisierungen im Spiel. Ist man sich dessen bewusst, ist nicht länger offensichtlich, wie eine Faktivitätsbedingung für explanatorisches Verstehen lauten soll.

Hilfe bietet die Diskussion über die (Nicht-)Faktivität von objektualem Verstehen. Übertragen auf explanatorisches Verstehen aufgrund einer Erklärung E gilt dieses gemäß einer starken Version genau dann als faktiv, wenn alle Aussagen, die E konstituieren, wahr sind. Die Faktivitätsbedingung würde demnach wie folgt lauten (vgl. Kvanvig 2003: 191):

(8) Wenn S aufgrund von E (in einem gewissen Ausmaß) versteht, warum p, dann sind alle Aussagen von E wahr.

Aber selbst Vertreter der Faktivitätsthese anerkennen, dass einige wenige periphere Falschheiten zwar das Verständnis mindern, aber nicht gänzlich aufheben. Das scheint für explanatorisches Verstehen genauso zu gelten wie für objektuales. Die Faktivitätsbedingung ist deshalb wie folgt abzuschwächen (Kvanvig 2003: 201–202; vgl. Elgin 2007: 36):

(9) Wenn S aufgrund von E (in einem gewissen Ausmaß) versteht, warum p, dann sind die meisten Aussagen von E und alle zentralen Aussagen von E wahr.

Diese Bedingung ist in zwei Hinsichten vage. Sie legt weder fest, wie viele Aussagen falsch sein können, ohne das Verständnis aufzuheben, noch, welche Aussagen als zentral und welche als peripher zu gelten haben (Kvanvig 2009: 341). Die Vagheit lässt beträchtlichen Spielraum, um die Faktivität von Verstehen zu retten.

Wie Catherine Elgin aufgezeigt hat, gibt es dennoch zwei Typen klarer Fälle, in denen explanatorisches Verstehen nicht faktiv ist, zumindest wenn wir akzeptieren, dass wir den Verstehensbegriff in einer solchen Weise konstruieren sollten, dass wir den Wissenschaften zumindest ein gewisses Verständnis der Phänomene, mit denen sie sich beschäftigen, zuschreiben können (Elgin 2007: 36–39).

Erstens ist Verstehen in einem stärkeren Ausmaß graduell als Vertreter der Faktivitätsthese meinen, da selbst Falschheiten, die durchaus zentral sind, manchmal das Verständnis bloß schmälern, ohne es gänzlich aufzuheben. In alltäglichen Lernprozessen ebenso wie in der wissenschaftlichen Ausbildung starten wir mit Charakterisierungen, die strikt genommen falsch sind, uns aber angemessen auf die Phänomene ausrichten, um zu Überzeugungen zu gelangen, die der Wahrheit näher kommen. Eine solche Entwicklung mag mit wahren Überzeugungen enden, aber selbst frühere Phasen liefern ein gewisses Verständnis. Ähnlich verhält es sich mit den Wissenschaften und der Abfolge der Erklärungen, die sie entwickeln. Denken wir an eine Ptolemäische, eine Kopernikanische und eine zeitgenössische Erklärung der Planetenbewegungen. Obwohl Kopernikus fälschlicherweise annahm, dass sich die Erde kreisförmig um die Sonne bewegt, stellt seine Erklärung einen großen Verständnisfortschritt dar gegenüber der Ptolemäischen Erklärung (Elgin 2007: 37). Wir können aber nur dann anerkennen, dass Kopernikus zumindest ein gewisses Verständnis der Planetenbewegungen hatte, wenn wir zugeben, dass selbst zentrale Falschheiten das Verständnis zwar mindern,

aber nicht gänzlich aufheben, wenn sie in der richtigen Umgebung sind. Überdies bestehen selbst aktuelle wissenschaftliche Erklärungen – der Planetenbewegungen ebenso wie anderer Phänomene – kaum weitestgehend aus Wahrheiten mit einigen recht unbedeutenden Falschheiten an der Peripherie. Wir können deshalb nur dann anerkennen, dass die gegenwärtigen Wissenschaften zumindest ein gewisses Verständnis der Phänomene haben, mit denen sie sich beschäftigen, wenn Verstehen nicht immer faktiv ist.

Kvanvig hat demgegenüber eingewendet, dass wir „Verstehen" in solchen Fällen in einem bloß honorierenden Sinn verwenden, genau wie „Wissen", wenn wir vom „gegenwärtigen Stand wissenschaftlichen Wissens" sprechen und dabei eingestehen, dass einiges, was dazu gehört, falsch ist. Honorierende Verwendungen epistemischer Termini gehören ihm zufolge aber zur Pragmatik und nicht zur Semantik epistemischer Terminologie. Entsprechend haben honorierende Verwendungen von „Verstehen" so wenig einen Einfluss darauf, was Verstehen ist, wie honorierende Verwendungen von „Wissen" einen Einfluss darauf haben, was Wissen ist (Kvanvig 2009: 341–342). Diese Analogie vermag jedoch nicht zu überzeugen. Nach unserer gewöhnlichen Verwendungsweise ziehen wir einen Wissensanspruch zurück, wenn wir entdecken, dass die fragliche Überzeugung falsch ist. Es ist deshalb vernünftig, propositionales und infolgedessen explanatorisches Wissen faktiv zu konstruieren. Unsere gewöhnliche Verwendung von „Verstehen" im Zusammenhang mit Erklärungen (oder ganzen Theorien) ist aber flexibler. Wir stimmen sicherlich zu, dass Kopernikus nicht gewusst hat, dass die Erde sich kreisförmig um die Sonne bewegt, es scheint aber unangemessen, ihm jedes Verständnis der Planetenbewegungen abzusprechen. Wir gehen typischerweise davon aus, dass jemand ein gewisses Verständnis von etwas haben kann, auch wenn einige recht zentrale Aussagen seiner Erklärung (oder Theorie) etwas von der Wahrheit abweichen. Es ist deshalb vernünftig, „Verstehen" in seiner explanatorischen (und objektualen) Verwendung nicht-faktiv zu konstruieren (vgl. Elgin 2007: 39).

Zweitens machen Erklärungen häufig von Idealisierungen Gebrauch. Das ideale Gasgesetz beispielsweise erklärt das Verhalten tatsächlicher Gase, indem es das Verhalten eines Gases beschreibt, das aus vollkommen elastischen, ausdehnungslosen Massepunkten besteht, die keinerlei Kräfte aufeinander ausüben. Ein solches Gas kann es nicht geben. Dennoch wird das Verhalten tatsächlicher Gase in Fällen, in denen die Abweichung vom Ideal vernachlässigbar ist (grob gesagt in Fällen monoatomischer Gase bei hoher Temperatur und geringem Druck), mit Bezug auf die Idealisierung erklärt (Elgin 2007: 38). Solche Idealisierungen sind strikt genommen falsch, aber sie können weder aus wissenschaftlichen Erklärungen entfernt noch an deren Peripherie verbannt werden. Wir können deshalb nur dann anerkennen, dass Erklärungen, die Idealisierungen enthalten, ein gewisses explanatorisches Verständnis liefern können, wenn wir akzeptieren, dass dieses nicht immer faktiv ist.

Kvanvig hat eingewendet, dass die Verwendung von Idealisierungen uns nicht auf die Nicht-Faktivität von Verstehen festlegt, wenn wir einsehen, dass das Objekt des Verstehens nicht einfach das Modell ist, sondern auch die Beziehung zwischen dem Modell und der Realität beinhaltet, inklusive Informationen darüber, welche Aspekte der Realität das Modell beleuchten soll und in welchem Ausmaß es eine Idealisierung ist (Kvanvig 2009: 342–343; vgl. Greco 2012: 126–127). Dem ist entgegenzuhalten, dass das Objekt des Verstehens weder das Modell noch das Modell und seine Beziehung zur Realität ist, sondern das Phänomen in der Welt. Ein solches Verständnis setzt ein Verständnis des fraglichen Modells und damit ein Bewusstsein davon voraus, dass dieses eine Idealisierung darstellt. Es scheint aber nicht in jedem Fall ein Wissen darüber zu beinhalten, wie genau das Modell von der Realität abweicht. Andernfalls wäre kaum einsichtig, was der Witz einer Idealisierung ist. Klimawissenschaftler zumindest wissen gerade nicht genau, wie die Klimamodelle von der Realität abweichen; dennoch haben sie ein gewisses Verständnis der globalen Erwärmung.

### *3.2 Festlegungsbedingung*

Die Nicht-Faktivität explanatorischen Verstehens impliziert eine weitere Differenz zwischen Verstehen und Wissen. Wenn eine Erklärung von Idealisierungen Gebrauch macht, kann jemand aufgrund dieser Erklärung ein gewisses explanatorisches Verständnis von ihrem Gegenstand haben, ohne die Erklärung (resp. alle ihre Aussagen) für wahr zu halten. Wenn nun die Einstellung des Überzeugtseins als die Einstellung des Für-wahr-Haltens verstanden wird, dann impliziert Verstehen im Gegensatz zu Wissen nicht, dass man vom Inhalt der Erklärung überzeugt ist. Verstehen impliziert aber sicherlich, dass man sich in irgendeiner Weise auf die Erklärung festlegt. Da man sich mehr oder weniger auf eine Erklärung festlegen kann, resultiert die folgende Festlegungsbedingung:

(10) Wenn S aufgrund von E (in einem gewissen Ausmaß) versteht, warum p, dann legt S sich (in einem gewissen Ausmaß) auf E fest.

Natürlich kann „S legt sich auf E fest“ nicht verstanden werden als „S hält E (oder die Aussagen von E) für wahr“. Dennoch kann man die Festlegung auf eine Erklärung anhand des Begriffs der Überzeugung verstehen. Ich habe dafür argumentiert, dass explanatorisches Verstehen in einer eher losen Weise mit Wahrheit verbunden und zudem auf eine Vielfalt explanatorischer Desiderata bezogen ist, die nicht nur insofern Desiderata sind, als sie das Erwerben von Wahrheiten befördern. Als Folge davon kann „S legt sich auf E fest“ verstanden werden als „S ist überzeugt, dass E den Tatsachen gerecht wird und optimal ist in Bezug auf die relevanten explanatorischen Desiderata“ (resp. als „S ist überzeugt, dass E die beste verfügbare und im gegebenen Kontext eine hinreichend gute Erklärung für p ist“).

### *3.3 Anti-Glück-Bedingung?*

Nach Pritchard und Hills zeigen Beispiele, bei denen epistemisches Glück im Spiel ist, dass es möglich ist, zu verstehen, warum etwas der Fall ist, ohne zu wissen, warum es der Fall ist (Pritchard 2010: 78–90; Hills 2010: 106). Liegen sie richtig, ist dies ein weiterer Grund dafür, dass explanatorisches Verstehen nicht einmal eine Art von Wissen ist.

Betrachten wir ein Beispiel, bei dem das epistemische Glück allein damit zu tun hat, dass man sich in einer aus epistemischer Sicht unfreundlichen Umgebung befindet, weshalb Pritchard von „environmental epistemic luck“ spricht (Pritchard 2010: 78). Nehmen wir an, ich erarbeite mir ein Verständnis davon, warum die globale Mitteltemperatur angestiegen ist, indem ich ein zuverlässiges Buch studiere. Nehmen wir zudem an, dass alle anderen Bücher über die globale Erwärmung ganz und gar unzuverlässig und voller Falschheiten sind, aber für einen Laien wie mich genauso wissenschaftlich daherkommen, so dass es ein bloßer Zufall ist, dass ich das zuverlässige Buch ausgewählt habe. Das in einem solchen Fall involvierte epistemische Glück schließt es nach einer weitverbreiteten Ansicht aus, dass ich Wissen erwerbe, da ich leicht ein unzuverlässiges Buch hätte kaufen und eine falsche Erklärung geben können. Wenn jemand Wissen hat, ist es dieser Ansicht zufolge aber ausgeschlossen, dass seine wahre Überzeugung leicht hätte falsch sein können. Das involvierte epistemische Glück unterminiert aber nicht mein Verständnis, schließlich erfasse ich eine korrekte Erklärung der globalen Erwärmung und habe gute Gründe für sie und bin daher im Besitz der notwendigen Erklärungs- und Rechtfertigungsfähigkeiten. Dies wäre ich selbst in einem Fall, in dem mein eigenes Buch das Ergebnis bloßen Ratens und daher ganz und gar unzuverlässig, seine Erklärung der globalen Erwärmung und aber zufälligerweise vollkommen richtig ist. Das Glück, das in diesem zweiten Fall im Spiel ist, nennt Pritchard „standard Gettier-style epistemic luck“; anders als im ersten Fall „schaltet“ es sich zwischen das Verstehenssubjekt und die Realität, aber in einer solchen Weise, dass seine Erklärung dennoch korrekt ist (vgl. Pritchard 2010: 36). Im Gegensatz zu Pritchard neige deshalb mit Hills (2010: 196, Anm. 13) zur Behauptung, dass explanatorisches Verstehen auch mit diesem zweiten Typ von epistemischem Glück vereinbar ist. Wenn dies zutrifft, dann ist noch klarer, dass man

verstehen kann, warum etwas der Fall ist, ohne zu wissen, warum es der Fall ist. Während gelegentlich bezweifelt wird, dass Wissen tatsächlich unvereinbar ist mit *environmental epistemic luck* (z.B. bei Grimm 2006: 527–529 und Roberts/Wood 2007: 57), ist weitgehend unbestritten, dass es unvereinbar ist mit *standard Gettier-style epistemic luck*.

Eine Überlegung von Kvanvig liefert einen weiteren Grund für die Annahme, dass Verstehen mit beiden Typen von epistemischem Glück verträglich ist. Ihr zufolge unterscheiden sich Verstehen und Wissen bezüglich ihrer Verträglichkeit mit epistemischem Glück, weil wir bei ihnen Unterschiedliches im Fokus haben. Beim Verstehen fokussieren wir auf das Erfassen explanatorischer Verbindungen und damit, kann man ergänzen, auf das Haben bestimmter Fähigkeiten. Beim Wissen fokussieren wir auf das Überzeugtsein von einer Aussage, die nicht leicht hätte falsch sein können, und damit auf die Nicht-Zufälligkeit. Dass man die Überzeugung aufgrund eines glücklichen Zufalls erworben hat, unterminiert deshalb das explanatorische Wissen; dass man die fraglichen Fähigkeiten aufgrund eines glücklichen Zufalls erworben hat, unterminiert aber nicht das explanatorische Verstehen. In dieser Hinsicht gleicht Verstehen-warum-etwas-der-Fall-ist dem Wissen-wie-etwas-zu-tun-ist, das ebenfalls mit epistemischem Glück kompatibel ist. Wenn wir beurteilen, ob jemand weiß, wie etwas zu tun ist, fragen wir auch ausschließlich, ob er die fragliche Fähigkeit besitzt, nicht aber wie er sie erworben hat (vgl. Grimm 2011: 92–93). Ich gestehe einem Freund auch dann gerne zu, dass er weiß, wie mein Computer wieder zum Laufen gebracht werden kann, wenn ich später erfahre, dass er sein *Know-how* aufgrund glücklicher Zufälle erworben hat.

Wenn epistemisches Glück, wie es in Gettier-Beispielen im Spiel ist, explanatorisches Verstehen nicht unterminiert, dann verlangt eine Definition explanatorischen Verstehens über die Rechtfertigungsbedingung hinaus keine zusätzliche Anti-Glück-Bedingung.

## 4. Definition explanatorischen Verstehens

Aufgrund meiner Überlegungen schlage ich die folgende Definition vor: S versteht (in einem gewissen Ausmaß), warum p gdw.

(a) S legt sich (in einem gewissen Ausmaß) auf eine Erklärung E fest, die aus dem Explanandum p und einem Explanans besteht, das zeigt, wie p von q abhängt;

(b) S erfasst E (in einem gewissen Ausmaß); das heißt: S ist (in einem gewissen Ausmaß) fähig, indem S von den relevanten Elementen von E Gebraucht macht,

  (i) zu schließen, dass p (oder dass wahrscheinlich p), gegeben, dass q,

  (ii) gegeben, dass p, p anhand von q zu erklären,

  (iii) für p* und q*, die ähnlich aber nicht identisch sind mit p und q, zu schließen, dass p* (oder dass wahrscheinlich p*), kontrafaktisch angenommen, dass q*,

  (iv) kontrafaktisch angenommen, dass p*, p* anhand von q* zu erklären;

(c) S ist (in einem gewissen Ausmaß) fähig, E zu rechtfertigen; das heißt: S ist (in einem gewissen Ausmaß) fähig zu zeigen, dass E

  (i) kohärent ist mit den Hintergrundüberzeugungen von S,

  (ii) den verfügbaren Belegen entspricht und

  (iii) explanatorische Desiderata (wie Reichweite, Einfachheit, Präzision und Mechanismus) optimiert; und

(d) E wird (in einem gewissen Ausmaß) den Tatsachen gerecht.

Meine Definition hat strukturelle Ähnlichkeiten mit der traditionellen Konzeption des propositionalen Wissens: (a) – vielmehr als (b) – entspricht der Überzeugungsbedingung, (c) der Rechtfertigungsbedingung und (d) der Wahrheitsbedingung. Die Erfassensbedingung (b) ist dagegen spezifisch für Verstehen. Die Definition lässt sich für objektuales Verstehen eines Gegenstands- oder Themenbereichs anhand einer Theorie adaptieren. In (a), (c) und (d) braucht „Erklärung E" im Wesentlichen bloß durch „Theorie T" substituiert zu werden (wobei die Charakterisierung der Erklärung in (a) natürlich entfällt, die Kohärenzbedingung in (c) durch Kohärenz innerhalb der Theorie zu ergänzen ist und die explanatorischen Desiderata durch allgemeinere theoretische Desiderata ersetzt werden müssen). Nur (b) verlangt nach einer substantiellen Erweiterung, da das Erfassen einer Theorie sich in einer Reihe zusätzlicher Fähigkeiten manifestiert. Das Erfassen einer Theorie der globalen Erwärmung anstatt bloß einer Erklärung ihrer Ursachen beinhaltet einerseits mehr Fähigkeiten derselben Art, da eine solche Theorie weitere Erklärungen umfasst, zum Beispiel der Auswirkungen der globalen Erwärmung auf natürliche und soziale Systeme. Andererseits beinhaltet es wohl auch zusätzliche Fähigkeiten, wie beispielsweise Emissionsszenarien entwickeln zu können, diese verwenden zu können, um zukünftige Treibhausgaskonzentrationen und den weiteren Temperaturverlauf vorauszusagen, Unsicherheiten solcher Voraussagen aufgrund von Klimamodellen einschätzen zu können und verschiedene Lösungen des Klimaproblems (wie Vermeiden, Anpassen und Geoengineering) beurteilen zu können. Zu bestimmen, welche Fähigkeiten in einem spezifischen Fall von objektualem Verstehen involviert sind, ist keine Aufgabe für Erkenntnistheoretiker; zudem wird sich dies von Fall zu Fall stark unterscheiden.

## 5. Propositionales Verstehen

Abschließend ziehe ich aus meiner Diskussion einige Konsequenzen für propositionales Verstehen und propositionales Wissen. Gemäß zwei weiterverbreiten Behauptungen ist explanatorisches Verstehen eine Art von propositionalem Verstehen und dieses wiederum ist äquivalent mit propositionalem Wissen:

(11) S versteht, warum p, gdw. S versteht, dass q eine korrekte Antwort auf die Warum-Frage ist (vgl. Kvanvig 2003: 189–190).

(12) S versteht, dass p, gdw. S weiß, dass p (vgl. Elgin 2007: 34; Grimm 2011: 85).

Weil „dass p" in (12) durch „dass q eine korrekte Antwort auf die Warum-Frage ist" in (11) ersetzt werden kann, implizieren (11) und (12) zusammen, dass explanatorisches Verstehen eine Art von propositionalem Wissen ist:

(13) S versteht, warum p, gdw. S weiß, dass q eine korrekte Antwort auf die Warum-Frage ist.

Die Behauptung (13) ist attraktiv, da sie die Möglichkeit verspricht, den schwer fassbaren Begriff des explanatorischen Verstehens über den viel besser verstandenen Begriff des propositionalen Wissens zu erklären. Wenn aber meine Überlegungen in diesem Aufsatz korrekt sind und explanatorisches Verstehen also weder identisch mit noch eine Art von Wissen ist, dann ist (13) falsch. Wenn (13) falsch ist, dann muss man entweder (11) oder (12) oder beide Behauptungen aufgeben.

Ich meine, dass (11) aufgegeben werden sollte, da die Zurückführung des explanatorischen Verstehens auf das propositionale Verstehen entweder unmöglich oder nutzlos ist. Sie ist unmöglich, wenn der Inhalt des propositionalen Verstehens sich in einer Aussage darüber erschöpft, von welchen Faktoren p abhängt. Verstehen, warum p der Fall ist, verlangt, dass man auch erfasst, wie p von diesen Faktoren abhängt. Die Zurückführung ist nutzlos, wenn propositionales Verstehen sehr anspruchsvoll konzipiert wird, so dass sein Inhalt durch eine komplexe Aussage gebildet wird, zu der auch Aussagen darüber gehören, wie p von den

spezifizierten Faktoren abhängt, und es zudem die erwähnten Erklärungsfähigkeiten und Rechtfertigungsfähigkeiten beinhaltet. Propositionales Verstehen wird dann keineswegs besser verstanden als explanatorisches Verstehen; das zweite auf das erste zurückzuführen, bringt uns deshalb nicht weiter. (11) aufzugeben macht es möglich, an (12) festzuhalten, einer Behauptung, die von fast allen an der Verstehensdebatte Beteiligten akzeptiert wird. Zudem erübrigt es (12), nach einer zusätzlichen Definition für propositionales Verstehen zu suchen.[3]

**Christoph Baumberger**
ETH Zürich & Universität Zürich
christoph.baumberger@env.ethz.ch

## Literatur

Elgin, C. Z. 1996: *Considered Judgment*. Princeton: Princeton University Press.

— 2007: „Understanding and the Facts“, *Philosophical Studies* 132, 33–42.

Grimm, S. R. 2006: „Is Understanding a Species of Knowledge?“, *British Journal for the Philosophy of Science* 57, 515–35.

— 2011: „Understanding“, in S. Bernecker und D. Pritchard (Hrsg.): *Routledge Companion to Epistemology*, New York: Routledge, 84–94.

— 201*: „Understanding as Knowledge of Causes“, erscheint in A. Fairweather (Hrsg.): *Virtue Scientia: Virtue Epistemology and Philosophy of Science*, Dordrecht: Springer.

Greco, J. 2010: *Achieving Knowledge. A Virtue Theoretic Account of Epistemic Normativity*. Cambridge: Cambridge University Press.

— 2012: „Intellectual Virtues and Their Place in Philosophy“, in C. Jäger und W. Löffler (Hrsg.): *Epistemology: Contexts, Values, Disagreement*, Frankfurt am Main: Ontos, 117–30.

Haddock, A., A. Millar und D. Pritchard (Hrsg.) 2009: *Epistemic Value*. Oxford: Oxford University Press.

Hills, A. 2010: *The Beloved Self. Morality and the Challenge form Egoism*. Oxford: Oxford University Press.

Khalifa, K. 2011: „Understanding, Knowledge, and Scientific Antirealism“, *Grazer Philosophische Studien* 83, 93–112.

Kim, J. 1994: „Explanatory Knowledge and Metaphysical Dependence“, in *Essays in the Metaphysics of Mind*, Oxford: Oxford University Press 2010, 167–86.

Kvanvig, J. 2003: *The Value of Knowledge and the Pursuit of Understanding*. New York: Cambridge University Press.

— 2009: „Response to Critics“, in A. Haddock, A. Millar und D. Pritchard (Hrsg.) 2009, 339–51.

Lipton, P. 2004: *Inference to the Best Explanation*. New York: Routledge.

Morris, K. 2011: „A Defense of Lucky Understanding“, *British Journal for the Philosophy of Science* 0, 1–15.

Pritchard, D. 2010: „Knowledge and Understanding“, in D. Pritchard, A. Millar und A. Haddock: *The Nature and Value of Knowledge. Three Investigations*, Oxford: Oxford University Press, 1–88.

[3] Ich danke Georg Brun und Gertrude Hirsch Hadorn für ihre hilfreichen Kommentare.

Riggs, W. D. 2003: „Understanding ‚Virtue' and the Virtue of Understanding", in M. DePaul und L. Zagzebski (Hrsg.): *Intellectual Virtue*, Oxford: Clarendon Press, 203–26.

— 2009: „Understanding, Knowledge, and the *Meno* Requirement", in A. Haddock, A. Millar und D. Pritchard (Hrsg.) 2009, 331–38.

Roberts, R. C. und W. J. Wood 2007: *Intellectual Virtues. An Essay in Regulative Epistemology*. Oxford: Clarendon Press.

Ruben, D. H. 1990: *Explaining Explanation*. London: Routledge.

Schmechtig, P. 2011: „Der epistemische Wert des Verstehens und Wissen-wie", *Dresdner Berichte in theoretischer Philosophie und philosophischer Logik* 38, 1–36.

Woodward, J. 2003: *Making Things Happen. A Theory of Causal Explanation*. New York: Oxford University Press.

Zagzebski, L. 2001: „Recovering Understanding", in M. Steup (Hrsg.): *Knowledge, Truth and Duty*, New York: Oxford University Press, 235–51.

# How Gettier Helps to Understand Justification

Frank Hofmann

Epistemic justification shares an important structure with knowledge: being success from ability. Therefore, justification also allows for Gettier cases. I will describe such a Gettier case for justification and present a diagnosis, relying on virtue epistemology. A virtue-evidentialist account of justification arises quite naturally from these considerations.

## 1. Introduction

I would like to argue for the thesis that knowledge and (epistemic) justification share a common structure, namely, the structure of 'success from ability', i.e., success which is explained by the exercise of ability. The structure of 'success from ability' is well-known from discussions of knowledge, as a virtue-theoretic structure (cf. Sosa 2007, Greco 2009). I will not try to defend this view here, but simply suppose that it is basically correct. My – original – claim that I will try to defend is that the very same virtue-theoretic structure attaches to justification as well. Thus, justification also allows for Gettier cases, i.e., cases in which the success is reached but is not reached from (or due to) ability. The subject reaches the relevant success luckily – in a certain sense of 'luck', of course: in the case of knowledge it is 'veritic luck' (i.e., lucky truth; cf. Pritchard 2005), whereas in the case of justification it is 'reasons luck', as we can call it (i.e., luckily being supported or backed up by some evidence).

I will describe a Gettier case for justification (section 2). Then I will present a diagnosis that is framed within virtue epistemology (section 3). And finally, I shall sketch a view of justification and evidence which is quite attractive and allows us to make sense of all the relevant phenomena discussed here (section 4).

Before I begin with describing the Gettier case for justification, however, a word on what is meant by '(epistemic) justification' is in order. I take it that the term 'justified belief' is not univocal. It can be used to denote various (similar but) different phenomena: blameless and/or responsible belief, rational belief, reasons-related belief, entitled belief, and possibly even further phenomena.[1] It is rather uninteresting to debate which one really deserves to be called 'justified belief'. What is important for epistemology is to see the differences and commonalities that are relevant for the epistemic status of a belief. Here I will focus on 'justified belief' in the sense of reasons-related belief, i.e., belief which is appropriately linked to some objective piece of evidence (reason). (I will spell out below in more detail what I mean by 'evidence' and what the 'appropriate link' involved is.) Concerning *this* phenomenon I will argue that it has a success-from-ability structure and, thus, allows for Gettier cases. The central question is: how exactly does a belief have to be related to a reason in order to count as justified? The claim about the success-from-ability structure is supposed to answer this question.

[1] David (2005) presents a quite similar list.

## 2. A Gettier Case for Justification

I will now describe a Gettier case for justification. It will involve a subject – call her 'Daniela' – who is veridically hallucinating. In veridical hallucination, a subject undergoes a perceptual experience with a content that is in fact satisfied by the world. The world perceptually appears a certain way to the subject, and the world actually is that way. But the satisfaction of the experience's content is entirely a matter of sheer luck; there is no systematic dependence of the experience on the world (as there is in cases of genuine perception). We can suppose that the subject is entirely unaware of the fact that she is hallucinating. She thinks that she is perceiving (or has no thought or belief about this matter at all, if that is possible). The following case of Daniela will be such a case of unrecognized veridical hallucination.[2]

Daniela is an ordinary subject, equipped with ordinary perceptual and conceptual capacities. However, she has taken a hallucination-inducing drug, unrecognizedly. Presently she undergoes a hallucinatory experience as of a red tomato in front of her. By sheer coincidence, there is in fact such a red tomato in front of her (exactly as she is hallucinating it). The world is the way it appears to her. She is veridically hallucinating. On the basis of her experience, Daniela forms the belief that there is a red tomato in front of her. Now suppose that in Daniela's context the facts that she is veridically hallucinating (certain color and shape facts) are really evidence or reasons for believing that there is a red tomato (believing that p). (These facts reliably indicate the presence of a red tomato, in Daniela's context at least.) So there is evidence for her belief that p. We can say that her belief is 'evidenced', to coin a phrase.

Intuitively, Daniela is in possession of evidence, since she veridically represents it in her experience. She grasp the evidence, or is somehow aware of it (though the awareness need not be phenomenal consciousness). And Daniela respects her evidence, by rationally forming the belief that p on the basis of her experience. (We can safely suppose that she has no further evidence as to whether there is a red tomato in front of her or not.) Her belief that p is rational, intuitively. Yet something is missing. There is something suboptimal about her belief. Intuitively, her belief that p is not justified. This comes out quite clearly if we compare her case with the corresponding good case in which she is genuinely perceiving the same facts (*ceteris paribus*). In the good case, her belief would be justified. But actually it is not.

Let me clarify what I mean by 'evidence' and 'reasons'. Here I am not using these terms in a generic, unspecific sense. Rather I will use them in the sense of objective facts (or other kinds of entities) that speak in favor of the truth of certain beliefs. Typically, the evidence is non-psychological (even though it can be psychological in special cases). Fingerprints and traces are paradigm examples of evidence. They are reasons, epistemic reasons. So by 'reasons' I do not mean anything whatsoever which makes a belief epistemically justified or contributes to its being justified. These justificatory factors are not what should be called reasons. Rather, I am using the terms 'evidence' and 'reasons' in a specific sense which accords with a specific strand of thinking of 'evidence' and 'reasons' that is present in our ordinary thinking and reasoning. And I will use '(piece of) evidence' and 'reason' interchangeably and synonymously here. For example, a reliable process of belief formation may or may not be a reason or piece of evidence for the subject's own belief that is the output of that process. But even if it is evidence, unless the subject is somehow aware of this fact (that her belief is the output of that reliable process) she is not in possession of this evidence.[3]

A variation of Daniela's case would be a corresponding case in which things are the same except for the fact that there is no tomato in front of her but a wax object that perceptually duplicates the genuine object. Then Daniela's belief would be false. – This shows that the

[2] For veridical hallucination see Lewis (1980).

[3] For more on reasons and evidence, see Hofmann (2013).

issue is independent of the truth of the resulting belief and, thus, that the truth of the resulting belief is not relevant. But we can focus on the simpler case, as described above, where the belief that p is true.

Now consider the following first reaction, or objection, to the case. One might think that Daniela's belief is not simply not justified, but is less justified than the same belief in the good case, or is justified but *not fully* justified. – This first objection, however, is not really convincing. The phenomenon is not a matter of degree but of quality. Daniela is entirely luckily in possession of evidence, as luckily as could be. And in the corresponding good case she would be entirely systematically and reliably in possession of evidence (since genuinely perceiving). So Daniela's case (i.e., the bad case of veridical hallucination) is not just a case of less than full justification. There is simply no justification. The difference is just like the difference between being flat and being bumpy. Ultimately, it may be a matter of degree – just as vastly many thing are ultimately a matter of degree. But the concept of justification that we are working with is such that it allows us to make the flat-out, 'absolute' claim that her belief is not justified. It would be inappropriate to describe her belief as justified but not fully justified. (If one wished to introduce a dimension of degree in one's conception of justification, one could do so, presumably by linking the degree of justification to the strength of the evidence that the subject possesses. But the strength of the evidence is not the issue in the case of Daniela's belief. One could make it as strong as possible if one wanted to.)

## 3. A Virtue-theoretic Diagnosis

We can draw the following two lessons from the case. (1.) Possession of evidence comes in two varieties, lucky and non-lucky possession of evidence. (2.) Justification requires non-lucky possession of evidence. The reason why Daniela's belief lacks justification is that her possession of the evidence is lucky.

It remains to clarify what kind of *luck* we are dealing with here. It is of course a certain kind of epistemic luck. This kind of epistemic luck is not to be confused with evidential luck. Evidential luck is familiar from discussions of luck in connection with knowledge. One may come luckily into possession of some piece of evidence, for example, by simply turning one's head for the fun of it and, thus, coming to see something (which is a piece of evidence). This kind of luck does not undermine knowledge.[4] But this kind of luck, evidential luck, is not what we are dealing with here in our discussion of justification. Seeing something (perceiving something) is not at all a lucky way of coming into possession of evidence in the sense which is relevant for justification. Quite the contrary, it is usually a highly reliable and systematic, non-lucky way of possessing evidence. It may be a matter of luck that one engages in perceiving at all. But the perceiving is not a lucky connection to the world (the perceived entities and facts). So the kind of epistemic luck that undermines justification is not to be confused with the sort of evidential luck just described. What matters for justification is whether the correct representation of the piece of evidence is formed in a reliable, competent way or not. One may accidentally believe something which is a reason for a certain other belief (for example, by lucky guessing), and one may then form this other belief on the basis of the former one. Then one is actually in possession of the evidence, and one exploits one's evidence in the appropriate, rational way. But this will not be sufficient for achieving justification. Possessing evidence by a lucky guess is not good enough for arriving at a justified belief.

Once we recognize that possession of evidence comes in these two forms, the lucky and the non-lucky form, a further step of theorizing seems almost obvious. A '*virtue-evidentialist account of justification*' suggests itself. According to such an account, one's justification

[4] See, for example, Pritchard (2005), ch. 5.

depends on competent possession of evidence. More precisely, the following basic claim about justification seems natural and very attractive:

(VE) A subject's belief that p is *prima facie* justified iff the subject is non-luckily in possession of evidence for the belief that p and bases her belief that p on this evidence.

This embodies our two lessons. And of course, VE is structurally similar to the virtue-theoretic accounts of knowledge known from the works of virtue epistemologists like Ernest Sosa and John Greco.[5] One reaches a certain success in a reliable, competent way. In the case of knowledge, the success is the belief's being true. In the case of justification, it is the belief's being evidenced. Exercise of the competence explains the success feature. If one competently reaches an evidenced belief, the fact that there is some piece of evidence in its favor (i.e., the fact that the belief is evidenced) is not a mere coincidence or accident. The exercise of the relevant competence is what explains the success. Perception is a paradigmatic way of competent possession of evidence. What we see and hear is our salient evidence, in many contexts. We exercize perceptual abilities and thereby come to possess evidence. The evidence consists in the perceived entities and facts. They are evidence for beliefs about natural and artificial kinds, since they speak in favor of the truth of corresponding beliefs about these kinds, at least in normal or typical conditions.[6]

Considering a further objection can be helpful in order to avoid some misunderstandings. The objection is this one: Daniela behaves epistemically rational and well in forming her belief that there is a tomato in front of her. So her belief is justified – contrary to what has been claimed above. After all, Daniela treats her 'information' as evidence, and it is really evidence for what she comes to believe.

This objection has to be rejected. It rests on a confusion of (epistemic) rationality and (epistemic) justification. As already mentioned at the beginning, we are not working with a generic, unspecific concept of justification here. Rather we are concerned with the specific concept of justification that links a belief in some interesting, appropriate way to objective evidence. It has already been granted that Daniela proceeds rationally. But the rationality in question consists essentially in a way of responding to her perceptual experience (the perceptual hallucination). This is what she gets right and, therefore, there is something about her belief that is epistemically good. But in conformance with a quite general picture of rationality, rationality is a certain, good way of responding to 'information'.[7] Rationality does not cover the condition that needs to be satisfied by this 'information' – the right relation to the world. Even a brain in a vat could behave entirely rationally, by forming the (essentially)

---

[5] Here I rely on the so-called 'reliabilist' branch of virtue epistemology (for knowledge), not on the so-called 'responsibilist' branch (paradimatically proposed by Linda Zagzebski).

[6] One should point out that what the exercise of the ability explains is not just the formation of the belief but its having the relevant success feature. We are opting for '*strong* virtue epistemology', as Pritchard calls it. See Pritchard (forthcoming). Strong virtue epistemology is favored by Sosa (2007) and Greco (2010) in their accounts of knowledge. Here I sketch a strong virtue-theoretic account of justification. Strong virtue epistemology has the important advantage of providing an explanation of the surplus value of the achieved epistemic status. Reaching success from ability is more valuable than reaching it accidentally, *ceteris paribus*. For a discussion of some of the questions concerning the explanation of the surplus value of knowledge, see Hofmann (forthcoming). A strong virtue-theoretic account of justification promises an explanation of the surplus value of justified belief, as compared to unjustified belief. Quite strikingly, this kind of surplus value has not been discussed much in the recent literature. It is mentioned by some, for example, in David (2005), but the usual focus is on knowledge and its surplus value. The 'new Meno problem' (as one might call it), concerning justification, is largely unexplored.

[7] Numerous authors could be cited as favoring such a conception of rationality, for example, Parfit (1997).

same beliefs as we do in response to his or her perceptual experiences.[8] But the connection to the objective evidence would be missing. And this connection is essential for the concept of justification that we are dealing with here, i.e., the concept of reasons-related justification.

In other words, justification requires an appropriate relation to the evidence, whereas the function of rationality is 'merely' to make use of the evidence that one possesses in the form of suitable mental states (and it can be exercized even if one is not in possession of evidence at all). The intuition that there is something (epistemically) good about Daniela's belief can thus be fully captured by ascribing rationality (and possession of evidence). But it should not be described as concerning justification. Distinguishing justification from rationality is quite crucial for getting at the epistemologically important structures. The objection misdescribes the (correct) underlying intuition.

## 4. More on a Virtue-theoretic Account of Justification

We are heading toward a virtue-theoretic account of justification. VE lies at the heart of it. Let us now note some important features of the structure that underwrites justification. (All of the following is tentative and sketchy, of course.)

First of all, justification requires the possession of evidence, in the form of suitable mental states (such as genuine perceptions, for example).

Second, on the question of what possession of evidence consists in we can occupy an intermediate position between a Williamsonian conception on the one hand, and an internalist conception on the other hand. For a Williamsonian, possession of evidence consists in knowledge: 'E = K'.[9] But arguably, this is too high a requirement. One can possess evidence by simply *perceiving* the relevant facts, even in the absence of corresponding beliefs.[10] But this does not necessarily lead into an internalist conception of evidence possession, such as, for example, Richard Feldman's.[11] One need not think that possessing evidence is entirely determined by the subject's mental states. In fact, the virtue-theoretic account is robustly externalist, since objective evidence is required. A brain in a vat (having an inaccurate perceptual experience as of p) does not possess p as a piece of evidence if p is not the case. But still, one need not believe (and know) the evidence in order to possess it; perceiving it is good enough. (Knowing it is, of course, also good enough.) Thus an intermediate position is available which seems to have all of the advantages of the 'objective' Williamsonian, knowledge account without requiring the implausibly high standard of knowledge.

Third, justification exhibits the same structure of 'success from ability' as knowledge. (Again, the 'from' here is to be understood in a certain, strong sense – according to 'strong' virtue epistemology.) A belief is successful because of the exercise of the relevant competences. Arguably, these competences can be understood as (complex) dispositions.[12] What makes a disposition a competence (or 'epistemic virtue') is its reliability.[13] Having evidence in its favor takes the place of being true, the success feature in the case of knowledge. The structure is the same.[14]

---

[8] Of course, the brain in a vat could not form the very same *de-re* beliefs as we can, since the objects are missing.

[9] See Williamson (2000).

[10] Non-epistemic perception (i.e., perception without corresponding belief) is of course what is crucial here. The issue is not whether the content of perceptual states is (always) conceptual. What matters is the absence of belief. For some discussion of non-epistemic perception and Williamson's equation 'E=K' see Hofmann (ms).

[11] See Feldman (2005), for example.

[12] This is how Sosa and Greco understand competences, and I agree. See Sosa (2010), Greco (2010), 77.

[13] For some arguments about how this is to be understood see Hofmann (forthcoming).

[14] The structure of success-from-ability extends beyond doxastic matters and epistemology, and can be found in the agential realm. Sosa (2007) has emphasized the presence of this structure in artistic

Forth, the virtue-theoretic account can be extended in a quite natural way such as to take defeaters into account. (This is a major advantage over pure reliabilism.) Possession of counter-evidence makes for defeaters.

Fifth, the view counts as a form of *reliabilism*. Reliability lies at the heart of epistemic competences and 'virtues', as understood here. Truth conduciveness is guaranteed by the favoring relation: the evidence speaks in favor of the belief in question and thereby makes the belief likely to be true (in an objective sense of likelihood).

A lot more would have to be said in order to arrive at a full theory of justification. But it seems that we have an attractive beginning that deserves to be taken seriously. Intuitions about cases are one thing, and they are typically not universally agreed upon. But the theoretical perspective of a structure of 'success from virtue' is a further thing that seems to be so promising that one can hardly resist engaging with it.

**Frank Hofmann**

University of Luxembourg, department of philosophy
frank.hofmann@uni.lu

## References

David, M. 2005: 'Truth as the primary epistemic goal: a working hypothesis', in E. Sosa, M. Steup (eds.) 2005: *Contemporary Debates in Epistemology*, Oxford: Blackwell, 296-312.

Feldman, R. 2005: 'Justification is internal', in E. Sosa, M. Steup (eds.) 2005: *Contemporary Debates in Epistemology*, Oxford: Blackwell, 270-84.

Greco, J. 2010: *Achieving Knowledge*. Cambridge: Cambridge University Press.

Hofmann, F. forthcoming: 'Epistemic value and virtues', in T. Henning, D. Schweikard (eds.) 2013: *Knowledge, Virtue, and Action*, Oxford: Routledge.

Hofmann, F. 2013: 'Three kinds of reliabilism', *Philosophical Explorations* 16(1), 1-22.

Hofmann, F. ms: 'E=K and non-epistemic perception', manuscript, 2012.

Lewis, D. 1980: 'Veridical hallucination and prosthetic vision', *Australasian Journal of Philosophy* 58, 239-49.

Mantel, S. 2012: 'Acting for reasons, apt action, and knowledge', *Synthese* DOI: 10.1007/s11229-012-0230-8.

Parfit, D. 1997: 'Reasons and motivation', *Proceedings of the Aristotelian Society Suppl. Vol.* 71, 99-130.

Pritchard, D. forthcoming: 'Anti-luck virtue epistemology', *The Journal of Philosophy*.

Pritchard, D. 2005: *Epistemic Luck*. Oxford: Clarendon Press.

Sosa, E. 2010: 'How competence matters in epistemology', *Philosophical Perspectives* 24, 465-75.

Sosa, E. 2007: *A Virtue Epistemology: Apt Belief and Reflective Knowledge, Vol. 1*. Oxford: Oxford University Press.

Williamson, T. 2000: *Knowledge and Its Limits*. Oxford: Oxford University Press.

---

performances and the sports. For an account of acting for a reason in terms of 'apt action' see Mantel (2012).

# Contextualism and Gradability – A Reply to Stanley

Romy Jaster

Contextualism in epistemology is the claim that the knowledge predicate is contextsensitive in the sense that it has different truth conditions across different contexts of use. Jason Stanley objects against this view that if it were correct, then "know" should be gradable in the same way as gradable adjectives. Since it lacks gradability, it also lacks the postulated contextsensitivity. Or so Stanley argues. In this paper, I show that the contextualist is not committed to the gradability of the knowledge predicate in the first place. I will distinguish between what I will call *pure threshold predicates*, which either apply *simpliciter* or not at all in each context, and *impure threshold predicates*, for which context determines whether they apply *simpliciter*, but which can also be satisfied to certain degrees. Threshold predicates are not gradable, but many of exhibit just the kind of contextsensitivity that is postulated for "know". Pace Stanley, three claims are going to be established: that the lack of gradability of the knowledge predicate (i) does not jeopardize its contextsensitivity, (ii) does not dismantle the analogies contextualists have claimed to hold between "know" and gradable adjectives, and (iii) is perfectly consistent with the idea of varyingly high epistemic standards.

Contextualism in epistemology (henceforth "contextualism") is the view that knowledge claims – paradigmatically statements of the form "S knows that p" or "S doesn't know that p" – are contextsensitive in the sense that they have different truth conditions across different contexts of use, depending on the epistemic standards obtaining within each context. As a result, "S knows that p" as uttered in a context with low standards does not express the same proposition as "S knows that p" as uttered in a context with high standards.[1] Witness Keith DeRose:

> [T]he truth conditions of knowledge-ascribing and knowledge-denying sentences (...) vary in certain ways according to the context in which they are uttered. What so varies is the epistemic standards that a subject must meet (or, in the case of a denial of knowledge, fail to meet) in order for such a statement to be true. In some contexts, "S knows that P" requires for its truth that S have a true belief that P and also be in a very strong epistemic position with respect to P, while in other contexts, the very same sentence may require for its truth, in addition to S's having a true belief that P, only that S meet some lower epistemic standards. Thus, the contextualist will allow that one speaker can truthfully say "S knows that P", while another speaker, in a different context, where higher standards are in place, can truthfully say "S doesn't know that P", though both speakers are talking about the same S and the same P at the same time. (DeRose, 2000: 91)

Contextualists have often emphasized that these features of the semantics of "know" are not unique to the knowledge predicate. Rather, they can also be found in gradable adjectives: the truth conditions of "S is tall" vary across contexts with varying standards for tallness. As a result, "S is tall" can be true in one context and false in another. The analogy between "know"

[1] There are versions of contextualism which do not involve a commitment to higher and lower standards. As Stanley himself points out, these views are immune to his objection.

and gradable adjectives has thus often served as an important motivation and illustration of contextualism.

Jason Stanley argues that this line of thought spells trouble for contextualism. Specifically, he argues that if the knowledge predicate were in fact contextsensitive in the postulated way then it should also be analogous to gradable adjectives when it comes to gradability; like "tall" and "flat", it should exhibit the property of being gradable. Since it does not exhibit this property, contextualism fails. Let's have a closer look at the line of thought Stanley develops. He writes:

> According to [contextualists], knowledge ascriptions come in varying degrees of strength. In other words, knowledge ascriptions are intuitively gradable. Contextualists speak (...) of higher and lower standards for knowledge. Comparative adjectives are one natural kind of gradable expressions. It is therefore no surprise that epistemologists (...) have been exploiting the analogy between "know" and adjectives such as "flat" and "tall". But (...) the attempt to treat "know" as a gradable expression fails. First, it shows that one cannot appeal to the context-sensitivity of adjectives to justify the context-dependence of knowledge ascriptions. Secondly, it casts doubt upon the claim that knowledge comes in varying degrees of strength (...). (Stanley, 2005: 35f.)

In this passage, Stanley indicates that contextualists are somehow committed to the claim that knowledge ascriptions are gradable. But this claim is false, according to Stanley. Its falsity can be shown by applying two tests for gradability to the case of the knowledge predicate.

> First, if an expression is gradable, it should allow for modifiers. For example, predicative uses of gradable adjectives allow for modification, as in:
>
> (1) (a) That is *very* flat.
>
> (b) That is *really* flat.
>
> (...) Secondly, if an expression is gradable, it should be conceptually related to a natural comparative construction. So, for "flat" "tall", and "small" we have "flatter than", "taller than", and "smaller than". (ibid.: 36)

Stanley then argues that "know" fails both of these tests; constructions such as "I don't really know it" (modification) or "I know it better" (comparative) and are not to be taken literally. Genuine modifier uses allow for constructions like "This is flat, but not really flat". Genuine comparatives allow for constructions like "I am tall, but you are taller". None of these constructions are available in the case of "know": "I know it, but I don't really know it" and "I know it, but you know it better" seem infelicitous.

On the basis of these findings, I think one should grant that "know" is indeed not gradable. Where Stanley goes astray is in thinking that this finding casts doubt upon contextualism as such. In fact, as I am going to show, neither the analogies that are said to obtain between gradable adjectives and "know" nor the idea of varying epistemic standards commit contextualists to the gradability of "know" in any way.

What contextualists say is that in different contexts different standards for what counts as knowledge obtain. In some contexts they are high, in others they are low. That is, in some contexts, the epistemic position a subject must be in to count as knowing must be stronger than in others. As a consequence, the truth conditions of knowledge sentences vary. In *these* respects, "know" works just like gradable adjectives.

But from this it does not follow that "know" itself should be gradable. For the very same analogies contextualists emphasize between "know" and gradable adjectives hold between "know" and clearly non-gradable, yet obviously contextsensitive expressions, such as "tall enough" or "sufficiently flat".

All the crucial things the contextualist says about "know" can be said about these expressions as well. In different contexts, different standards for counting as sufficiently flat or tall enough obtain. In some contexts they are high, in others they are low. If, for instance, one tries to decide whether a certain lawn can be used for a boccia game, the standards for sufficient flatness are rather lax. If physicists try to decide whether a certain surface can be used for an experiment, the standards may be much higher. The same goes for "tall enough". A rather high standard obtains in conversations about reaching the ceiling, and a rather low standard obtains in conversations about reaching the windowsill. That is, in different context, a subject must have different heights to count as tall enough and a surface must have varyingly few bumps in order to count as sufficiently flat. Consequently, the truth conditions of "x is sufficiently flat" and "S is tall enough" vary across contexts. But neither "sufficiently flat" nor "tall enough" is gradable.

If this is true, however, and the analogies contextualists have emphasized between "know" and gradable adjectives also hold for clearly non-gradable contextsensitive expressions like "tall enough" and "flat enough" then gradability does not seem to be a relevant feature in this analogy. It seems to be an independent feature of some expressions, which are in other respects analogous to the predicate "know".

Why does Stanley think that contextualists are committed to the gradability of "know" in the first place then? Looking for Stanley's reasons is instructive; I take it that underlying his gradability objection is a misled understanding of contextualism itself. Invoking certain remarks by Stewart Cohen and Keith DeRose, Stanley tells us that it is a *core claim* of contextualism that knowledge comes in varying degrees of strength. Since Stanley does not give us bibliographical references with respect to these remarks, it is hard to see whether Cohen and DeRose have actually formulated their view in these words. If so, they have chosen a very misleading way of doing so. It is quite natural to expect a predicate to be gradable if we can truly state that the property it picks out comes in varying degrees of strength. If a core claim of contextualism really were that knowledge comes in varying degrees of strength it would therefore indeed be surprising if "know" were not gradable.

Luckily, contextualists are not committed to what Stanley takes to be their core claim. According to contextualists, a subject has to meet varyingly high standards to *count as knowing*. This is a metalinguistic claim about the application conditions of the knowledge predicate. These *standards* are gradable and so are the epistemic positions a subject must be in to meet these standards: both can be lined up along a scale in ascending order; we can line up standards from low to high and epistemic positions from weak to strong. In each case, both of Stanley's tests for gradability – modifier use and comparative construction – are satisfied: we can say that a standard is high, but not *really* high, or that someone is in a strong epistemic position, but not in a *really* strong epistemic position. Likewise, we can say that a given standard is *higher* than some other standard, or that someone is in a *better* epistemic position with respect to p than someone else.

In spite of these gradable elements, however, it does not make any more sense to say that knowledge comes in varying degrees of strength than to say that being tall enough or being sufficiently flat come in degrees. There is a gradable element to these predicates as well: relative to different standards, different heights count as tall and different degrees of flatness count as flat. We can arrange heights and degrees of flatness along a scale. And in a *derived* sense we can line up different instances of being tall enough or being sufficiently flat along a scale as well. We can say that the property of being tall enough is satisfied to a higher degree the taller the subject is. And we can say that the property of being sufficiently flat is satisfied to a higher degree the flatter the subject is.

In the same *derived* sense we can grade knowledge relations. We can say that they are realized to a higher degree the stronger the required epistemic position is. But we should keep in mind that this is just a very loose formulation of what is really claimed. In speaking this

way, we are not really grading "tall enough", "sufficiently flat" or "know". What we are grading are heights, degrees of flatness and epistemic positions, things that have to reach a certain degree for "tall enough", "sufficiently flat", or "know" to apply *simpliciter*. Strictly speaking, being tall enough and being sufficiently flat do not come in degrees. Neither does knowledge. We therefore should not expect "know" to pass Stanley's tests for gradability.

The upshot thus far is that in certain respects, "know" functions like *modified* gradable adjectives such as "tall enough" or "sufficiently flat" rather than functioning like gradable adjectives such as "tall" or "flat".[2] It is what I would like to call a *pure threshold predicate*: In each context, the obtaining standard fixes the conditions that have to be satisfied for a subject to exceed the threshold. Once the threshold is exceeded, the predicate applies. This is the only purpose the predicate has. It marks whether the threshold is surpassed or not. So the predicate is either satisfied *simpliciter* or not at all. It cannot be satisfied to certain degrees.

Gradable adjectives are more complex. They are *threshold predicates*, but *impure* ones. Context determines the standard that has to be met for them to apply *simpliciter*, but in contrast to pure threshold predicates they can also be satisfied to certain degrees. That is why we can speak of one person being taller than another, thereby indicating that the first person satisfies the tallness predicate to a higher degree than the second one. Pure threshold predicates do not admit of that: we cannot speak of one person being more tall enough or knowing something to a higher degree than another person.[3]

Stanley thinks that this line of thought is thoroughly mistaken and offers a brief argument against it in his book. In the remainder of this paper, I will argue that his objection is severely flawed in a variety of ways. Here is what he says:

> One reaction (...) is to maintain that I have focused on the wrong model (...). Instead of "know" being analogous to "flat" or "tall", the contextualist claim is rather that "know" is analogous to "flat *enough*" or "tall *enough*". (...) It is not clear to me in what sense "know" is supposed to be analogous to "tall enough" (...) or even "justified enough". There are all sorts of disanalogies (...). Most alarmingly, one standard use of these expressions is to convey that something has the property for a sufficient degree for present purposes, though it does not in fact have the property. (Stanely 2005: 43)

In accordance to this standard use, Stanley argues, we can felicitously say things like the following:

> (23) He isn't tall, but he's tall enough. (ibid.: 44)
>
> (24) I may not be justified in my suspicion, but I'm justified enough to investigate further. (ibid.)

He then goes on to argue:

> If "know" is supposed to be synonymous with something like "is justified enough in one's true belief" then (...) one would expect to be able smoothly to say things like:
>
> (25) John isn't justified in his belief that the bank is open, but he knows that the bank is open. (ibid.)

Since this is infelicitous, Stanley concludes that

> 'know that p' simply doesn't behave as 'is justified enough in one's true belief that p'. These disanalogies are sufficient to undermine the plausibility of the proposal. (ibid.)

In this passage, Stanley makes a variety of very dubious moves, the most dubious of which is that he distorts his opponent's proposal considerably in the course of the passage I have just

---

[2] This point is also made by Halliday (2007).
[3] Halliday (2007: 390) makes a very similar distinction in terms of two senses of "gradability".

quoted. The proposal he explicitly considers at first – and which is in fact what I (and Halliday 2007) have argued for – is that know is in many ways *analogous* to "tall enough", "flat enough" or – for all I care – "justified enough". This claim is then distorted to the claim that "'know' is *synonymous* with something like 'is justified enough in one's true belief'", which is a very different claim. Synonymity requires substitutability. Analogies don't. That "is justified enough" cannot be substituted by "knows" in the exemplary sentences is therefore not decisive at all against the proposal at issue.

The important question is: is the infelicity of (25) problematic for the actual proposal, according to which there are important *analogies* between "know" and "tall enough"? Not as far as I can see. The claim that there are important, or even crucial, analogies between two things in *some* respects can hardly be criticized on the basis of the finding that there are disanalogies between the two things in *other* respects. The proposal at issue is that gradability is not to be expected in the case of "know" because "know" - like "tall enough" - is a pure threshold predicate. This does not imply that there are no *dis*analogies between the terms. It is therefore strange that "[i]t is not clear to [Stanley] in what sense 'know' is supposed to be analogous to 'tall enough'" solely on the basis of the fact that "[t]here are all sorts of disanalogies".

Nevertheless, there is one very legitimate worry left to answer. I am not sure whether this worry is actually what guides Stanley's criticism, but I take it to be the most promising way to attack the view I have presented: the worry is that the disanalogies that can be observed between "know" on the one hand and "tall enough" and "sufficiently flat" on the other arise from the very property that I have postulated to be analogous between them. If that were the case – if what I have claimed to be analogous between "know" and "tall enough" indeed gave rise to a behaviour of "tall enough" that cannot also be observed in the case of "know" - then, of course, this would strongly count against the postulated analogy between them. As I will show, however, this is not the case.

The property I have identified as analogous between "know" and "tall enough" is that both are pure threshold predicates. That means that the predicate either applies *simpliciter* or not at all, and whether or not it applies depends on whether or not a contextually determined threshold for its application is surpassed or not. Hence, the question we need to turn to is whether *their being pure threshold predicates* gives rise to the fact that "tall enough", "sufficiently flat", and "justified enough" can be combined with a denial of "tall", "flat", or "justified" - a denial, that is, of their underlying unmodified predicates?

The answer is straightforward: there are lots of pure threshold predicates whose attribution cannot felicitously be combined with a denial of their underlying unmodified predicates. Just consider "*very* tall". It clearly is a pure threshold predicate, but nevertheless it is infelicitous to say "He's very tall, but he's not tall". The same goes for "absolutely flat", and "highly justified". It also goes for covertly modified predicates such as "excellent" and "vanished", which I take to be roughly equivalent to "extremely good" and "completely gone", respectively. In all of these cases, the context determines a threshold above which the predicates apply. None of them is gradable, and none of them can be felicitously combined with a denial of their underlying unmodified predicates. If something is absolutely flat, it is also flat. If something is highly justified, it is also justified. If something is excellent, it is also good. If something is vanished, it is also gone.

Now, if the scale on which "know" marks a certain threshold is a scale of better or worse epistemic statuses – or of higher or lower justification – but "know" requires an excellent, very good, absolute, extremely good or simply a high level of justification, then it is not to be expected that "know" can occur in statements such as "He knows, but he's not justified". The threshold for "know" can never be lower than the threshold for "justified", just as the threshold for "very tall" can never be lower than the threshold for "tall". In this respect, pure threshold predicates can differ: the threshold for "tall enough" can be – and often is – lower

than the threshold for "tall", as the felicity of Stanley's statement (23) shows. But this is not a consequence of "tall enough" being a pure threshold predicate. Rather, it is once more an independent feature of *some* pure threshold predicates. I therefore conclude that the disanalogy Stanley observes between "know" and "tall enough" does not arise from the feature that I have identified as analogous between the two. As a consequence, the disanalogy does not cast doubt upon the claim that they have this feature in common.

What my discussion has shown is this: Stanley is right. "Know" is not gradable, whereas some contextsensitive adjectives such as "tall" are. But none of the consequences Stanley draws from this finding follow from it. First, and most importantly, it does not follow that "know" is not a context-sensitive term. As I have shown, the contextualist is not committed to the gradability of "know" in the first place. Secondly, Stanley's finding that "know" is not gradable does not threaten the analogy between "know" and gradable adjectives. It shows that they are not analogous with respect to gradability, all right. But this is perfectly compatible with the line of thought typically endorsed by contextualists. The analogy they stress is the sensitivity to standards which obtains in both cases. And this sensitivity to standards has nothing to do with gradability, as clearly non-gradable but nonetheless standard-sensitive expressions – pure threshold predicates such as "tall enough", "very flat", "highly justified", and "excellent" - show. Third, and finally, Stanley's finding does not cast doubt upon a core claim of contextualism. It might cast doubt upon the claim that knowledge comes in degrees. But this claim is not more than a distortion of an essential thesis of contextualism. It is a loose and misleading formulation of the claim that *what is required for"know" to apply* comes in degrees. This is what contextualists are committed to. But as other cases of pure threshold predicates show, this does not commit them to the gradability of "know" either.

**Romy Jaster**

Humboldt-Universität zu Berlin
jasterro@hu-berlin.de

## References

DeRose, K. 2000: 'Now You Know It, Now You Don't', *Proceedings of the Twentieth World Congress of Philosophy* (Philosophy Documentation Center) Vol. V, Epistemology, 91-106.

Halliday, D. 2007: 'Contextualism, Comparatives, and Gradability', *Philosophical Studies* 132, 381–393.

Stanley, J. 2005: *Knowledge and Practical Interests.* New York and Oxford: Oxford University Press.

# Intuitions, Heuristics, and Metaphors: Extending Cognitive Epistemology

Eugen Fischer

Psychological explanations of philosophical intuitions can help us assess our warrant for accepting them. To explain and assess conceptual or classificatory intuitions about specific situations, some philosophers have suggested explanations that invoke heuristic rules proposed by cognitive psychologists. This approach offers a promising alternative to the standard approach of experimental philosophy. The present paper develops this alternative in fresh directions: It motivates the proposal of a fresh heuristic, and shows that this heuristic can explain a class of influential intuitions that have been neglected in current debates in the epistemology of philosophy. By integrating results from two hitherto disconnected strands of psychological research, on intuitive judgment and on analogy and metaphor, respectively, the paper motivates the proposal of a 'metaphor heuristic'. Second, it shows that this heuristic can explain general factual intuitions influential in the philosophies of mind and perception. The paper shows that the proposed heuristic satisfies the key requirements imposed by cognitive psychologists in the relevant research traditions, and that explanations employing this new heuristic can reveal whether particular philosophical intuitions are due to the proper exercise of cognitive competencies or constitutive of cognitive illusions.

This paper will develop an approach that forms part of a research program we can helpfully dub 'cognitive epistemology'. This is a kind of naturalised epistemology, which shares a central ambition with experimental philosophy: It seeks to develop psychological explanations of philosophically relevant intuitions that help us assess philosophers' warrant for accepting them (cp. Knobe and Nichols 2008: 8). Cognitive epistemology pursues this aim by drawing on experiments and theories already available from cognitive and social psychology. Work in this nascent tradition seeks to explain – and assess – philosophically relevant intuitions as the results of cognitive processes for which psychologists have already provided experimental evidence (Fischer 2011, Gerken 2011, Nagel 2010, 2011, Spicer 2007).

One approach that has been tentatively tried (e.g. by Hawthorne 2004 and Williamson 2005) is to explain some relevant intuitions by reference to heuristic rules posited by cognitive psychologists working within the 'heuristics and biases program' (Tversky and Kahneman 1974, Kahneman and Frederick 2005, Kahneman 2011). This may reveal that compelling intuitions are due to seductive fallacies and can be disregarded. The moment we pool the conceptual resources of this research program with that of the 'adaptive behaviour and cognition' program (Gigerenzer and Todd 1999, Gigerenzer 2008), also from cognitive psychology, we can explain intuitions and assess their evidentiary value more widely, both when they lack and when they possess such value. Crucially, we can do this without – controversial (Cappelen 2012, Williamson 2007) – recourse to conflict, sensitivity, or instability results from surveys of the kind experimental philosophers tend to conduct (cp. Alexander and Weinberg 2007).

This paper will develop the approach of intuition assessment through heuristics-based explanation, and present two extensions of it: the proposal of a new heuristic, the metaphor heuristic, and its application in the explanation and assessment of a philosophically important class of intuitions that has been neglected in current debates. Debates in the epistemology of philosophy have focused on conceptual, classificatory, or modal intuitions

mainly about specific situations presented in thought experiments (e.g. Gettier cases). By contrast, I would like to focus on general factual intuitions about how things actually are (e.g. about the workings of the mind and perception). In various areas of the subject, philosophers have accepted such intuitions without much, if any, argument, even when these intuitions clash with common-sense convictions or among each other. Such conflicts are at the bottom of various philosophical problems, including familiar problems from the philosophies of mind and perception (Fischer 2011).

This paper will present the approach of assessing intuitions through heuristics-based explanation (section 1), build up to the new metaphor heuristic (section 2) that can explain influential general factual intuitions of the kind freshly targeted (section 3), and show that such explanations can expose the intuitions explained as cognitive illusions (section 4) – and thus help resolve philosophical problems these intuitions appeared to raise.

## 1. Assessing and Explaining Intuitions through Heuristics

First, what do we seek to explain (and assess)? In philosophy as in psychology, intuition is contrasted with deliberate reflection. Practically all psychological research on intuitions conceptualises these as a kind of judgments – which we may, but need not, be entitled to accept and which may, but need not, provide evidential support for other judgments. In two respects, intuitive are like perceptual judgments, though they do not involve the use of our five senses in anything like the same way: We (i) do not control the processes that give rise to intuitive judgments (Mercier and Sperber 2009), and (ii) we are not even conscious of those processes but only of the judgments in which they issue (Sloman 1996). When we focus (like Gigerenzer 2007: 16) on judgments 'strong' or compelling 'enough' for thinkers 'to act upon', in deed, word, or thought, we are talking about intuitions: non-perceptual judgments thinkers (i) make spontaneously, (ii) without being aware of making any inference or rehearsing any reasoning, and (iii) find plausible or compelling.

Intuition research in psychology is dominated by two partially complementary, partially competing programs: the 'heuristics and biases program' (pioneered by Tversky and Kahneman 1974) and the 'fast and frugal heuristics' or 'adaptive behaviour and cognition' (ABC) program (pioneered by Gigerenzer et al. 1999). Both programs seek to explain intuitions as the result of largely automatic application of heuristic rules. While *normative rules* (of logic, probability theory, morality) define, determine, or constrain which answer, solution, decision, or action is right or reasonable, *heuristic rules* are rules of thumb which typically yield reasonably accurate results in certain ranges of application, but do not define or constrain what is right or reasonable. While we employ many such rules in explicit reasoning, they also govern automatic cognition (Gigerenzer 2008).

Automaticity is a complex notion which has been tied to several independent features that are individually gradable, not dichotomic (Bargh 1994, Moors and De Houwer 2006). A cognitive process is

- *effortless* to the extent to which its execution requires no attention or other limited cognitive resource – and is hence not impaired when the subject is distracted by tasks requiring such resources (e.g. keeping in mind long numbers),
- *unconscious* to the extent to which the subject is unable to report the course of the process, as opposed to articulating its outcome (judgment, decision, etc.),
- *non-intentional* to the extent to which the process is initiated regardless of whether or not the subject wants to, namely regardless of what aims or goals she pursues,
- *autonomous* or '*uncontrolled*' iff once the process has started, the subject cannot alter its course or terminate it before it has run its course.

Effortless processes also tend to be executed more rapidly than effortful processes. Effortlessness is widely used for an operational definition of 'automatic' processes, more generally (Evans 2008). In social psychology, cognitive processes that are rapid, effortless, unconscious, and non-intentional are called '*spontaneous*' (Uleman et al. 2008).

Many fully automatic processes get routinely or continually carried out in sense-perception or language-comprehension, even in the absence of specific tasks set. These are known as *natural processes* (Tversky and Kahneman 1983). For example: In a modest sense of the term, 'recognition' is automatically assessed by a natural process – when hearing a name or seeing a person we cannot help having a gut feeling about whether we have encountered this name or person before (regardless of whether we recall where and when). Judgment heuristics are simple strategies for obtaining answers to a variety of questions (Which of two cities is bigger? Which of two players is more successful? etc.) from such 'natural' or 'basic assessments' that are generated by natural processes (Tversky and Kahneman 1983, Kahneman 2011).

Judgment heuristics work where the outcomes of natural processes are correlated with true answers to a question. By and large, bigger cities and more successful athletes receive more media attention, so we are more likely to have encountered their names before. Hence, in game shows and elsewhere, it makes good sense to employ this *recognition heuristic* (here in a simple version for paired choices): If you recognise one of two objects (cities, athletes, etc.) but not the other, judge that the recognised object has the higher value (is bigger, more successful, etc.) (Goldstein and Gigerenzer 2002).

To support the hypothesis that thinkers actually use a particular such strategy in largely automatic – and unconscious – cognition, psychologists derive from it surprising fallacies or effects, and reproduce these in behavioural experiments. The recognition heuristic, e.g., predicts that under certain conditions subjects who possess less relevant information than others will make more correct judgments, and such less-is-more effects were reproduced in striking experiments (op. cit.).

Chronometric and multi-tasking studies suggest the application of judgment heuristics is effortless and rapid (De Neys 2006, Pachur and Hertwig 2006, Volz et al. 2006): Apparently participants first apply the pertinent heuristic in an effortless and rapid process, and some of them then correct the outcome, where necessary, in a further, effortful step which takes a few seconds and falls by the wayside under the pressures of multi-tasking (Kahneman and Frederick 2005, Kahneman 2011).

Intuitive judgments are frequently in need of justification, e.g., when they are inconsistent with common-sense-convictions, scientific findings, or other intuitions (of the same or other subjects). As I have shown in detail elsewhere, philosophers often accept such paradoxical or otherwise justification-needy intuitions without argument, or at any rate without any non-circular argument (Fischer 2011). Where this happens, they are justified in accepting an intuition precisely to the extent to which this intuition has *probative force*, i.e., precisely to the extent to which the mere fact that a given thinker has this intuition speaks for its truth.

Heuristic-based explanations of intuitive judgments help us determine whether a thinker's intuitions have such force. Negatively, such an explanation can reveal that the intuition explained is constitutive of a cognitive illusion – which has no probative force. Positively, it can reveal that the intuition explained is due to the exercise of an 'epistemic virtue'. One prominent proposal conceives of such virtues as cognitive competencies 'to discriminate the true from the false reliably (enough) in some subfield of ... propositional contents' (Sosa 2007: 58). The following proposal allows for epistemic virtues that equip us for many but not all relevant judgments, and are manifest in responses beyond mere true-false judgments: To possess an epistemic virtue, I suggest, is to be competent to offer true or accurate judgments

implying truth, falsity, accuracy, or inaccuracy of propositional contents in such a subfield, for a significant proportion of such contents.

A cognitive competence can consist in mastery of a rule. To what extent mastery of a heuristic rule (like the recognition heuristic) renders us competent to make judgments of criterion (say, relative size) for items (say, cities) depends upon two things. First, how often can we apply the heuristic, e.g., how often is its cue available to us? (How often do we recognise precisely one member within pairs from the reference class?) Second, in how many of these cases does the heuristic let us get things right? (How often is the one item I recognise indeed bigger?) The proportion of true to overall judgments a subject can obtain by applying the heuristic to the given reference class is known as its *ecological validity* for that class (Gigerenzer and Sturm 2012). Comprehensive mastery of such a rule can be constitutive of an epistemic virtue, namely, to the extent to which the heuristic is applicable for the subject to the domain of a given subfield of propositional contents, to the extent to which the heuristic is ecologically valid in the reference class provided by that domain, and to the extent to which the subject's spontaneous application of the rule is sensitive to its differing ecological validity for different reference classes (cp. Pohl 2006).

This takes us to our second, negative, possibility: Illusions are predictable and reproducible deviations of perceptions, judgments, or memories, from relevant facts or normative standards. In *cognitive illusions*, (i) thinkers make spontaneous judgments violating relevant normative rules (e.g. of logic or probability theory); (ii) thinkers do so in a predictable, rather than random fashion; while these misjudgments can be modified and even completely corrected by conscious reflection, (iii) they are automatic or involuntary in origin and (iv) subjects typically find them intuitively compelling even once they have realised they cannot be right (Pohl 2004, 2-3). The spontaneous application of heuristic rules generates intuitions that are cognitive illusions where they predictably lead to judgments that are not merely false but violate normative rules. The probably best known example is the conjunction fallacy engendered by the representativeness heuristic proposed by Kahneman and Tversky and apparently reproduced by their famous Linda study (Tversky and Kahneman 1983).

By and large, intuitions that issue from a cognitive competence have probative force, and intuitions constitutive of cognitive illusions do not. This, however, is no neat dichotomy: Just as competent observers who can typically make correct visual judgments of relative length are liable to fall prey to the Müller-Lyer illusion even so, a thinker with comprehensive mastery of a heuristic rule is liable to fall prey to specific cognitive illusions engendered by its application or the interference of other automatic processes. Intuition assessment through heuristic-based explanation seeks to identify generally reliable cognitive competencies and specific cognitive illusions even competent thinkers may fall for, to assess precisely where our philosophically relevant intuitions have probative force. We shall now further develop this approach, for starters by building up to a new heuristic, the metaphor heuristic.

## 2. The Metaphor Heuristic

Metaphorical reasoning is a kind of analogical reasoning (as when physicists think of atoms as analogous to solar systems). According to all major psychological theories of analogical reasoning, such reasoning about a target domain (say, atoms) involves at least three steps (Holyoak 2012): First, a suitable source-model (e.g. the solar system) is identified and knowledge about it is retrieved from memory. Second, model and target are aligned and elements of the source-model (planets, sun, relations between them) are mapped onto elements of the target domain (electrons, nucleus, etc.). Third, a procedure known (since Holyoak et al. 1994) as 'copying with substitution and generation' is used to obtain conclusions about the target domain from familiar premises about the source.

The first two steps are subject to the two constraints of semantic and structural similarity (Gentner et al. 1993). To simplify greatly, we are most likely to retrieve knowledge about source domains to which we apply some of the same terms (Wharton et al. 1996), and will then first map relations in the source-domain on their namesakes in the target-domain. This initial mapping then gets expanded and knocked into shape by enforcing one-to-one mapping and parallel connectivity: when mapping relations and properties, you also need to map their relata and arguments (e.g. Falkenhainer et al. 1989, Forbus et al. 1995).

A mapping that satisfies these constraints brings out a relational structure which source and target domain share. The actual inference then proceeds by spontaneous completion of this pattern:

**Copying with substitution and generation (CSG):**

1. [*Copying*] When a source-domain element A (individual, property, or relation) has been mapped onto a target-domain element B, all the relations in which A stands in the source domain, plus relevant relata, are transferred onto B, into the target domain.

2. [*Substitution*] As far as possible, relations and relata from the source domain are replaced by those elements of the target domain onto which they have been mapped.

3. [*Generation*] Where an element of the source domain could not be mapped onto any element of the target domain, it is simply carried over identically, and new elements are postulated in the target domain, to the extent necessary to complete the transferred structure.

Such analogical reasoning may lead to the metaphorical extension of terms, e.g., from the comparatively concrete source-domain of visual search or manual operation to the rather more abstract target-domain of goal-directed intellectual effort (cp. Jäkel 1995). About the intellectual achievement of understanding somebody's action, e.g., we can say such things as:

> It is *clear* to me why you acted that way, when I manage to *see* your reasons – and *obscure* when I fail. I may *look for* reasons where these are *hidden* or *be blind to* reasons *in plain view*. An *illuminating* explanation *throws new light* on your action and lets me *discern* reasons I had previously *overlooked*, *get a fuller picture*, or at least *catch some glimpse*, of reasons about which I previously *was completely in the dark*.

The same inferential relations between the involved terms obtain both in the concrete source- and the more abstract target-domain. E.g., regardless of whether I am looking for my keys or your reasons, I cannot see what is lying in the dark, and something that sheds light may help me see it.

The underlying analogical reasoning is highly systematic. It employs a series of related mappings unfolded by the most elementary analogical inferences from 'basic mappings' such as the mapping of seeing onto knowing:

(1) S sees x $\rightarrow$ S knows x

The first mappings made in analogical reasoning correlate elements of the different domains, to which the same concepts apply (Gentner et al. 1993). Prior to metaphorical extension, they hence map attributes and relations that obtain in both source- and target-domain onto themselves; according to the rules of CSG, these attributes and relations get 'substituted' by themselves (and are, in effect, simply copied). The most *elementary CSG inferences* involve only copying with substitution, no generation, and employ only such 'mappings onto self' plus a basic mapping like (1), while (non-relational) logical and modal operators, which also pertain to both domains, simply get copied.

Such elementary CS inferences can proceed from either closed or open sentences. When they proceed from open sentences, attributes and relations designated by the premises can be mapped onto those designated by the conclusions, and this yields further mappings:

(2) S does not see x → S does not know x

(3) It is possible for S to see x, i.e., x is visible for S → It is possible for S to know x

(4) It is not possible for S to see x, i.e., x is invisible for S → It is not possible for S to know x

(5) X makes it possible for S to see y, i.e., x makes y visible for S → X makes it possible for S to know y

(6) X makes it impossible for S to see y, i.e., x makes y invisible to S → X makes it impossible for S to know y

(7) S tries to get to see x, i.e., S looks for x → S tries to get to know x

The basic mapping and the mappings obtainable from it through elementary CS inferences are jointly *constitutive of* a *conceptual metaphor* – like the present metaphor *Intellectual Effort as Visual Search* (Fischer 2011: 22-28 and 41-49).

Conceptual metaphors are realised in pictures and language (Lakoff 1993). In language, they facilitate the systematic metaphorical extension of terms from their source to their target domains. Such extension is motivated by implications: We frequently say things by implying them without stating them. For example, we associate attributes like strength, courage, and nobility with lions, and reliably infer from 'x is a lion' that x is strong, courageous, or noble. Hence the predicate came to be metaphorically extended from animals to heroes and humans, where 'Achilles is a lion' means that the hero is strong, courageous, or noble. Metaphorical extension of relational terms to their salient implications can motivate the basic mappings of conceptual metaphors: Typically, when you see something happening, you know it happens.

Conceptual metaphors facilitate a plethora of CSG inferences which forge fresh implications that motivate metaphorical extension of further terms. The relevant inferences proceed from conditional open sentences about the source domain, generate or carry over the antecedent, and substitute the consequent. They give source-domain terms (which remain in the antecedent of the conclusion) metaphorical implications about the target domain, which are specified by the substituting terms in the consequent. Take a familiar fact about the present source domain of visual search, stated by using the term 'obvious' in the literal sense in use until the 18th century: 'standing in the way, positioned in front of, opposite to, facing' (OED). When things are obvious in this sense, when they stand right in front of us, they are, by and large, easily visible. Straightforward CSG inference takes us from this conditional premise to the conclusion that when things are 'obvious', they are easily knowable. All we need is an elementary CS inference from the basic mapping to the mapping (3*): It is readily possible for S to see x (x 'is easily visible for S') → It is readily possible for S to know x. One can conveniently represent such CSG inferences through 'grids':

| | **Source-domain premise** | **Operation** | **Target-domain conclusion** |
|---|---|---|---|
| 1 | X is obvious | Generation | X is obvious |
| 2 | implies (1;3) | Substitution: identical | implies (1;3) |
| 3 | X is easily visible | Substitution: mapping 3* | X is easily knowable |

This inference endows 'obvious' with a metaphorical implication in the intellectual target-domain, which lets us say that the point I have been belabouring here is really pretty obvious. In these ways, conventional and metaphorical implications motivate the metaphorical

extension of terms, in talk of heroes and solutions, respectively. Through such extension, conceptual metaphors become realised in a language.

This two-page crash course on analogy and metaphor has familiarised us with the ingredients needed to construct a heuristic. To appreciate this, we need to briefly turn to experiments on so-called 'metaphor consistency effects' (e.g. Gentner et al. 2002): Where texts first employ metaphorical expressions realising one conceptual metaphor and then suddenly a metaphorical expression realising another, reading time for the new 'inconsistent' metaphorical expression increases. The experimenters interpreted this as evidence of largely automatic 'non-intentional analogical reasoning' (Day and Gentner 2007) employing the second conceptual metaphor. Such evidence has been found for freshly coined metaphorical expressions as well as for perfectly conventional spatial time metaphors (Gentner and Bowdle 2008). This has one crucial implication: CSG inferences which proceed from premises about the source-domain of conceptual metaphors and employ their constitutive mappings are routinely made automatically, in language comprehension, at least whenever metaphorical expressions are freshly coined – and beyond. In other words: If this is correct, such CSG inferences amount to the kind of 'natural process' judgment heuristics put to use (see above). Second, there is a rationale for making such *metaphorical CSG inferences*, as I would like to call them: Linguistic realisation of a conceptual metaphor indicates higher-order analogy between domains, which we can profitably exploit in analogical reasoning. By and large, extensive linguistic realisation is indicative of wide and deep mapping: It makes visible a mapping of several related first-order relations plus constraining higher-order relations between them.

In science, including physics (Hesse 1966) and psychology (Gentner and Grudin 1985), analogical reasoning issuing in, and proceeding from, metaphors is also used consciously and deliberately. I would therefore like to conceptualise whatever unconscious use we may make of such reasoning outside language development, in problem-solving and judgment-making, as driven by the largely automatic application of a heuristic rule that can also be deliberately employed in conscious reasoning:

> **Metaphor Heuristic:** To obtain conclusions about a domain D, choose a linguistically realised conceptual metaphor that takes D as target domain, retrieve from memory knowledge about its source domain $D^S$, and infer conclusions about D from known facts about $D^S$, through copying with substitution and generation, employing mappings constitutive of the metaphor.

There already is empirical evidence for the use of this heuristic. For a start, experimental work on analogical reasoning showed that subjects are particularly likely to draw analogical inferences and accept their conclusions, where mappings are wide and deep and between semantically similar domains: where mappings map several first- plus constraining higher-order relations, from a source- to a target-domain to which we apply many of the same terms (Gentner and Markman 2005, Lassaline 1996). Conceptual metaphors are wide and deep mappings, and linguistic realisation ensures the same terms are applied to both. Second, the heuristic accurately predicts surprising fallacies in the reasoning of competent thinkers, including the conflation of ideas and their intentional objects (Fischer under review, cp. Fischer 2011). Third, experiments have documented unconscious inferences consistent with the heuristic in problem-solving (Thibodeau and Boroditsky 2011).

There is not only initial evidence for the actual use of the metaphor heuristic in automatic reasoning. The proposed heuristic also satisfies the requirements imposed by the ABC program, which requires that heuristics be

> (a) ecologically rational, i.e. [exploit] structures of information in the environment, (b) founded in evolved psychological capacities such as memory and the perceptual system, (c) fast, frugal, and simple enough to operate effectively when time, knowledge

> and computational might are limited, (d) precise enough to be modelled computationally, and (e) powerful enough to model both good and poor reasoning. (Goldstein and Gigerenzer 2002: 75)

(a) The metaphor heuristic exploits structural information embedded in our publicly shared language, namely information about structural or higher-order analogies between different domains, which is both acquired and deployed in coming to understand – some – metaphorical expressions.

(b) It is founded on exemplary 'evolved psychological capacities' (cp. Gigerenzer 2007: 58): language comprehension and memory.

(c) The heuristic is fast enough to operate in real time: It draws on processes demonstrably executed in little over a second, in everyday conversational settings (Gentner et al. 2002: 555).

(d) Several computational models of the three processes invoked by the metaphor heuristic are currently available (Gentner and Forbus 2011).

(e) While contributing to an explanation of poor reasoning (below), the heuristic can account also for good reasoning (like that studied in Thibodeau and Boroditsky 2011).

We can therefore conclude: The hypothesis that subjects employ the metaphor heuristic is sufficiently well supported to warrant further study. Second, the heuristic can be studied with ABC tools, to determine its applicability, ecological validity, and whether it is constitutive of an epistemic virtue. Third, the heuristic can also be studied with the aims of the heuristics and biases program, to predict fallacies and expose cognitive illusions. Most of this research remains to be done. In the remainder of this paper, I would like to get started by showing that the proposed metaphor heuristic helps us explain and assess influential factual intuitions about the mind. The story-line will be different, though, from that of familiar heuristics-and-biases narratives about heuristics like representativeness or availability: We will encounter a tale of how another natural process interferes with the application of a potentially helpful heuristic and how this interference generates cognitive illusions.

## 3. Memory-based Processing

Social psychologists seek to explain a wide variety of spontaneous inferences as the outcome of simple associative processes which can duplicate achievements of complex reasoning (Uleman et al. 2008). Cognitive psychologists are reconceptualising automatic applications of familiar judgment heuristics as driven by automatic associative processes in memory (Kahneman 2011, Sloman 1996). Associationist models of analogy and metaphor processing are available (e.g. Leech et al. 2008); some of them (e.g. Budiu and Anderson 2004) are predictively successful and compatible with the account above (section 2). If this wide range of work in social and cognitive psychology is on the right track, we should expect that also the proposed metaphor-heuristic is spontaneously applied through associative processing.

The slippery notion of associative processing can be rendered more precise by reference to semantic or to connectionist networks (Gigerenzer and Regier 1996). Let's use simple semantic networks (Anderson et al. 2004). They double as information storage and inference facilitators. The nodes of a *semantic network* stand for concepts and activation spreads to and from them automatically. Nodes which represent things or events that are spatiotemporally contiguous or share attributes or relations come to be linked. The network thus comes to link potential causes and their typical effects as well as things and their properties, and things and the categories to which they belong. When a node is activated, the

activation spreads out from it with decreasing strength to the several nodes directly or indirectly linked to it, subject to the principles that familiarity attracts and use strengthens: The more often a subject is exposed to a concept, the more strongly its node is activated each time; and the more frequently a link is activated, the more activation it gradually comes to pass on, while gradually atrophying upon disuse. An activated concept or proposition becomes conscious when the node representing it is activated above a certain threshold and more strongly than competitors. Association can therefore duplicate inferences, by spreading activation in sufficient strength from connected nodes representing one proposition to nodes that jointly represent another.

Associative processing within such a network is governed by the principle of partial matching (Kamas et al. 1996). To encounter it, answer the question: 'How many animals of each kind did Moses take on the ark?' In a classic experiment (Erickson and Mattson 1981), subjects were asked to read this and other questions out aloud, and either answer them or indicate that something is wrong with them, as appropriate. Over 80% answered "2" to the present question – even though they knew that the protagonist of the ark story is Noah, not Moses. In an attenuated form, the phenomenon persists when statements replace questions: In an analogous task, 40% of knowledgeable subjects assented to the corresponding statement about Moses. It seems subjects spontaneously interpret a question or statement about Moses as stating a familiar fact or prior belief about Noah. So why is that? Such semantic illusions are not explicable by reference to Gricean principles; but they may be due to partial matching in associative processing which ensures that relevant answers can be retrieved to differently phrased questions (Park and Reder 2004): Activation spreads from the node for the stimulus concept ('Moses') via nodes for attributes and relations possessed by its bearer (Biblical figure, leader, covenant with God, etc.) to nodes for bearers sharing these attributes (including Noah). When one of these similar bearers' node is activated also through other channels (say, upon mention of the ark), propositions about this similar bearer may be retrieved from memory.

A relevant process is outlined by the theory of *interpretation-based processing INP* (Budiu and Anderson 2004). Its interpretation process exploits the possibilities of partial matching and is incremental: It begins the moment the first semantic unit is read and parsed; results are updated after each unit is read. At each point, propositions with semantically similar concepts in matching positions are activated, where a concept is '*semantically similar*' to another, to the extent to which the things (individuals, stuffs, properties, etc.) they stand for are believed to share the same attributes or relations. At each point, the most strongly activated proposition is chosen as 'candidate interpretation' which specifies what the processed statement states. As the subject reads each semantic unit (verb, adverb, noun phrase), activation spreads to representations of several propositions with concepts semantically similar to the one represented by that unit: about other Biblical figures, leaders, etc. The most strongly activated candidate interpretation at a given point, say, 'Moses took his people out of Egypt', retains such rapidly decaying activation – only – while the next semantic unit is read, so that momentarily both the candidate and the new stimulus spread activation. This may activate more strongly the representation of another proposition, which thus replaces the previous candidate, and so on, until the entire sentence is read. In this way, a semantically similar proposition the subject already believes true (*Noah took two animals of each kind onto the ark*) may come to be accepted as content of the statement ('Moses took...').

This process is liable to interact in two ways with spontaneous applications of the metaphor heuristic: INP will turn some clearly wrong conclusions of spontaneous analogical inferences (conclusions 'about Moses', as it were) into apparent truisms ('about Noah'). And it will help forge fresh mappings which facilitate novel analogical inferences – and are not recommended by the metaphor heuristic. This interaction is best explained through an example. Various scholars have already connected the deliberate or unconscious use of visual metaphors with

the development of the modern (post-Aristotelian) concept of the mind (Rorty 1980, Lakoff and Johnson 1999). Let's consider a key component of this development: the transformation of the faculty of reason, intellect, or understanding, into an organ of sense which peers into a distinct perceptual space, called 'the mind' (a Rylean category mistake, if ever there was one).

The intuitions to be explained can be generated by several parallel and mutually reinforcing analogical (CSG) inferences which employ distinct but related visual metaphors and proceed from source-domain premises about different acts and achievements of visual perception. One of the most important of these conceptual metaphors is the metaphor *Thinking-about as looking-at* which motivates metaphorical talk of 'looking hard at the problem', 'looking at the issue from different sides', etc. A CSG inference which employs its basic mapping takes us from the source-domain premise that when we look at things we use our eyes to the conclusion that

$C_0$ When we think about something, we use our eyes.

| | **Source-domain premise** | **Operation** | **Target-domain conclusion** |
|---|---|---|---|
| 1 | S looks at X | Substitution | S thinks about X |
| 2 | Implies (1;3-4) | Substitution: identical | implies (1;3-4) |
| 3 | S uses Y | Substitution: identical | S uses Y |
| 4 | S's eyes (Y) | Generation | S's eyes (Y) |

INP is set to interpret fresh conclusions as expressing propositions the thinker already believes true, namely in case some such proposition is semantically sufficiently similar to the input. Both common sense and faculty psychology furnish us with a suitable belief, a belief we can express by sentences which employ the same terms as $C_0$, almost throughout: 'When we think about things we use our...' – 'reason', 'intellect', or 'understanding'. In faculty psychology as in ordinary language, these terms are used to refer not to any organ of sense but to a faculty, power, or ability, namely to the 'faculty of comprehending or reasoning' or the 'power or ability to understand' (as the *Oxford English Dictionary* puts it). But 'intellect' and 'understanding' can, like eyes, be said to be 'used' by subjects who 'have' them; like eyes, they may be more or less 'sharp', etc. They share enough attributes and relations to enjoy a degree of semantic similarity. The concepts employed earlier in the resulting statement are identical with those filling the same thematic roles in $C_0$, and both statements fill the same number of roles. Since the semantic similarity between two propositions is a function of (i) the similarities of the concepts filling the same thematic roles in the different propositions and (ii) the number of different such roles that get filled in either proposition (Budiu and Anderson 2004: 38), the proposition

$C_0^*$ When we think about something, we use our understanding

is semantically highly similar to $C_0$. Therefore INP will activate $C_0^*$ throughout, and eventually strongly, so that it is accepted as interpretation of $C_0$, and the utterance of $C_0$ is interpreted as expressing the prior belief $C_0^*$.

This interpretation process is bound to trigger new mappings: INP aligns candidate interpretations with the claim to be interpreted, and compares expressions in matching roles for semantic similarity. In processing the present CSG conclusion, it aligns:

S – uses – his eyes.

S – uses – his understanding.

This alignment facilitates mapping. In non-intentional analogical reasoning, we map not merely first-order relations onto other such relations (as the conceptual metaphor did) but readily map also *relata* of such relations. In such reasoning, subjects automatically first correlate the source- and target-domain elements to which the same concepts apply, and subsequently add mappings that correlate the hitherto unmapped *relata* of mapped relations. Hence the present alignment will have them automatically map 'x uses y' onto its target-domain homonym, and then map the 'rear' *relatum*, yielding the

**New mapping N:** eyes → understanding.

Such interplay of non-intentional analogical reasoning and INP can also yield this mapping's twin. Most ordinary uses of 'the mind' are motivated by a conceptual metaphor that builds on the mapping of spatial inclusion onto remembering and thinking-of (Fischer 2011: 41-45): To remember something is to 'retain' it in one's vicinity, to 'keep' or 'have' it 'in' a personal space, 'the mind', from which it may 'slip', etc. The present processes lead to the integration of this personal space into visual metaphors. CSG transforms truisms about the visual source domain into conclusions about the target domain. It takes us, e.g., from 'When we look at things, they are in our visual field', to the wild conclusion 'When we think about things, they are in our visual field'. For the reasons explained, INP aligns this with the semantically similar proposition 'When we think of things, they are in our mind', and facilitates the fresh

**Mapping M:** visual field → mind.

Mappings M and N are not constitutive of the visual metaphors we have considered: They cannot be obtained from their basic mappings through elementary CS inferences. These fresh mappings – which are not recommended by the metaphor heuristic – facilitate a plethora of CSG inferences that jointly transform 'mind' and 'understanding' by taking us from truisms like 'When we look at things, things are before our eyes' to conclusions like (non-identical substitutions underlined):

C1 When we <u>think about</u> things, things are before our <u>understanding</u>.

C2 When we <u>think about</u> things, things are in our <u>mind</u>.

C3 Things before our <u>understanding</u> are in our <u>mind</u>.

In non-philosophical discourse, the verb 'perceive' ordinarily applies in the same generic sense, 'to apprehend with the mind or senses' (as the *OED* puts it), to epistemic achievements brought off by using either one's wit or one's senses, no matter which. 'S perceives X' thus stands for a generic epistemic relation that is an element of both the present source- and target domain. It gets mapped onto itself in analogical reasoning, and is 'substituted' by itself in further CSG inferences like the inference from:

When we <u>look at</u> things, we perceive things with our <u>eyes</u>, in our <u>visual field</u>. To:

C4 When we <u>think about</u> things, we perceive things with our <u>understanding</u>, in our <u>mind</u>.

Together with C1 to C3, this transforms the understanding from an intellectual faculty into an organ of sense employed in thought, and the mind into this organ's perceptual field.

## 4. From Explanation to Assessment

This explanation facilitates epistemological assessment of the philosophically relevant intuitions explained. The present explanation allows us to identify a crucial mistake in non-intentional analogical reasoning leading to C1 to C4, and related intuitions, second, to show that the intuitions due to this mistake are constitutive of cognitive illusions, and possibly,

third, to resolve seminal versions of a familiar mind-body problem. We will now gradually build up to the mistake at issue, by pointing out a specific risk in analogical reasoning, how we commonly mitigate this risk, and how INP-generated mappings prevent our default mitigation strategy.

In analogical reasoning, the first mappings to be made connect source- and target-domain elements to which the same terms apply (section 2). In the case of first-order analogies (e.g. the supposed analogy between atoms and the solar system), relevant relational terms (like 'x orbits y') stand for the same relations in both domains, so that these relations get mapped onto themselves. This is different where linguistically realised conceptual metaphors are built on second-order analogies between concrete source- and more abstract target-domains: Here, the same terms typically stand for radically different first-order relations in the two domains. E.g. 'x looks at y' stands for *looking-at* in one domain and *thinking-about* in the other. *Relata* of one relation (say, John who looks at Joan) will typically stand in a host of further relations (like: x stands in front of y) which are not, or cannot be, shared by the *relata* of the other relation (John may think of absentees, and of problems and risks without physical location). Hence CSG inferences that involve generation as well as substitutions licensed by a conceptual metaphor are very risky, where we move from a concrete to a more abstract domain.

But this need not lead to false conclusions: In ordinary discourse, we mitigate the present risk by placing a metaphorical interpretation on generated terms, wherever possible, as a default: We interpret them in the light of their metaphorical implications (cp. section 2):

> p has the **metaphorical implication** q* iff p→q* can be obtained through CSG from a truth p→q about the source-domain of a conceptual metaphor whose constitutive mappings license substitution of or in q yielding q*, but license no substitution of or in p. ('→' designates de- or inductive inference.)

For instance, things before your eyes are, by and large, easy for you to see, and simple CSG with the conceptual metaphors considered takes us from this premise to 'If things are before your eyes, it is easy for you to get to know them'. Similarly, when something is outside my range of vision, I cannot see them, so simple CSG yields: 'If something is beyond my ken, it is impossible for me to know it'. This has us say, in ordinary speech, that things we cannot get to know or understand are 'beyond our ken'.

When followed by such metaphorical interpretation, CSG inferences without M or N take us from the premises of C1, C2, and C4 to the conclusions: 'When we think about things, we can easily get to know things', '...it is possible for us to get to know things', and '... we (get to) know things', respectively. Instead of C3, we get 'When it is easy for us to get to know things, it is possible for us to get to know them.' These conclusions do not even appear to refer to spaces or organs of perception, of any kind or description.

Use of INP-generated mappings like M and N, in CSG inference, prevents such metaphorical interpretation. CSG with mapping N transforms '...before my eyes' into '... before my understanding'. This has no implications in the source domain of sight – and hence no metaphorical implications. CSG inferences without the new mappings M and N have us apply to the intellectual target-domain predicates like 'x is before my eyes' or 'x is in our visual field', which include spatial terms but which, in their entirety, have source-domain implications that facilitate metaphorical interpretation. By contrast, CSG inferences with M and N replace 'your eyes' and 'my visual field', respectively, and carry over from the source domain only those spatial relations (x is before y, x is in y, etc.). They thus lead to conclusions that place elements of the intellectual target-domain into spatial relations, and are not amenable to metaphorical interpretation – but only to an insufficient ersatz treatment.

Early modern thinkers frequently explain that minds are meant to 'take up no space, have no extension', as Locke puts the common position (Locke 1700, II.ix.10). Thus, Berkeley

explains: 'When I speak of objects as existing in the mind ... I would not be understood in the gross literal sense, as when bodies are said to exist in a place' (Berkeley 1734, 250). These thinkers continue to accept mappings M and N and continue to posit mind-spaces and understanding-organs in us. They then seek a non-literal interpretation merely for the spatial terms ('x is in y' and 'x is before y'). Proper metaphorical interpretation removes from CSG conclusions employing vision-to-thought analogies all reference to perceptual spaces and organs of any kind or description. By contrast, early modern *ersatz* metaphorical interpretation leaves us stranded with notions of non-physical 'locations' and 'spaces' without extension – which gives rise to seminal versions of the ontological mind-body problem.

This paper has addressed a possible source of this problem: CSG inferences with mappings M and N carry over spatial relations and posit entities which stand in the same spatial etc. relations to subjects and the objects they think about as visual fields and eyes, stand to visual observers and the things they see:

- a 'mind' in which those objects are located [C2]
- before an 'understanding' [C1, C3]
- with which they are perceived [C4].

*Ersatz* metaphorical interpretation manifests knowledge that there are no such entities. When accepted by thinkers who offer such interpretative glosses, these intuitive conclusions hence violate what we may call the '*no assumed false lemma rule*' (not to be confused with the 'no false lemma rule' proposed in response to the Gettier problem): Do not rely in your judgments, intuitive or deliberate, on assumptions you believe to be false. The intuitions explained rely on existence assumptions thinkers know to be false.

If correct, our explanation reveals that these intuitions possess all four defining characteristics of cognitive illusions (section 1): They violate the uncontroversial 'no assumed false lemma rule' which constrains what a thinker has warrant to conclude or believe, that is, a normative rule. Second, the violations exposed are predictable, not random: Both INP-generated mappings and the transformation of CSG conclusions through INP can be predicted with: the proposed metaphor heuristic, information about which conceptual metaphors are linguistically realised in the thinker's language, and information about the thinker's prior beliefs and their subjective semantic similarity to conclusions inferable with the heuristic. Third, both interacting processes, namely, application of the metaphor heuristic and INP, are largely automatic in character – even if their joint outputs may be subject to effortful modification (including the explanations quoted, which evidence better knowledge). Fourth, even once they have realised that these claims cannot be right, thinkers find these intuitions intuitively compelling enough to presuppose them in further reasoning (Fischer 2011).

The exposure of such cognitive illusions can perhaps help us resolve a number of philosophical problems. Ontological problems arise where we spontaneously project spatial properties onto non-physical things we know not to have them, and seminal versions of the ontological mind-body problem might arise in this way. Where we do not re-interpret paradoxical intuitions or conclusions q but accept them despite their apparent conflict with extant convictions p, or among each other, we raise reconciliation problems, typically, though not invariably, articulated by questions of the form: How is it possible that p (given that q, and that $q \rightarrow \neg p$)? We can, I submit, resolve some problems of these kinds by showing that they are raised only by cognitive illusions.

**Eugen Fischer**
University of East Anglia
E.Fischer@uea.ac.uk

## References

Alexander, J. and J. Weinberg 2007: 'Analytic Epistemology and Experimental Philosophy', *Philosophy Compass 2*, 56-80.

Anderson, J. R., D. Bothell, M.D. Byrne, S. Douglass, C. Lebiere, and Y. Qin 2004: 'An Integrated Theory of the Mind', *Psychological Review 111*, 1036-1060.

Bargh, J.A. 1994: 'The Four Horsemen of Automaticity', in R. Wyer and T. Srull (eds.): *Handbook of Social Cognition, vol.1*, Hillsdale; Earlbaum, 1-40.

Berkeley, G. 1734: 'Three Dialogues', in *Philosophical Works*, ed. M. Ayers, London: Dent, 1996.

Budiu, R. and J.R. Anderson 2004: 'Interpretation-based Processing: A Unified Theory of Semantic Sentence-comprehension', *Cognitive Science 28*, 1-44.

Cappelen, H. 2012: *Philosophy without Intuitions*. Oxford: OUP.

Day, S.B. and D. Gentner 2007: 'Non-intentional Analogical Inference in Text-comprehension', *Memory and Cognition 35*, 39-49.

De Neys, W. 2006: 'Automatic-heuristic and Executive-analytic Processing during Reasoning: Chronometric and Dual-task Considerations', *Quarterly Journal of Experimental Psychology 59*, 1070-1100.

Erickson, T. and M. Mattson 1981: 'From Words to Meaning: A Semantic Illusion', *Journal of Verbal Learning and Verbal Behaviour 20*, 540-551.

Evans, J.S.B.T. 2008: 'Dual-processing Accounts of Reasoning, Judgment and Social Cognition', *Annual Review of Psychology 59*, 255-78.

Falkenhainer, B., K.D. Forbus, and D. Gentner 1989: 'The Structure-mapping Engine: Algorithm and Examples', *Artificial Intelligence 41*, 1-63.

Fischer, E. 2011: *Philosophical Delusion and its Therapy*. New York: Routledge.

Fischer, E. under review: 'Philosophical Intuitions, Heuristics, and Metaphors'

Forbus, K.D., D. Gentner, and K. Law 1995: 'MAC/FAC: A Model of Similarity-based Retrieval', *Cognitive Science 19*, 141-205.

Gentner, D. and B.F. Bowdle 2008: 'Metaphor as Structure-mapping', in R. Gibbs (ed.): *The Cambridge Handbook of Metaphor and Thought*, New York: CUP, 109-128.

Genter, D. and K.D. Forbus 2011: 'Computational Models of Analogy', *WIREs Cognitive Science 2*, 266-276.

Genter, D. and J. Grudin, 1985: 'The Evolution of Mental Metaphors in Psychology', *American Psychologist 40*, 181-192.

Gentner, D., M. Imai, and L. Boroditsky 2002: 'As Time Goes By: Evidence for Two Systems in Processing Space-time Metaphors', *Language and Cognitive Processes 17*, 537-565.

Gentner, D. and A. Markman 2005: 'Defining Structural Similarity', *Journal of Cognitive Science 6*, 1-20.

Gentner, D., M. Ratterman, and K. Forbus 1993: 'The Roles of Similarity in Transfer: Separating Retrievability from Inferential Soundness', *Cognitive Psychology 25*, 527-575.

Gerken, M. 2011: 'Epistemic Focal Bias', *Australasian Journal of Philosophy*, DOI: 10.1080/00048402.2011.631020.

Gigerenzer, G. 2007: *Gut Feelings*. London: Allen Lane.

Gigerenzer, G. 2008: *Rationality for Mortals*. Oxford: OUP.

Gigerenzer, G. and T. Regier 1996: 'How Do We Tell an Association from a Rule?', *Psychological Bulletin 119*, 23-26.

Gigerenzer, G. and T. Sturm 2012: 'How (Far) Can Rationality Be Rationalised?', *Synthese 187*, 243-268.

Gigerenzer, G., P.M. Todd, and the ABC Research Group 1999: *Simple Heuristics that Make Us Smart*. Oxford: OUP.

Goldstein, D. and G. Gigerenzer 2002: 'Model of Ecological Rationality: The Recognition Heuristic', *Psychological Review 109*, 75-90.

Hawthorne, J. 2004: *Knowledge and Lotteries*. Oxford: OUP.

Hesse, M. 1966: *Models and Analogies in Science*. Notre Dame: University of Notre Dame Press

Holyoak, K.J. 2012: 'Analogy and Relational Reasoning', in K.J. Holyoak and R.G. Morrison (eds.): *Oxford Handbook of Thinking and Reasoning*, Oxford: OUP, 234-259.

Holyoak, K.J., L.R. Novick, and E.R. Melz 1994: 'Component Processes in Analogical Transfer', in K.J. Holyoak and K.J. Barnder (eds.): *Advances in Connectionist and Neural Computation Theory, vol. 2*, Norwood: Ablex, 113-180.

Jäkel, O. 1995: 'The Metaphorical Concept of Mind', in J.R. Taylor and R.E. McLaury (eds.): *Language and the Cognitive Construal of the World*, Berlin: de Gruyter, 197-229.

Kahneman, D. 2011: *Thinking, Fast and Slow*. London: Allen Lane.

Kahneman, D. and S. Frederick 2005: 'A Model of Heuristic Judgment', in K.J. Holyoak and R. Morrison (eds.): *Cambridge Handbook of Thinking and Reasoning*, Cambridge: CUP, 267-293.

Kamas, E.N. and L.W. Reder 1995: 'The Role of Familiarity in Cognitive Processing', in R.F. Lorch and E.J. O'Brien (eds.): *Sources of Coherence in Reading*, Hillsdale: Earlbaum, 177-202.

Lakoff, G. 1993: 'The Contemporary Theory of Metaphor', in A. Orthony (ed.): *Metaphor and Thought*, 2nd ed., Cambridge: CUP, 202-251.

Lakoff, G. and M. Johnson 1999: *Philosophy in the Flesh*. New York: Basic Books.

Lassaline, M.E. 1996: 'Structural Alignment in Induction and Similarity', *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*, 754-770.

Leech, R., D. Mareschal, and R.P. Cooper 2008: 'Analogy as Relational Priming', *Behavioural and Brain Sciences 31*, 357-414.

Locke, J. 1700: *An Essay Concerning Human Understanding*, 4th ed., ed. by P.H. Nidditch. Oxford: Clarendon Press, 1975.

Mercier, H. and D. Sperber 2009: 'Intuitive and Reflective Inferences', in J. Evans and K. Frankish (eds.): *In Two Minds: Dual Processes and Beyond*, Oxford: OUP, 149-170.

Moors, A. and J. De Houwer 2006: 'Automaticity: A Theoretical and Conceptual Analysis', *Psychological Bulletin 132*, 297-326.

Nagel, J. 2010: 'Knowledge Ascriptions and the Psychological Consequences of Thinking about Error', *Philosophical Quarterly 60*, 286-306

Nagel, J. 2011: 'The Psychological Basis of the Harman-Vogel Paradox', *Philosophers' Imprint 11(5)*, 1-28

Pachur, T. and R. Hertwig 2006: 'On the Psychology of the Recognition Heuristic', *Journal of Experimental Psychology: Learning, Memory, and Cognition 32*, 983-1002

Park, H. and L.M. Reder 2004: 'Moses Illusion', in Pohl (ed.) 2004, 275-291

Pohl, R.F. (ed.) 2004: *Cognitive Illusions*. New York: Psychology Press

Pohl, R.F. 2006: 'Empirical Tests of the Recognition Heuristic', *Journal of Behavioural Decision Making 19*, 251-271.

Rorty, R. 1980: *Philosophy and the Mirror of Nature*. Oxford: Blackwell.

Sloman, S.A. 1996: 'The Empirical Case for Two Systems of Reasoning', *Psychological Bulletin 119*, 3-22.

Sosa, E. 2007: 'Intuitions: Their Nature and Epistemic Efficacy', *Grazer Philosophische Studien 74*, 51-67.

Spicer, F. 2007: 'Knowledge and the Heuristics of Folk Psychology', in V. Hendricks and D. Pritchard (eds.): *New Waves in Epistemology*, London: Palgrave Macmillan.

Thibodeau, P.H. and L. Boroditsky 2011: 'Metaphors We Think With: The Role of Metaphor in Reasoning', *PLoS ONE 6*, e16782, doi: 10.1371/journal.pone.0016782

Tversky, A. and D. Kahneman 1974: 'Judgment under Uncertainty: Heuristics and Biases', *Science 185*, 1124-1131.

Tversky, A. and D. Kahneman 1983: 'Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment', *Psychological Review 90*, 293-315.

Uleman, J.S., S.A. Sairbay, and C.M. Gonzales 2008: 'Spontaneous Inferences, Implicit Impressions, and Implicit Theories', *Annual Review of Psychology 59*, 329-60.

Volz, K.G., R.I. Schubotz, M. Raab, L.J. Schooler, G. Gigerenzer, and D.Y. Cramon, 2006: 'Why You Think Milan Is Larger than Modena', *Journal of Cognitive Neuroscience 18*, 1924-36.

Wharton, C.M., K.J. Holyoak, and T.E. Lange 1996: 'Remote Analogical Reminding', *Memory and Cognition 24*, 629-643.

Williamson, T. 2005: 'Contextualism, Subject-sensitive Invariantism, and Knowledge of Knowledge', *Philosophical Quarterly 55*, 213-35.

Williamson, T. 2007: *The Philosophy of Philosophy*. Oxford: Blackwell

.

# What are Epistemic Duties?

Andrea Kruse

Epistemic duties are important for answering certain questions within normative epistemology. Besides their role as criteria of epistemic evaluation, they are also supposed to guide the intellectual conduct of epistemic agents. Recently, epistemic duties have become very unpopular in epistemology. One reason is a famous argument against epistemic duties given by Alston. In this paper, I will develop an alternative approach to epistemic duties, which does not fall prey to Alston's argument. To illustrate this, I will sketch Alston's argument and discuss two responses to it and their shortcomings. In addition, I will give an alternative response to Alston's argument. This response presupposes an alternative approach to epistemic duties which I will describe in this paper. Moreover, I will characterize epistemic duties in more detail by giving and discussing necessary conditions for them.

## 1. Epistemic Duties in Epistemology

The notion of an epistemic duty is an important notion within the area of normative epistemology. Many epistemologists take the norm that one should believe in accordance with one's evidence as a paradigmatic instance of an epistemic duty. Epistemic duties or epistemic norms are important for the "ethics of belief", if it is taken as a theory of normative epistemology[1]. The central question(s) of the "ethics of belief" are "What ought we to believe?" and "How ought we to believe?". Answering these questions from an epistemic perspective calls for searching or establishing epistemic duties or epistemic norms. Recently, the most popular epistemic theory of the "ethics of belief" is evidentialism. Evidentialism answers the question of what one should believe with the requirement that one should always believe in accordance with one's evidence (cf. Feldman 2002).

One of the most important aims of epistemology is to give an intuitively appealing account of knowledge, which can deal in a satisfactory way with epistemic problems like the Gettier problem, skepticism and the value problem. Goldman claims that besides the aim of finding an appropriate account of knowledge, there are also other aims in epistemology. According to Goldman "[...] one central aim of epistemology is to guide or direct our intellectual conduct" (Goldman 2001: 116). For the realm of epistemology, which is concerned with this aim, the search for epistemic duties or epistemic norms appears to be necessary. The considerations above show that notions like epistemic duty or epistemic norm are of epistemic importance.

In what follows I will distinguish epistemic duties from epistemic norms in the following way. Epistemic duties as well as epistemic norms are ought-sentences (Oφ), which serve as criteria of epistemic evaluation. However, an epistemic norm is only an epistemic duty, if, besides serving as a criterion of epistemic evaluation, it also serves to guide epistemic agents in their intellectual conduct. Thus, epistemic duties have an evaluative function as well as a guidance function. For an epistemic norm to satisfy the guidance function, it has to satisfy some further

[1] The "ethics of belief" goes back to John Locke's "*An essay concerning human understanding*" and has become very popular with Clifford's famous article "*The ethics of belief*" (cf. Chignell 2013). In that article the question, "What one should believe?" is answered from the perspective of moral considerations, which locates the "ethics of belief" in the ethical realm rather than in epistemology. However, if one considers the question from an epistemic perspective, "the ethics of belief" becomes a discipline of normative epistemology.

conditions, which I will explain in the last part of this paper. An epistemic norm, which satisfies the guidance function, will be called a *regulative epistemic norm*. From Goldman's quote it becomes clear that the search for and the establishment of epistemic duties are what is needed to deal with the epistemological aim of guiding our intellectual conduct (cf. Goldman 2001: 116). I will characterize epistemic duties in more detail below. From what has been said so far, we get two assumptions about epistemic duties. The first assumption is that epistemic duties are epistemic norms, but not vice versa. The second assumption is that epistemic duties are regulative epistemic norms, i.e. norms with a guidance function. This is in accordance with what Feldman says about epistemic duties.

> Epistemological duties are duties that one must carry out in order to be successful from an intellectual (or epistemological) perspective. (Feldman 2002: 376)

### *1.2 An Argument against Epistemic Duties*

There is an argument against epistemic duties which can be found in Alston (1988)[2] and also in Feldman (2001). The argument against epistemic duties can be paraphrased in the following way:

(1) There are epistemic duties only if epistemic duties have a non-empty domain, i.e. they apply at least to some epistemic agents in some cases.

(2) Epistemic duties take doxastic attitudes as their objects. If Oφ is an epistemic duty, then φ is a doxastic attitude.

(3) If an epistemic agent has an epistemic duty Oφ (i.e. Oφ is incumbent on the epistemic agent), then the epistemic agent is able to bring about φ voluntarily. (*Epistemic "ought implies can" principle*)

(4) Epistemic agents are not able to voluntarily bring about doxastic attitudes.

(5) Therefore, epistemic duties do not apply to any epistemic agent and hence, epistemic duties have an empty domain.

(6) Hence, there are no epistemic duties.

The first premise seems to be plausible, because epistemic duties, which do not apply to any epistemic agent, do not make sense from the start. The second premise rests on the idea that only duties which have doxastic attitudes as their objects can be characterized as epistemic duties. Feldman shows that to have a doxastic attitude as an object is only a necessary, but not a sufficient condition for Oφ to be an epistemic duty (cf. Feldman 2002: 373). For now, I will accept this premise. The third premise captures the idea that epistemic norms, which are supposed to guide us in our intellectual conduct, need to take the cognitive limits of epistemic agents into account. How else could we say that epistemic agents can be guided by epistemic norms? That an epistemic duty to φ is only incumbent on the epistemic agent if the agent is able to voluntarily bring about φ[3], is one way to take the cognitive limits of epistemic agents into account. Therefore, an epistemic "ought implies can" principle appears to be a necessary condition for Oφ to be an epistemic duty. Moreover, the epistemic "ought implies can"

---

[2] Alston (1988) does not argue directly against the existence of epistemic duties. This argument is a byproduct of his arguments against the viability of a deontological notion of epistemic justification, which relies on his famous argument for the psychological incapacity of epistemic agents to bring about doxastic attitudes voluntarily, i.e. his rejection of doxastic voluntarism.

[3] Please note that to prevent a common criticism against "ought implies can", namely that if "ought implies can" holds, then agents can too easily release themselves from having a certain duty in a certain situation by bringing about that they are not (any longer) able to bring about the required state of affairs (cf. Sinnott-Armstrong 1984), I support a more refined version of "ought implies can" as it is given by Howard-Snyder (2006: 235f.)

principle is also necessary for ascribing epistemic blame and praise, which is also important for guiding the intellectual conduct of epistemic agents[4].

Alston himself gives one of the most famous arguments for the fourth premise. He argues that epistemic agents are not able to form doxastic attitudes voluntarily because of the epistemic-reason-responsive nature of the cognitive processes, which are used to form doxastic attitudes (Alston 1988: 263-268). Alston identifies having voluntary control over one's doxastic attitudes with having an effective choice about what doxastic attitude one takes toward a proposition p in a certain situation (Alston 1988: 261). According to Alston voluntarily bringing about a state of affairs φ implies that one has the control to bring about φ as well as some incompatible alternative to φ[5] (cf. Alston 1988: 261). He then asks us to consider whether we have an effective choice, given we take our evidence to speak (conclusively) in favor of p? He answers that if we take our evidence to speak (conclusively) in favor of p, we do not have the power to bring about another doxastic attitude toward p than the belief that p. The same holds for the formation of the disbelief that p if we take our evidence (conclusively) to speak against the truth of p. If we take our evidence to speak neither against nor for the truth of p, Alston argues that we do not have the power to form a different doxastic attitude toward p than the suspension of judgment about p (cf. Alston 1988: 263-268). From these considerations Alston concludes that we do not have an effective choice over our doxastic attitudes and, therefore, we are not psychologically able to form doxastic attitudes voluntarily (cf. Alston 1988: 263).

Premise three and four together with modus tollens give us five. Five and premise one together with modus tollens give us the conclusion that there are no epistemic duties. Since epistemic duties unlike epistemic norms in general serve to guide the intellectual conduct of epistemic agents, the third premise does not hold for epistemic norms in general. Therefore, the argument is not an argument against the existence of epistemic norms. However, the argument raises doubts on the viability of the "ethics of belief", taken as an epistemic theory, and the viability of the epistemic realm which is concerned with the epistemological aim to guide our intellectual conduct. As far as the viability of those research disciplines is dependent on a viable notion of epistemic duty, Alston's argument threatens the viability of those disciplines as well.

### *1.3 The Norm of Evidentialism*

Especially, proponents of evidentialism like Feldman (2001, 2002) and Steup (1998), who take evidentialism to be the only viable epistemic theory of the "ethics of belief", have objected to Alston's argument. In what follows, I will shortly describe how each of them tries to reject Alston's argument. After that I will explain why they fail to show that the norm of evidentialism is an epistemic duty.

Feldman characterizes the evidentialist norm in the following way:

> For any person S and time t, if S considers p at t, then S has the duty to have the attitude toward p that fits the evidence S has at t concerning p. (Feldman 2002: 368)

Beside Feldman and Steup, many epistemologists consider the evidentialist norm as the paradigmatic epistemic duty. According to this norm, there is only one doxastic attitude which the epistemic agent ought to have at t toward a certain p, she is considering at t. The epistemic agent ought to have the doxastic attitude toward p which fits her available evidence at t (cf. Feldman 2000: 680). Feldman agrees with Alston that we do not have direct

[4] I will explain this in more detail in the third section of this paper.

[5] Whether Alston's notion of voluntary control is neutral regarding compatibilist and incompatibilist approaches to the free will need not concern us here, because there are compatibilist as well as incompatibilitist interpretations of the condition of having an effective choice.

voluntary control over our doxastic attitudes and claims that we have indirect voluntary control[6] only in very rare cases (cf. Feldman 2001: 80f.)[7]. Feldman rejects the argument of Alston by rejecting the third premise. According to Feldman, to have an epistemic duty Oφ does not imply that the agent can voluntarily bring about φ. He admits that if one rejects that an epistemic "ought implies can" principle holds for epistemic duties (premise three), it follows that such epistemic duties do not serve to ascribe epistemic blame or praise (cf. Feldman 2001: 89). The ascription of epistemic praise or blame through other epistemic agents can be seen as one way in which epistemic duties can indirectly guide our intellectual conduct. Moreover, the holding of "ought implies can" for epistemic duties ensures that the duties are sensitive to the cognitive limitations of epistemic agents. Epistemic norms, for which an epistemic "ought implies can" principle does not hold, cannot guide the intellectual conduct of epistemic agents, since they can neither be used to ascribe epistemic blame or praise nor are they sensitive regarding the agents cognitive limitations. By rejecting the assumption that an epistemic "ought implies can" principle holds for the evidentialist norm, this norm becomes a mere evaluative epistemic norm. So Feldman's way to reject Alston's argument has the consequence that the evidentialist norm cannot be characterized as an epistemic duty, but that it is rather a mere criterion of epistemic evaluation.

Steup (1998) is also a proponent of an evidentialist account of epistemic justification. He rejects the argument of Alston not by rejecting the epistemic "ought implies can" principle for epistemic duties, but by rejecting the fourth premise, i.e. that epistemic agents are psychologically not able to bring about doxastic attitudes voluntarily (cf. Steup 2008). Steup proposes a compatibilist account of voluntary doxastic control[8] in the following way.

> S's attitude A toward p is free [i.e. S voluntarily brings about her attitude A toward p] iff (i) S has attitude A toward p, and (ii) S wants to have attitude A toward p; (iii) S's having taken attitude A toward p is the causal outcome of a reason-responsive mental process. (Steup 2008: 380 [parenthesis, A.K.])

According to Steup, the epistemic-reason-responsive nature of belief-forming processes enables that we have voluntary doxastic control over our doxastic attitudes. Moreover, he claims that doxastic attitudes are outcomes of deliberative processes. Within the process of deliberation, which accompanies our consideration of p, we interpret, evaluate, and weight our evidence for and against p. The outcome of this deliberation determines what doxastic attitudes we take toward p (Steup 2008). For example, if I consider p at t and if, after deliberating about the evidence I have for p at t, I come to the conclusion that my evidence supports p, I will form the belief that p accordingly. As mentioned before, Alston (1988) assumes that for an epistemic agent to form a doxastic attitude toward p voluntarily, it is necessary that she could have brought about an alternative doxastic attitude toward p in the "same" situation[9]. Let's call this *the principle of doxastic alternatives*[10]. A compatibilist

---

[6] An epistemic agent exercises indirect voluntary control over her doxastic attitude D toward p if and only if her doxastic attitude D toward p is the immediate (and intended) result of her bringing about a certain state of affairs voluntarily (cf. Alston 1988, Feldman 2001).

[7] The standard example for an agent exercising indirect voluntary doxastic control consists in an epistemic agent who is in her office, in which the light is off. Under normal conditions the agent is able to bring about the belief that the light in her office is on through exercising indirect voluntary doxastic control. The agent is able to bring about the belief that the light in her office is on by merely switching the light-switch. Since, given the switch is functioning well and the agent is cognitively well functioning and nothing covers her eyes, by switching the light-switch she will probably get the evidence needed for forming the belief that the light in her office is on (cf. Feldman 2001: 82).

[8] Note that according to the terminology of Steup (2008) an agent has voluntary control over her doxastic attitude if and only if the doxastic attitude is free.

[9] How we characterize the *sameness of situations* is crucial to the question of whether the principle is understood in a compatibilist or an incompatibilist way.

version of this principle would be as follows: an epistemic agent S forms a doxastic attitude toward p voluntarily *only if* it holds that if the deliberative processes would have come to a different outcome, for example by weighing the evidence differently, the epistemic agent would have formed a different doxastic attitude toward p. The consequent of this implication is true within Steup's approach due to the epistemic-reason-responsive nature of belief-forming processes[11]. However, Steup's approach to voluntary doxastic control does not ensure that epistemic agents can bring about the state of affairs voluntarily, which is to form the doxastic attitude toward a considered proposition p at t that fits the evidence the agent has for p at t. For epistemic agents to bring about the state of affairs voluntarily, which is required by the evidentialist norm, Steup has to make another assumption, which is typical for accessibilist internalism about epistemic justification. For every doxastic attitude and every proposition p it holds that if S has a doxastic attitude D toward p at t, then S has cognitive access to her available evidence for p at t and the epistemic principles, which are relevant to evaluate the evidence for p at t, by mere reflection about the epistemic status of her doxastic attitude toward p at t (cf. Steup 1998). From this, it follows that an epistemic agent can recognize the epistemic status of her doxastic attitude at a time t by mere reflection (cf. ibid.), call this the *transparency condition*. Since whether we reflect upon our evidence for a certain proposition at t is under our voluntary control, the transparency condition together with the reason-responsive nature of belief-forming processes ensure that epistemic agents can voluntarily bring about the doxastic attitude toward p at t, which is required by the norm of evidentialism at t. Moreover, within Steup's evidentialist approach to epistemic justification it is possible to ascribe epistemic blame to agents who violate the norm of evidentialism (without an excuse), since an epistemic "ought implies can" principle holds for the evidentialist norm. According to Steup (1998) an epistemic agent is blameworthy for holding a doxastic attitude toward a proposition p at t, which violates the norm of evidentialism, since, if she would have reflected about the epistemic status of her doxastic attitude at t, she would have recognized that her doxastic attitude is not supported by her evidence at t due to the transparency condition. By recognizing that her doxastic attitude is not supported by the available evidence at t, the epistemic agent would have revised her doxastic attitude toward p accordingly due to the reasons-responsive nature of belief-forming processes. Given Steup's assumptions are correct, he can reject Alston's argument and keep the assumption that an epistemic "ought implies can" principle holds for the evidentialist norm, which is necessary for this norm to be an epistemic duty.

The problem with Steup's account is that the transparency condition is too demanding. We are not always able to recognize the epistemic status of our doxastic attitudes by mere reflection, since it is not the case that the evidence, which we have in a certain situation, and the epistemic principles, which are relevant for assessing the available evidence, are always recognizable for us by mere reflection. I suppose that epistemic agents are able to recognize the epistemic status of their doxastic attitudes only in rare cases. Even the most popular internalist approaches to epistemic justification like the ones proposed by Feldman and Conee (2001) and Wedgwood (2002) do not assume something like the transparency condition. Given this is true, the evidentialist norm would only apply in very rare cases and would not be suitable to ground the notion of epistemic justification. The only available option for Steup to keep the idea that the evidentialist norm grounds an evidentialist notion of epistemic justification is then to reject the transparency condition. However, within Steup's approach the transparency condition is necessary for epistemic agents to have voluntary

[10] I call this the principle the principle of doxastic alternatives in analogy to its close cousin the principle of alternate possibilities.

[11] I doubt that this is a viable approach to voluntary doxastic control, but I cannot discuss this approach in more detail here. For a critical discussion of Steup's approach to voluntary doxastic control, see Buckareff (2006).

control over the state of affairs required by the evidentialist norm. Hence, if Steup rejects the transparency condition, he has to reject the assumption that an epistemic "ought implies can" principle holds for the evidentialist norm as well, since otherwise he would fall prey to Alston's argument against epistemic duties. But, if Steup rejects the epistemic "ought implies can" principle for the norm of evidentialism, then this norm cannot guide the intellectual conduct of epistemic agents and is not any longer an epistemic duty, but is a mere evaluative epistemic norm. Thus, Steup faces a dilemma. If the norm of evidentialism is supposed to be an epistemic duty, then it applies to epistemic agents only in very rare cases, which would disqualify the norm to ground a viable account of epistemic justification. However, if Steup rejects the assumption that an epistemic "ought implies can" principle holds for the evidentialist norm, then the norm of evidentialism is not an epistemic duty.

From the discussion of the evidentialist norm above, we can conclude that the attempts of Feldman and Steup to reject Alston's argument and to take the evidentialist norm to be an epistemic duty have failed. This might give us reasons to think that the evidentialist norm is not an epistemic duty, but only an evaluative epistemic norm.

## 2. Are There Epistemic Duties?

According to what has been said so far, one might wonder whether there are epistemic duties at all. The conclusion of the last section was that the evidentialist norm, which was taken to be the paradigmatic epistemic duty, is not an epistemic duty. In what follows, I will argue that there are epistemic duties. I will reject Alston's argument by rejecting premise two of Alston's argument. An alternative approach to epistemic duties will be introduced, within which the second premise of Alston's argument does not hold. It will be argued that it is not the case that $O\varphi$ is only an epistemic duty, if $\varphi$ is a doxastic attitude. This allows me to claim that an epistemic "ought implies can" principle holds for epistemic duties, which is necessary for epistemic duties to guide our intellectual conduct, without the need to assume doxastic voluntarism, i.e. that epistemic agents are able to bring about doxastic attitudes voluntarily. Moreover, within the alternative approach to epistemic duties it is assumed that epistemic duties are standards of epistemic evaluation and that they are principles, which guide our intellectual conduct.

But, what are epistemic duties if their object is not necessarily a doxastic attitude? Since I assume that an epistemic "ought implies can" principle holds for epistemic duties, I will argue that the states of affairs which are objects of epistemic duties are epistemically significant states of affairs, which can be brought about voluntarily by normal epistemic agents. A state of affairs is epistemically significant if and only if it has a positive or a negative influence on the pursuit of epistemic goals, like the truth-goal or the goal of understanding.

Although epistemic agents are not able to bring about doxastic attitudes directly voluntarily, they are nevertheless able to influence the quality of their doxastic attitudes as well as the quality of the belief-formation and -revision in various ways (cf. Nottelmann 2007)[12]. How we gain new information and how we deal with it is in a certain sense up to us and has an epistemically significant influence on our pursuit of epistemic goals. Thus, by bringing about certain epistemically significant states of affairs voluntarily, epistemic agents are able to influence their pursuit of epistemic goals in an active way. States of affairs, which influence the evidence available for an agent like searching for and collecting information, and states of affairs, which influence how we form or revise our doxastic attitudes and how we deal with new information, are types of epistemically significant states of affairs. Epistemic agents exercise doxastic control by bringing about such states of affairs voluntarily. More generally,

[12] Nottelmann gives an interesting overview over exercisable kinds of doxastic control (cf. Nottelmann 2007).

an epistemic agent exercises *doxastic control* by voluntarily bringing about states of affairs, which have an epistemically significant impact on her *doxastic practices* or her belief-system. *Doxastic practices* of an epistemic agent consist in dispositions, methods, cognitive processes of belief-formation and –revision, and of actions, which influence the evidence available to the agent. I will call the ability to exercise doxastic control over one's doxastic attitudes *doxastic agency*.

Epistemic agents can exercise their doxastic agency properly or negligently. The way, in which epistemic agents exercise their doxastic agency, has an influence on how well they pursue the epistemic goals. I claim that epistemic duties guide the appropriate exercise of doxastic agency. Thus, the objects of epistemic duties are epistemically significant states of affairs which can be brought about voluntarily by epistemic agents. One might wonder whether such an approach to epistemic duties is still genuinely epistemic or merely prudential or moral.

Are duties, which have epistemically significant states of affairs as their objects that can be brought about voluntarily by epistemic agents, genuine epistemic norms? To answer this question, we first need to know what epistemic norms are. In the following, I will give two different approaches to epistemic norms. The first approach claims that the states of affairs (φ), which are required by an epistemic norm Oφ, are (types of) doxastic attitudes. The second approach defines epistemic norms from the perspective of epistemic goals or aims.

The first approach claims that Oφ is an epistemic norm only if φ is a (kind of) doxastic attitude (cf. Feldman 2002). According to this approach, the norm of evidentialism would be an epistemic norm, since it requires from an epistemic agent to form the doxastic attitude toward a proposition p at t which fits the evidence the agent has at t. The duties, whose objects are non-doxastic, but epistemically significant states of affairs would not count as epistemic norms within such an approach, since they do not have doxastic attitudes as their objects. But also the norms of externalist justification would not count as epistemic norms within such an approach. Externalist norms of epistemic justification do not require to have a certain (kind of) doxastic attitude toward a certain proposition p given a certain evidential situation, but they require to form doxastic attitudes in certain ways. For example, traditional reliabilism requires from agents to form doxastic attitudes via reliable processes or methods (cf. Goldmann 1979) and the norms of agent-reliabilism require from epistemic agents to form doxastic attitudes by exercising epistemic virtues for being epistemically justified (cf. Greco 1999). Since such externalist norms of epistemic justification do not have (kinds of) doxastic attitudes as their objects, they would not count as epistemic norms as well. This is counterintuitive. It is common to take such externalist norms of epistemic justification as genuinely epistemic, since satisfying these norms is supposed to be necessary for gaining knowledge, at least according to externalism about justification. Thus, I conclude that having a (kind of) doxastic attitude as an object is not a necessary condition for Oφ to be an epistemic norm.

The second approach characterizes epistemic norms in an alternative way. A norm can be characterized as epistemic, if it serves to pursue epistemic goals or aims (cf. Johnson and Hall 1998: 130). According to this approach, one can define the notion of an epistemic norm as follows: a norm is epistemic if and only if the satisfaction of the norm has a positive and epistemically relevant influence on the doxastic practices of the agent or her belief-system under normal conditions. As I have argued before, by exercising doxastic control epistemic agents influence their pursuit of epistemic goals. Thus, the duties which guide the proper exercise of doxastic agency satisfy the alternative definition of being an epistemic norm and are, thus, genuinely epistemic.

Since, within the approach given above, epistemic duties are supposed to guide the intellectual conduct of epistemic agents, I will take epistemic duties to be regulative epistemic norms. Moreover, according to this approach, epistemic duties are duties which take epistemically significant states of affairs as their objects, which can be brought about

voluntarily by epistemic agents. I have argued that such an approach to epistemic duties does not fall prey to the argument against epistemic duties given by Alston (1988). Because, even though I assume that an epistemic "ought implies can" principle holds for epistemic duties, I am not committed to doxastic voluntarism, since the second premise of Alston's argument does not hold within the approach of epistemic duties described in this section.

## 3. What are Epistemic Duties?

In section two I have presented an alternative account for epistemic duties, for which an epistemic "ought implies can" principle holds and which does not fall prey to Alston's argument. In what follows, I will characterize in more detail what epistemic duties are according to this approach and how they serve as criteria of epistemic evaluation and guidance principles for the intellectual conduct of epistemic agents. For this aim, I give necessary conditions for Oφ to be an epistemic duty.

Since epistemic duties are supposed to be epistemic norms, they are criteria of epistemic evaluation. However, epistemic duties are supposed to guide the intellectual conduct of epistemic agents as well. That is why I suppose that in contrast to the norms of epistemic justification, epistemic duties are not criteria to evaluate doxastic attitudes, but rather that they are criteria to evaluate epistemic agents for how they exercise their doxastic agency. Therefore, epistemic duties are supposed to ground the ascription of epistemic blame or epistemic praise[13] rather than the evaluative notion of epistemic justification. For arguments why the ascription of epistemic praise or blamelessness is neither sufficient nor necessary for being epistemically justified see Alston (1988) and Engel (1992). Nottelmann argues that epistemic blameworthiness justifies reactive attitudes, which epistemic agents have toward how an epistemic agent has exercised her doxastic agency (cf. Nottelmann 2007: 3). If epistemic duties allow the ascription of epistemic blame to an agent in case the agent violates an epistemic duty without an excuse, the idea is that the epistemic duties can guide the intellectual conduct of the agent, at least indirectly by the criticism of other epistemic agents toward the way in which the agent has exercised her doxastic agency.

In ethics it is common to assume that it is unfair to blame an agent for the occurrence of a state of affairs φ, whose bringing about was not under the agent's direct or indirect voluntary control. That is why duties, which ground ascriptions of blame, have to satisfy an epistemic "ought implies can" principle. Thus, within my approach an *epistemic "ought implies can" principle* holds for epistemic duties.

(1) If Oφ is an epistemic duty, then Oφ is only incumbent on an epistemic agent, if the agent is able to bring about φ voluntarily[14].

Thus, the holding of an epistemic "ought implies can" principle for epistemic duties is necessary for epistemic duties to ground the ascription of epistemic blame. Therefore, it is necessary for epistemic duties to guide the intellectual conduct of epistemic agents at least indirectly through the ascription of epistemic blame (criticism) by other epistemic agents.

Moreover, epistemically significant states of affairs, whose occurrence is either inevitable or impossible, need to be precluded as objects of epistemic duties. Norms, which have inevitable or impossible states of affairs as their objects, are neither viable criteria of epistemic evaluation nor are they viable guidance principles, since the former are always satisfied,

[13] Alston (1988) as well as Booth and Peels (2010) give reasons, why epistemic duties ground the evaluation of epistemic blame and blamelessness rather than epistemic praise and praiselessness. I cannot go into further detail here. I agree with them.

[14] For a refined version of an "ought implies can" principle, which is sensitive to time-parameters and certain aspects of the situation the agent is in, see Howard-Snyder (2006).

whereby the latter do not apply ever to any epistemic agent because of the "ought implies can" principle. That is why epistemic duties have to be *non-trivially satisfiable*. Oφ is *non-trivially satisfiable* if and only if there exists at least one kind of doxastic control[15], which is such that (i) epistemic agents are able to exercise it under normal conditions, and (ii) epistemic agents are able to bring about φ voluntarily (directly or indirectly) by exercising this kind of doxastic control.

(2) If Oφ is an epistemic duty, then Oφ is non-trivially satisfiable.

As I mentioned before, epistemic duties are supposed to be principles which guide the intellectual conduct of epistemic agents. This distinguishes epistemic duties from mere epistemic norms. One indirect way in which epistemic duties guide the intellectual conduct of an agent is by the reactive attitudes or (proper) criticism of other agents toward how the agent exercises her doxastic agency. One might wonder whether epistemic duties can also guide our intellectual conduct in a more direct way. In my opinion, an epistemic agent's exercise of doxastic agency is directly guided by an epistemic duty if and only if the agent is following or "acting" upon the epistemic duty. Thus, I suppose that for an epistemic duty to guide our intellectual conduct directly, it is a necessary condition that it is possible to "act" upon or to follow the epistemic duty. The *possibility to "act" upon or to follow an epistemic duty* is understood in the following way. It is *possible to follow or to "act" upon Oφ* if and only if for all epistemic agents it holds that if the epistemic agent knows that Oφ is incumbent on her in a certain situation, then it is possible that the agent follows or "acts" upon Oφ. To follow or to "act" upon an epistemic duty in a certain situation implies that one knows that Oφ is incumbent on one in this situation.

Admittedly, my notion of *the possibility to follow or to "act" upon an epistemic duty* is rather weak, since the possibility to follow an epistemic duty in a situation is conditional on the agent's knowledge that the duty is incumbent on her in that situation. A much stronger notion of the possibility to follow an epistemic duty would claim that an epistemic duty Oφ is only incumbent on an epistemic agent in a certain situation, if it is possible for the agent to follow this duty in this situation. The stronger version of this notion makes it – in my opinion – too easy for epistemic agents to be released from their epistemic duties in a certain situation.

There might be some worries regarding my weak notion of the possibility to follow an epistemic duty as well. My weak notion allows that there are some (but not all) epistemic agents, on whom an epistemic duty is incumbent in a certain situation, but they can't follow the duty in this situation. It might appear unfair that it is possible that an epistemic duty is incumbent on an epistemic agent, who is not able to follow this duty. I will evade this worry with the following considerations. In my opinion, there are two kinds of situations in which it is not possible for an epistemic agent to follow an epistemic duty Oφ, which is incumbent on the agent and which can be followed, if known. The first kind of situations are situations in which the agent is faultlessly not able to bring about the required state of affairs φ voluntarily and, thus, cannot follow Oφ due to her incapability. I agree that it would be a worrisome conclusion if, within my approach of epistemic duties, it would follow that an epistemic duty Oφ is incumbent on an agent, even though the agent is faultlessly not able to bring about φ voluntarily. But if the epistemic agent is faultlessly not able to bring about φ, then the epistemic duty Oφ is not incumbent on the agent due to the epistemic "ought implies can" principle, which holds for epistemic duties within my approach.

The second kind of situations are situations in which the agent doesn't know that the duty is incumbent on her, even though the duty is incumbent on the agent in the very situation. Since following an epistemic duty implies knowledge that the duty is incumbent on one, not to know that an epistemic duty is incumbent on one in a certain situation, in which the duty is

[15] For an overview of exercisable kinds of doxastic control, see Nottelmann (2007).

fact incumbent on one, implies that one cannot follow the duty in this situation. My weak notion of the possibility to follow an epistemic duty allows that there is an epistemic agent who is in a certain situation and does not know that an epistemic duty is incumbent on her in this situation, but the duty is nevertheless incumbent on her. If not-knowing that an epistemic duty Oφ is incumbent on one would release one from Oφ, it appears to me to be too easy to get released from one's duties. In my opinion not-knowing does not per se release one from epistemic blame or criticism.

To guide epistemic agents more directly, I claim that the following assumption holds within my approach to epistemic duties.

(3) If Oφ is an epistemic duty, then for all epistemic agents it holds that if the epistemic agent knows that Oφ is incumbent on her in a certain situation, then it is possible for the agent to follow or to "act" upon Oφ.

There is at least one more problem which arises from the weak notion of the possibility to follow an epistemic duty given above. If there are unknowable epistemic duties, i.e. epistemic duties which are such that no epistemic agent at any time can know that they are incumbent on her, then such duties trivially satisfy the condition of the possibility to follow the duty. As mentioned before, to follow an epistemic duty implies that one knows that the duty is incumbent on one. Epistemic duties, which cannot be known by anyone at any time, cannot be followed by anyone at any time. If no one can ever follow an unknowable epistemic duty, then it is hard to imagine that anybody can be guided directly by unknowable epistemic duties. To be guided directly by an epistemic duty implies that one follows the duty and for this it is necessary that one knows that the duty is incumbent on one. Thus, unknowable epistemic duties contravene my rationale for taking the possibility to follow an epistemic duty as a necessary condition for epistemic duties. I have argued that the possibility to follow Oφ is a necessary condition for Oφ to be an epistemic duty, since it enables epistemic agents to be guided directly by Oφ. Thus, within my approach of epistemic duties, it is necessary that Oφ is knowable for Oφ to be an epistemic duty.

(4) If Oφ is an epistemic duty, then Oφ is knowable.

An epistemic duty Oφ is knowable if and only if it is possible that there are epistemic agents, who know that Oφ is incumbent on them.

To sum up, to reject the argument of Alston against epistemic duties I have proposed an alternative approach to epistemic duties, in which the assumption that it is a necessary condition for being an epistemic duty to have a (kind of) doxastic attitude as an object does not hold. Within my approach the objects of epistemic duties are states of affairs, which are epistemically significant and can be brought about voluntarily by epistemic agents. I have argued that epistemic duties guide the proper exercise of doxastic agency and that the proper exercise of doxastic agency promotes the pursuit of epistemic goals. This allows me to argue that within my proposed account, epistemic duties are genuine epistemic norms. What distinguishes epistemic duties from epistemic norms is that epistemic duties are not only criteria of epistemic evaluation, but also principles which can guide the intellectual conduct of epistemic agents. For this, epistemic duties need to satisfy certain conditions, which I have explained in the third part of this paper.

For Oφ to be an epistemic duty, it is necessary that (1) Oφ is only incumbent on an epistemic agent S, if S is able to bring about φ voluntarily, and (2) Oφ is non-trivially satisfiable, and (3) for all epistemic agents S it holds that if S knows that Oφ is incumbent on her, then it is possible for S to follow Oφ, and (4) Oφ is knowable.

To conclude my paper I will give some examples of epistemic duties, which are also discussed in the literature. As I have argued in the first part of this paper, the norm of evidentialism is not an epistemic duty, since in the viable approaches to evidentialism an epistemic "ought

implies can" principle does not hold for this norm. In contrast to the evidentialist norm, there are certain kinds of epistemic norms that can be characterized as epistemic duties. For example, epistemic norms, which require from epistemic agents to exercise intellectual virtues, like to be intellectually courageous, to be intellectually fair, to be epistemically careful, and to be open minded (cf. Montmarquet 2008) are epistemic duties within my approach. Epistemic norms which require epistemically significant actions, like to reflect upon one's evidence or to search for and collect more information can also be characterized as epistemic duties within the approach that has been introduced in this paper.

**Andrea Kruse**
Ruhr-Universität Bochum
Andrea.Kruse@rub.de

## References

Alston, W. P. 1988: 'The Deontological Conception of Epistemic Justification', *Philosophical Perspectives* 2, 257–299.

Booth, A. R. and R. Peels 2010: 'Why Responsible Belief is Blameless Belief', *Journal of Philosophy* 107, 257–265.

Buckareff, A. 2006: 'Doxastic Decisions and Controlling Belief', *Acta Analytica*, 21, 102–114.

Chignell, A. (2013), 'The Ethics of Belief', *The Stanford Encyclopedia of Philosophy* (Spring 2013 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2013/entries/ethics-belief/>

Engel, M. 1992: 'Personal and Doxastic Justification in Epistemology', *Philosophical Studies*, 67, 133–150.

Feldman, R. and E. Conee 2001: 'Internalism Defended', *American Philosophical Quarterly* 38, 1–18.

Feldman, R. 2001: 'Voluntary Belief and Epistemic Evaluation', in M. Steup (ed.), *Knowledge, Truth, and Duty. Essays on Epistemic Justification, Responsibility, and Virtue*, New York: Oxford University Press, 77–92.

Feldman, R. 2002: 'Epistemological Duties', in P. K. Moser (ed.), *The Oxford Handbook of Epistemology*, Oxford: Oxford University Press, 362–384.

Goldman, A. I. 1979: 'What is Justified Belief?', in G. Pappas (ed.), *Knowledge and Justification*, Dordrecht: D. Reidel Publishing Company, 1–23.

Goldman, A. I. 2001: 'Internalism Exposed', in M. Steup (ed.), *Knowledge, Truth, and Duty. Essays on Epistemic Justification, Responsibility, and Virtue*, New York: Oxford University Press, 115–133.

Greco, J. 1999: 'Agent Reliabilism', *Philosophical Perspectives* 13, 273–296.

Hall, R. J. And C. R. Johnson 1998: 'The Epistemic Duty to Seek More Evidence' *American Philosophical Quarterly*, 35, 129–139.

Howard-Snyder, F. 2006: '"Cannot" implies "Not Ought"', *Philosophical Studies* 130, 233–246.

Montmarquet, J. 2008: 'Virtue and Voluntarism', *Synthese* 161, 393–402.

Nottelmann, N. 2007: *Blameworthy Belief. A Study in Epistemic Deontologism*, Dordrecht: Springer-Verlag.

Sinnott-Armstrong, W. 1984: ''Ought' conversationally implies 'Can'', *The Philosophical Review*, 93, 249–269.

Steup, M. 1998: 'A Defense of Internalism', in L. Pojman (ed.), *The Theory of Knowledge. Classical and Contemporary Readings*, Belmont: Wadsworth Publishing Company, 373–384.

Steup, M. 2008: 'Doxastic Freedom', *Synthese*, 161, 375–392.

Wedgwood, R. 2002: 'Internalism Explained', *Philosophy and Phenomenological Research* 65, 349–369.

# The Method of Reflective Equilibrium and Intuitions

Julia Langkau

Reflective equilibrium has been considered a paradigm method involving intuitions. Some philosophers have recently claimed that it is trivial and can even accommodate the sort of scepticism about the reliability of intuitions advocated by experimental philosophers. I discuss several ways in which reflective equilibrium could be thought of as trivial and argue that it is inconsistent with scepticism about the reliability of intuitions.

## 1. Introduction

Reflective equilibrium has been considered a paradigm philosophical method involving intuitions. It has been extensively discussed in normative ethics and political philosophy. The key idea as introduced by John Rawls (1971, 1974/1975) for moral and political philosophy is that we test our moral judgments (or intuitions) and moral principles against each other and revise and refine both when they are inconsistent.[1]

Without specifying what exactly they mean by it, philosophers in all areas of research frequently use the term 'reflective equilibrium' when they mention the methods and aims of their inquiry. It has been suggested that 'reflective equilibrium' is nothing more than a metaphor for the rational performance of philosophy: for taking into account all relevant information available and for working out the most plausible, coherent, and comprehensive theory of the subject matter under investigation.[2] Michael DePaul (2011) and Peter Singer (2005) suggest that it could even be compatible with scepticism about the reliability of intuitions, i.e., with a view according to which we ought not take our intuitions into account.

I am interested in the question whether the method of reflective equilibrium (MRE in what follows) is as trivial as some philosophers think or whether it gives us helpful methodological advice. I first present MRE as it has been discussed in moral philosophy (section 2), apply it to a case in epistemology (section 3), and specify what philosophers mean when they say that MRE is trivial (section 4). I then argue that the sceptical view according to which intuitions ought not be taken into account is not compatible with MRE (section 5).

## 2. The Method of Reflective Equilibrium

Two versions of MRE have been distinguished, *narrow* and *wide* MRE.[3] According to narrow MRE, we take (a) a set of considered moral judgments (i.e., judgments made or intuitions had in certain circumstances conducive to the truth of their content) held by a particular person

[1] Even though the term 'reflective equilibrium' was introduced by Rawls, Nelson Goodman was the first to discuss the method behind the name in 'The New Riddle of Induction' (originally 1953) in his *Fact, Fiction, and Forecast* (1955).

[2] E.g., DePaul (1998, 2011), Foley (1993), Singer (2005), and Williamson (2007).

[3] Daniels finds this distinction implicit in Rawls (1971) and explicit in Rawls (1974/1975).

and (b) a set of general moral principles and produce a coherent theory by adjusting either (a) or (b), or both. In wide MRE, we extend the area of considered judgments and principles we take into consideration to reach coherence among our *widest set of beliefs*. The following is the general idea which lies behind wide MRE, according to Norman Daniels:

> The method of wide reflective equilibrium is an attempt to produce coherence in an ordered triple of sets of beliefs held by a particular person, namely, (a) a set of considered moral judgments, (b) a set of moral principles, and (c) a set of relevant background theories. We begin by collecting the person's initial moral judgments and filter them to include only those of which he is relatively confident and which have been made under conditions conducive to avoiding errors of judgment. For example, the person is calm and has adequate information about cases being judged. We then propose alternative sets of moral principles that have varying degrees of 'fit' with the moral judgments. We do not simply settle for the best fit of principles with judgments, however, which would give us only a narrow equilibrium. Instead, we advance philosophical arguments intended to bring out the relative strengths and weaknesses of the alternative sets of principles (or competing moral conceptions). These arguments can be construed as inferences from some set of relevant background theories [...] (1979: 258)

Rawls, Daniels, and other proponents of MRE think that the inclusion of alternative sets of principles and background theories is essential to a method in moral and political philosophy.[4] My focus lies on the question of how MRE applies to areas such as epistemology, metaphysics, or philosophy of language, where the subject matter under investigation does not consist of moral or similar norms. It seems that background theories play a role when we solve conflicts between our intuitions and beliefs in these areas as well. Narrow MRE will therefore not be discussed any further, and I will use 'MRE' for wide MRE in what follows.

Daniels' (1980, 1996, 2003) formulation of MRE can be given in several steps. Following the literature, I take it that 'initial moral judgments' and 'considered moral judgments' can be replaced by 'intuitions', and I will take theories to be the counterparts of moral principles in non-moral philosophy.

> Step 1. *The relevant intuitions*. Amongst all our intuitions, we choose a set of intuitions of which we are relatively confident, and which we had under ideal conditions, such as having adequate information about the subject matter under investigation and being in a state of mind that is not conducive to error.
>
> Step 2. *The best-fitting theory*. We determine a theory that fits best with the set of intuitions, i.e., that is directly supported by the set of intuitions.
>
> Step 3. *Alternative theories*. We determine alternative theories that are not directly supported by the set of intuitions, but do concern the subject matter under investigation. In the case of a counterexample to an accepted philosophical theory, our currently accepted theory will be amongst these theories.
>
> Step 4. *The relevant background theories*. We look for empirical and philosophical background theories that deliver arguments for or against the competing theories.
>
> Step 5. *Restoring coherence*. We use arguments from our background theories and our intuitions to figure out the best and most coherent theory by either disregarding our intuitions, adjusting our accepted theory, or adjusting our background theories, or all three.

[4] See, e.g., Rawls (1971: 49). Goodman (1955), in contrast, defends narrow MRE for the justification of inductive and deductive forms of reasoning.

## 3. Reflective Equilibrium and Counterexamples

Whenever some philosopher comes up with an intuitive counterexample to an accepted philosophical theory, we confront the task of resolving the resultant inconsistency in one or the other way. MRE could be the method we ought to apply to regain consistency. I will go through the steps of MRE as presented above and apply them to a well-known case of inconsistency in epistemology: the JTB theory of knowledge and the Gettier intuitions.

Step 1. Let us imagine an epistemologist Sophie who lives in the year 1963. Sophie has been thinking about knowledge for a long time, and one day she gets to read Edmund Gettier's paper 'Is Justified True Belief Knowledge?'. Here is one of Gettier's cases (for present purposes, Gettier's second case is sufficiently similar that we do not need to describe it as well):

> Suppose that while Smith has strong evidence that his friend Jones owns a Ford, he has no idea where his friend Brown is. Smith randomly selects three place-names and constructs the following three propositions: either Jones owns a Ford or Brown is in Boston; either Jones owns a Ford or Brown is in Barcelona; either Jones owns a Ford or Brown is in Brest-Litovsk. Even though Smith has no idea where Brown is, he is justified in believing each of these three propositions, because he has correctly inferred them from a proposition for which he has strong evidence, namely that Jones owns a Ford. However, Jones does in fact not own a Ford, but is driving a rented car. Unknown to Smith, Brown happens to be in Barcelona. It seems that Smith does not know that either Jones owns a Ford or Brown is in Barcelona, even though it is true, Smith believes that it is true, and Smith is justified in believing that it is true.[5]

Sophie agrees with the description of the case. She has the intuition that Smith does not know that either Jones owns a Ford or Brown is in Barcelona. However, the theory of knowledge that Sophie holds, the JTB theory, predicts that Smith knows that either Jones owns a Ford or Brown is in Barcelona. Hence, the content of Sophie's intuition is clearly inconsistent with what follows from the JTB theory. Suppose that Sophie notices the inconsistency and strives to resolve it. Suppose furthermore that Sophie's intuition meets the relevant criteria: Sophie is relatively confident of the intuition, and she had it under ideal conditions. Her intuition therefore qualifies for MRE (in Daniels' terms, the intuition is a 'considered judgment' as opposed to an 'initial judgment').

Step 2 and step 3. Sophie finds the JTB theory supported by a set of her intuitions about cases where justified true belief seemingly is knowledge. For instance, Sophie has the following intuition: if she hears her cat meowing in the kitchen she is justified in believing that her cat is in the kitchen, and if it is also true that her cat is in the kitchen, then she knows that her cat is in the kitchen. Her intuitions in the Gettier Cases are not part of this set of intuitions, and Sophie now has to determine the theory that fits her Gettier intuitions best as well as alternative theories that compete with the best-fitting theory.

Sophie reads Michael Clark's response to Gettier, 'Knowledge and Grounds: A comment on Mr. Gettier's Paper'. She spends a long time thinking about the cases and about Clark's no-false-lemma reply to Gettier. In both Gettier Cases, Smith's reasoning is based on a false premise. In the case quoted above, Smith gains his justified true belief by reasoning from the justified false belief that Jones owns a Ford. The case Clark gives in reply has it that knowledge is justified true belief where the justification is not based on a false assumption. Sophie also comes up with her own theories, of which the first is a defeasibility analysis of knowledge. The defeasibility account has it that knowledge is justified true belief absent a

[5] Gettier (1963: 14-15).

defeater.[6] The second theory is a causal theory, on which the belief that P is knowledge only if it is appropriately causally connected to the fact that P.[7] The third theory Sophie comes up with involves a simple reliability condition. According to such an account, S knows that P if and only if S's belief that P is true and justified, where S's belief that P is justified if and only if the belief that P was produced by a reliable cognitive process.[8] Sophie then decides that Clark's adjusted JTB theory fits best with her intuitions in the Gettier Cases (step 2). The JTB theory, the causal theory, the defeasibility account, and the reliability account are alternative theories Sophie has to consider (step 3).

Step 4. Sophie needs to think about her background theories which will help her to decide between the JTB theory, Clark's account, and the alternative theories. Which theory of knowledge Sophie chooses certainly depends on what she thinks justification is. Theories of justification are relevant yet do not directly concern the question of what knowledge is. To keep Sophie's case simple, let us suppose that she has externalist views about justification, which means that she will not consider internalist theories of knowledge. Metaphysical accounts of causation might also influence Sophie's choice. Since metaphysics of causation suggests that abstract and future facts cannot be causes, this counts against the causal theory of knowledge. Other background theories will probably influence Sophie's preferences tacitly.

Step 5. What Sophie has done so far should help her to remove the inconsistency between the content of her intuition that Smith does not know the relevant proposition and her accepted theory of knowledge, the JTB theory.

Sophie thinks that Clark's account is the one that fits her Gettier intuitions best, but she now wonders whether it covers other cases as well and tries to come up with a counterexample. After a while of thinking, she comes up with a case where the justification does not rest on a false assumption, similar to Keith Lehrer's (1965) Nogot case. In Lehrer's case, Nogot in S's office has given S evidence that he, Nogot, owns a Ford. S directly moves from his evidence to the conclusion that someone in S's office owns a Ford, without arguing via the assumption that Nogot owns a Ford. Nogot does not own a Ford, however, someone else in the office, Havit, owns one, which makes S's belief true. Let us suppose that Sophie's case is very similar to this case in that the subject's reasoning somehow does not rest on a false assumption. Sophie comes to the conclusion that Clark's analysis is not correct, because it cannot account for some cases that are very similar to the original Gettier Cases.

Let us say that the causal theory is not consistent with Sophie's background theory on causality. The defeasibility analysis she came up with covers more cases than Clark's analysis, but Sophie thinks it is also extremely complicated.

Sophie thinks about the advantages and disadvantages of the JTB theory, the theory she thinks fits her Gettier intuitions best, and the alternative theories. She weighs them against each other and, given her views on justification and causation which she does not want to give up, decides that a reliabilist account of knowledge is the best choice. While Sophie's case is fictional and could have been told differently, it obviously roughly corresponds to how parts of the debate about the Gettier Cases were conducted in the literature, and some philosophers made a similar decision as Sophie.

---

[6] See Lehrer & Paxon: 'A defeasibility condition requires that there is no other true statement, d, such that the conjunction of S's present evidence for p with d would fail to make S justified in believing p.' (1969, p. 230).

[7] See, e.g., Goldman (1967).

[8] See, e.g., Goldman (1979).

## 4. Reflective Equilibrium and the Triviality Charge

Let us now look at two obvious ways in which MRE could be interpreted as non-trivial or misguided. First, some philosophers understand MRE as a theory of justification (e.g., Rawls (1951), Daniels (1979, 1980, 1996), Elgin (1996), Stich (1988), and Goodman (1955)). According to MRE as a theory of justification, the beliefs we reach as a result of applying MRE to a certain topic are thereby justified. For instance, if Sophie comes to the conclusion that Smith in the Gettier Case does not know that either Jones owns a Ford or Brown is in Barcelona, then Sophie's belief that Smith does not have knowledge of this proposition is justified if it is in reflective equilibrium, i.e., if it coheres with all her other beliefs about knowledge as a result of having appropriately followed steps 1 to 5. As such, MRE is most naturally understood as a coherentist theory of justification. According to coherentist theories of justification, our beliefs are justified through their relation to other beliefs.[9] Coherentist theories of justification are controversial, especially for the justification of our beliefs in areas other than moral or political philosophy. Clearly philosophers such as Timothy Williamson (2007) who claim that reflective equilibrium is trivial do not have reflective equilibrium as an account of justification in mind.

What else could MRE be if not a theory of justification? Besides merely aiming for true beliefs, we aim to build our theories or revise our beliefs in a rational manner. Ideally, we want a methodology in the sense of rules that guide us in the process of building a theory or revising our beliefs. MRE could simply be understood as a method of rational belief revision that does not carry any commitment with regard to what exactly we gain when we apply MRE.[10] As a method of rational belief revision, MRE is neutral with respect to epistemic theories such as foundationism or coherentism, and it is neutral with respect to epistemic externalism or internalism as well. Distinguishing a method of rational belief revision from the question whether our beliefs are justified or true accounts for the strong intuition that our opponents sometimes are as rational in holding the beliefs they hold as we ourselves are, even if either our opponents or we ourselves have unjustified and false beliefs.[11] Supposing that we are not in a sceptical situation, MRE as a method of rational belief revision could nevertheless be conducive to the justification of our beliefs. It is prima facie plausible that if a subject S is not deceived by a Cartesian demon and starts out with beliefs that are rational for her to hold, and if she revises her beliefs in a rational way to reach a theory T, then it is more likely that T is justified than that T is not justified. In different terms, if S is not victim of a Cartesian demon and starts out with beliefs that are rational for her to hold, and if she revises her beliefs in a rational way to reach a theory T, then this is at least some evidence for the truth of T. This does not mean that coherence of T is the only or even the best evidence for the truth of T. In what follows, I will take MRE to be a method of rational belief revision.

Here is a second way in which MRE could be interpreted as non-trivial: it could be understood as misguided because it idealizes our practice of revising beliefs. DePaul (2011) claims that this is the most serious problem MRE confronts, and that MRE can be understood

[9] See Daniels (2011).

[10] Philosophers have discussed several ways in which MRE could be a useful method in moral philosophy. Geoffrey Sayre-McCord (1996) mentions three approaches that do not involve the justification of beliefs. According to the first, MRE is a heuristic method, i.e., MRE is useful to discover the fundamental truths of morality, but it does not justify the beliefs reached through its application. According to a second approach, there is a moral obligation to act only upon moral principles that are in reflective equilibrium with all our other beliefs. According to a third and pragmatic view, it is in some sense useful to act upon a principle which is in reflective equilibrium with all other beliefs we hold. See also Kappel (2006), who is pessimistic concerning the role of MRE in epistemic justification, but mentions that a pragmatic or otherwise not truth-related account of MRE might be defendable.

[11] See also Kelly & McGrath (2010), who think that MRE ought to lead to beliefs that are rational for us to hold.

as idealizing our practice with regard to two aspects. First, the order in which we ought to proceed according to MRE does not correspond with what we in fact do: we hardly ever start out with first determining the relevant intuitions (step 1), figuring out a set of theories that matches (step 2), then taking alternative theories into consideration (step 3), and finally using our background theories (step 4) to build the most comprehensive and coherent theory on the subject matter (step 5). We rather 'naturally bring all kinds of considerations [...] into play helter skelter as they occur to us'[12]. Second, the quantity of intuitions and theories we take into account is limited. It is simply not possible for us to take all relevant alternative theories into account (step 3), and it is not possible to test our intuitions concerning all relevant aspects (step 1), because 'one would need to reflect upon and form [intuitions] about far too many kinds of hypothetical cases'[13]. Looking at Sophie's case, one might think this is indeed a problem. Surely the way Sophie applies MRE to the Gettier Cases is an idealization. No single philosopher did apply or could have applied MRE in the way Sophie does: Sophie starts out with step 1 and then goes through steps 2 to 5. She also takes far more theories into consideration in steps 2 and 3 before she moves on to the final step 5 than philosophers did in 1963.

There are two ways to reply to the idealization objection. First, we might say that MRE has to be understood as making claims about what we should ideally do, not as telling us what we ought to do given our temporal and intellectual constraints. MRE says that we should consider all possible alternatives, but this does not mean that it is not rational for us to stop at a certain point to settle for a coherent theory. Let us look at Sophie again. Even though she considers many alternative theories, one could still criticize her for not doing everything she ought to do in order to appropriately follow the advice MRE gives. Maybe Sophie should not have made a decision as to which theory to endorse, since the debate over the Gettier Cases is ongoing and philosophers are still coming up with theories to cover our intuitions in the Gettier Cases and numerous variations of the Gettier Cases. Ideally, Sophie would even be much smarter. However, just as it seems rational for Sophie to stop considering alternative thought experiments and alternative theories, other philosophers could be rational relative to their limits.

It is unproblematic to stop considering alternative theories because MRE can be applied to a minimal set of intuitions and theories. MRE can be re-applied whenever someone comes up with a new theory or a new thought experiment, and the accepted theory can be revised again. This is a reply to both aspects of the idealization objection, the worry that we do not always go through the steps in the right order and the worry that we cannot consider all relevant theories and intuitions. To engage in MRE could simply mean to engage in an ongoing series of applications of MRE. As Rawls claims: once a subject has reached a coherent theory, "[...] this equilibrium is not necessarily stable. It is liable to be upset by further examination [...] and by particular cases which may lead us to revise our judgments [...]"[14].

In fact, as I presented Sophie's case, Sophie comes up with a new thought experiment in step 5, namely with a case that is similar to Lehrer's Nogot case. The reason why I let Sophie come up with this case in step 5 is that she compares different theories in step 5, and hence she thinks about whether Clark's account can accommodate as true all or enough intuitions only in step 5. Strictly speaking, Sophie goes back to step 1. However, it seems that this is not a problem for MRE: the order in which we follow the steps does not seem to be crucial.

A second way to reply to the idealization objection is to say that MRE ought to be applied by a group of researchers instead of a single philosopher. Whereas Rawls, Daniels, and DePaul think that MRE is to be pursued by an individual researcher, Goodman (1955) defends such a

[12] DePaul (2011: xcii).
[13] DePaul (2011: xcii).
[14] Rawls (1971: 18).

collective MRE. The view is that as individual philosophers, we work on different ends: we discover inconsistencies, we develop theories that cover our intuitions best, and we develop background theories. We do all this in much detail, which might require a whole career or life-time. As a community of researchers, we might eventually reach the aim of step 5: one single theory wins. Maybe Sophie might better be understood as representing a group of researchers, not an individual philosopher.

DePaul argues that MRE cannot possibly be understood as the description of what we are supposed to do as a group of philosophers:

> Because of the way revisions are determined, [reflective equilibrium] must be a first-person inquiry. Propositions do not seem true to a group of people except in the derivative sense that they seem true to each member of the group. Any disagreement within a group and there will be nothing to determine how conflicts are to be resolved, and hence, nothing to determine the group's state of [reflective equilibrium]. Moral inquiry can be a joint endeavor according to [reflective equilibrium] only insofar as we agree or insofar as one person is willing to assist another in her individual attempt to bring her beliefs into equilibrium by doing such things as pointing out potential conflicts in her beliefs, presenting examples that might elicit interesting intuitions or proposing theories that might account for her [intuitions]. Alternatively, one might approach some other person as a subject, taking that person's beliefs and seemings as data and attempting to work out what that person's state of [reflective equilibrium] would be. (2011: 1xxxi)

DePaul thinks that disagreement in intuitions makes a collective MRE impossible. However, disagreement might simply show that we have not yet reached the final state of reflective equilibrium. There surely is a lot of disagreement in philosophy, and this disagreement concerns our theories as much as our intuitions. We could nevertheless all be concerned with the same project, namely with finding the best theory of a certain subject matter, and it is still possible that in the end one theory will win in the sense that it will be accepted by everyone.

While I do not think that disagreement makes collective MRE impossible, I agree with DePaul that MRE should be understood as a method an individual philosopher ought to pursue. The reason is that if MRE is what a group of philosophers ought to engage in, it is unclear what the advice for the individual philosopher would be and how we would access whether an individual philosopher is revising her beliefs in a rational way. This, however, is what we are interested in when we talk about a method of rational belief revision. We want such a method to give us advice on how to revise our beliefs as individual philosophers and we want to be able to decide whether an individual philosopher is revising her beliefs rationally. Even if other people's intuitions matter, it does not follow that MRE is a collective enterprise. It is plausible to think that we do not only take our own intuitions into consideration but rather rely on other philosophers with respect to intuitions as much as we do with respect to theories. Endorsing MRE as a method for an individual philosopher does not mean that we cannot rely on work other researchers have done, on their intuitions or on the theories they developed.

If we do not understand MRE as a theory of justification and if we think it does not have to be viewed as an idealization of our practice, the concern is that step 1 to 4 of MRE are trivial and step 5 gives us only very general instructions. According to Rawls, we have to go "back and forth", sometimes to adjust the principles to our judgments, sometimes to conform the judgments to our principles.[15] Similarly vaguely, Daniels claims that we are "expected to revise our beliefs at all levels as we work back and forth among them and subject them to various criticisms"[16]. It seems that MRE collapses into a trivial and uncontroversial claim

[15] Rawls (1971: 20).
[16] Daniels (2003).

about philosophical methodology, as Williamson expresses in the following passage about MRE:

> The question is not whether philosophers engage in the mutual adjustment of general theory and judgments about specific cases—they manifestly do—but whether such descriptions of it are sufficiently informative for epistemological purposes. (2007: 244)

Similarly, Foley thinks that MRE is too general to be useful:

> The problem with this recommendation is familiar. It is not so much mistaken as unhelpful. [...] It tells you essentially this: take into account all the data that you think to be relevant and then reflect on the data, solving conflicts in the way that you judge best. On the other hand, it does not tell you what kinds of data are relevant, nor does it tell you what is the best way to resolve conflicts among the data. It leaves you to muck about on these questions as best you can. (1993: 128)

DePaul (1998, 2011) goes so far as to argue that MRE is the only rational method in philosophy. In a nutshell, DePaul's version of MRE says that we should reflect upon the logical and evidential relations between all our relevant beliefs, and that we should resolve conflicts which might emerge during this process in the best possible way we can figure out. DePaul then argues that it is difficult to think of a rational alternative to MRE thus construed, and that an opponent of MRE would have to make one of the following claims: (A) we should abandon reflection altogether; (B) our method should direct the inquirer to reflect, but to do so incompletely, i.e., to leave certain beliefs, principles, or theories out of account; (C) our method should not allow the results of the inquirer's reflections to determine what the inquirer goes on to believe.[17] Unsurprisingly, DePaul concludes from his discussion of (A), (B), and (C) that any method endorsing at least one of these claims would be irrational.

One way of deciding whether MRE is trivial is to see whether it is compatible with different views concerning the role of intuitions in philosophy. If it is not, then it seems that MRE is not trivial. In the next section, I argue that MRE is not compatible with the sort of scepticism about the reliability of intuitions advocated by experimental philosophers.

## 5. Reflective Equilibrium and Scepticism

Some experimental philosophers draw a radical conclusion from their studies: using intuitions as evidence for or against philosophical theories is an unreliable method which should not be pursued.[18] Some philosophers have claimed that MRE can accommodate scepticism about the reliability of intuitions as advocated by experimental philosophers. I first look at a view according to which sceptical considerations are not supposed to be part of MRE because meta-theories in general are not supposed to be part of MRE (section 5.1). I then look at Singer's scepticism about intuitions (5.2) and at DePaul's view according to which scepticism is compatible with MRE because it could come as a result of applying MRE that we disregard every single intuition (section 5.3). I conclude that MRE is not compatible with scepticism about the reliability of intuitions.

### *5.1 MRE and Meta-Theories*

One way to argue that scepticism about the reliability of intuitions is not compatible with MRE is to exclude meta-considerations from the beginning. DePaul mentions that according to Rawls' original account, meta-considerations and arguments from philosophy of language and metaphysics that have been used for or against metaethical theories such as moral

[17] DePaul (1998: 301).

[18] E.g., Alexander & Weinberg (2007), and Weinberg (2007).

relativism, moral realism, or noncognitivism are not supposed to be considered in MRE. The reason is that while MRE is supposed to help us decide between different moral theories, meta-theories do not bear directly on the moral subject matter under consideration.[19] The same would apply to contemporary empirical research on the reliability of our intuitions, according to DePaul:

> Efforts to use results from psychology or neuroscience or evolutionary theory to question the reliability of some or all of our [intuitions] are now extremely prominent [...] As Rawls conceived of [wide reflective equilibrium], these background theories would not be part of the equilibrium. Because they provide premises for a broad skepticism regarding morality, they would not serve as premises of arguments for, or against, any particular [moral theory]. (2011: lxxxix)

Exactly the same would be true for results from psychology or experimental philosophy in epistemology. Take Sophie and her epistemic intuitions. Sophie is interested to know whether Smith has knowledge in the Gettier Case, and more generally what knowledge is. Questions concerning the relevance of empirical research on intuitions do not bear directly on the question whether Smith has knowledge of the relevant proposition or not, so it seems not to help Sophie to answer her question.

DePaul mentions that Rawls' main reason to exclude meta-theories from MRE is that he is interested in MRE as a method to detect our moral sensibilities rather than to detect the truth about moral issues. In order to determine which moral theory captures our moral sensibilities best, meta-theories are obviously not relevant. To compare it with a simple case: if I am interested to find out about my food preferences and notice that I do not like broccoli, the fact that broccoli is healthy and it would be much better for me to like broccoli is irrelevant to my concern.

However, if we are ultimately interested in the truth of our theories rather than merely in the systematization of our beliefs and intuitions, the question of whether our intuitions are reliable is highly relevant. Hence, if MRE told us to ignore doubts as to their reliability, it could not be a method of rational belief revision. Excluding meta-considerations such as scepticism about the reliability of intuitions from the beginning is thus not an option.

### *5.2 MRE without Intuitions*

In line with some experimental philosophers, Singer (2005) draws sceptical conclusions from recent empirical studies and defends a view according to which we should not assign any plausibility to our moral intuitions. Singer objects to "any method of doing ethics that judges a normative theory either entirely, or in part, by the extent to which it matches our moral intuitions"[20].

As an example of evidence for the insignificance of moral intuitions, Singer discusses Joshua Greene et al.'s (2001) studies on intuitions people have when confronted with different versions of the Trolley Case. In one version of the Trolley Case, we are asked whether a fat man should be pushed down a bridge to stop the trolley, in which case only one person dies and five people who would otherwise be killed survive. Most of us have the intuition that pushing the fat man off the bridge would be wrong. In another version, we are asked if the driver of the trolley should side-track the trolley, which would again kill only one person instead of five people. Most of us have the intuition that the driver should side-track the trolley. Greene et al. (2001) conducted brain scans of subjects while they had intuitive reactions to cases very similar to the two versions of the Trolley Case. The results suggest that different intuitive responses have to be explained by differences in the emotional pull of

[19] DePaul (2011: lxxxix).

[20] Singer (2005: 346).

situations which involve causing someone's death in a close-up, personal way vs. causing someone's death in a way which is at a distance and less personal.[21] Singer thinks that more research is likely to show that Greene has not only explained, but rather explained away the philosophical puzzle of why our intuitions in different versions of the Trolley Case differ. Based on Greene et al.'s results and arguments from evolution[22], Singer draws the following sceptical conclusions for moral intuitions and for MRE:

> [R]ecent scientific advances in our understanding do have some normative significance, and at different levels. At the particular level of the analysis of moral problems like those posed by trolley cases, a better understanding of the nature of our intuitive responses suggests that there is no point in trying to find moral principles that justify the differing intuitions to which the various cases give rise. Very probably, there is no morally relevant distinction between the cases. At the more general level of method in ethics, this same understanding of how we make moral judgments casts serious doubt on the method of reflective equilibrium. There is little point in constructing a moral theory designed to match considered moral judgments that themselves stem from our evolved responses to the situations in which we and our ancestors lived during the period of our evolution as social mammals, primates, and finally, human beings. (2005: 348)

We are interested in the second claim on a more general level, according to which empirical studies cast doubt on MRE. Even though Singer does not think it would be a good idea, he mentions that MRE could possibly be interpreted "wide enough" to accommodate a practice where we do not take any of our intuitions into account. In that case, MRE might be compatible with the idea that intuitions should not play any role in philosophy, but it would no longer be a distinctive method for normative ethics.[23]

Scepticism about the reliability of intuitions entails *prima facie* that step 1 is misguided: we ought not take our intuitions into account. In a concrete case of a thought experiment such as the Gettier Cases, this means that a subject ought not assign any plausibility to the fact that she has an intuition that P or to the content of the intuition that P. In general terms:

**No Plausibility**

The subject assigns no plausibility to her intuition that P.

A method that rules out intuitions (either the fact that we have an intuition that P or the content of the intuition that P) as evidence from the beginning does not seem to be compatible with MRE, at least not with MRE as presented in section 2. According to MRE, we remove an inconsistency between an intuition and our accepted theory either by disregarding the intuition and revising our judgment or by adjusting our accepted theory (or our background theories, or all three). The crucial point is that two options are available with respect to the intuition that P: either to accommodate it as true or to disregard it. In principle, MRE leaves Sophie the choice to either endorse that Smith has no knowledge, or to reject it and either stick with the JTB theory or endorse a different theory. Which option she chooses depends on her background theories and on how good the alternative theories she comes up with are, but neither is ruled out from the beginning.

---

[21] Greene et. al (2001: 2106).

[22] According to Singer, "[...] the salient feature that explains our different intuitive judgments concerning the two cases is that the footbridge case is the kind of situation that was likely to arise during the eons of time over which we were evolving; whereas the standard trolley case describes a way of bringing about someone's death that has only been possible in the past century or two, a time far too short to have any impact on our inherited patterns of emotional response." (2005: 348).

[23] Singer (2005: 347).

Proponents of MRE might disagree about how exactly the process of restoring coherence in step 5 ought to take place. We said that background theories are relevant to the areas we are interested in, i.e., the relevant MRE is wide MRE. This means that we use arguments from background theories in step 5, and it also means that background theories are amongst the theories that can be adjusted. However, that there are two ways of dealing with the intuition seems to be widely shared. Here is how Ernest Sosa puts it (using 'principle' for 'theory'):

> If a conflict pits the intuitive pull of an example against the tug of a familiar principle, we seek to remove or revise one or the other, so as to remove the tension. Sometimes the particular intuition(s) will win, but sometimes the tug of the principle must prevail. (1989: 262)

Roy Sorensen writes (using 'theoretical principles' for 'theories' and 'atheoretical judgment' for 'intuition'):

> You attain reflective equilibrium when your theoretical principles cohere with your atheoretical judgments. To reach this state, you must remove conflicts between them. Sometimes the conflict is resolved by giving up the principle and sometimes by giving up the intuition. (1992: 83)

It seems that step 5 of MRE, applied to thought experiments as counterexamples to philosophical theories, entails the following methodological claim with respect to intuitions.

> **Coherence**
>
> In case of a conflict between our intuitions and our accepted theory on a certain subject matter it is sometimes permissible to disregard an intuition that P and sometimes permissible to adjust theory to accommodate as true an intuition that P.

If we endorse No Plausibility, Coherence must be false. Since according to scepticism about the reliability of intuitions, we ought not assign any plausibility to our intuitions, it is not true that we have two options to regain coherence. According to this view, thought experiments simply would not be potential counterexamples and would not have to be taken into consideration at all. Sophie, for instance, would have to disregard her intuition in the Gettier Case and stick with her accepted theory, the JTB theory of knowledge—or use other kinds of evidence to support a different theory of knowledge. According to MRE, however, we have to assign at least some initial plausibility to our intuitions. Hence, if scepticism about the reliability of intuitions entails No Plausibility, it is not compatible with MRE.

### *5.3 Scepticism as a Consequence of MRE*

However, there might be a way to account for scepticism about the reliability of intuitions that does not entail No Plausibility. Let us look at DePaul's view. In contrast to Singer, DePaul thinks that moral intuitions play an important role, but he agrees that MRE could accommodate scepticism about the reliability of intuitions.[24] As a consequence, DePaul's characterization of MRE is indeed so general that it could be true of any kind of truth-directed activity: the essence of MRE is that it 'directs one to leave nothing out of consideration and to believe what seems likely to be true upon due consideration'.[25] It seems that MRE understood as broadly as this can no longer count as a method of rational belief revision since it does not entail any advice as to how to revise our beliefs.

DePaul thinks that scepticism about the reliability of intuitions could come as a result of applying MRE. If we take empirical research on particular intuitions such as psychological and neuroscientific evidence into account, it could turn out that we end up rejecting every

[24] DePaul (2011: c).

[25] DePaul (2011: cii).

single one of our intuitions. This seems to be what DePaul has in mind when he says the following:

> Suppose the data provided by the research seems much more likely to be true to S than any of her intuitive moral judgments, and that she follows the arguments from the data and has no doubts about them. So, S excises all normative moral beliefs from her overall system. She ends up accepting no moral theory; she makes no moral judgments. She only has beliefs about morality, e.g., that all the moral judgments she previously made were mistaken and that all the moral judgments other people make are mistaken. Does it follow that S would have abandoned the method of [reflective equilibrium]? Not at all—she would have done exactly what that method dictates. (2011: c)

Since he thinks that all moral theory in the end amounts to intuitions, DePaul concludes that if the subject rejected all her moral intuitions, she would have no moral views at all. We would not have to take this extreme stance on the role of intuitions in non-moral philosophy, but we could imagine ending up disregarding all intuitions from thought experiments, because it turned out that we have reasons to do so for every single intuition.

DePaul's idea is that since it would come as a result of applying MRE, disregarding all intuitions would be compatible with MRE. However, such a result would certainly raise serious doubts about the reliability of intuitions more generally. We would then have to adjust our meta-philosophical view: it would hardly be rational to consider further intuitions, and we would rather have to assign no initial plausibility to intuitions anymore. We would have to endorse No Plausibility, which entails the denial of Coherence (as discussed above). Hence, if MRE leads to the rejection of every single of our intuitions, it is self-defeating. However, as long as this does not lead us to disregard all our intuitions, MRE is compatible with taking empirical evidence about the reliability of particular intuitions into account.

## 6. Conclusion

The question I aimed to answer in this paper was whether MRE understood as a method of rational belief revision reduces to a trivial claim about philosophical methodology. I argued that contrary to what some philosophers think, MRE is not compatible with scepticism about the reliability of intuitions and is thus not trivial.

The claim that we ought to assign some plausibility to our intuitions is ambiguous. It can either mean that we take the content of the intuition that P or that we take the fact that we have an intuition that P to be evidence. In order to understand the role of intuitions in MRE, more work has to be done to specify how exactly they are supposed to come into play.

**Julia Langkau**

University of Zurich
julialangkau@gmail.com

## References

Alexander, J., and J. Weinberg 2007: 'Analytic epistemology and experimental philosophy', *Philosophy Compass* 2, 56–80.

Daniels, N. 1979: 'Wide reflective equilibrium and theory acceptance in ethics.' *The Journal of Philosophy* 76, 5 (1979), 256–282.

— 1980: 'Reflective equilibrium and archimedean points', *Canadian Journal of Philosophy* 10, 83–103.

— 1996: *Justice and Justification: Reflective Equilibrium in Theory and Practice*. New York: Cambridge University Press.

— 2003: 'Reflective equilibrium', http://plato.stanford.edu/entries/ reflective-equilibrium/ (2003; substantive revision 2011).

DePaul, M. R. 1998: 'Why bother with reflective equilibrium?', in *Rethinking Intuition. The Psychology of Intuition and Its Role in Philosophical Inquiry*, Rowman and Littlefield Publishers, Inc., 293–309.

— 2011: 'Methodological issues: Reflective equilibrium', in *The Continuum Companion to Ethics*, London, New York: C. Miller, Ed. Continuum International Publishing Group, Ixxv–cv.

Elgin, C. Z. 1996: *Considered Judgment*. Princeton University Press.

Foley, R. 1993: *Working Without a Net: A Study of Egocentric Epistemology*. New York: Oxford University Press.

Gettier, E. L. 1963: 'Is justified true belief knowledge?', *Analysis* 23, 121-123.

Goldman, A. 1967: 'A causal theory of knowledge', *Journal of Philosophy* 64, 357-372.

— 1979: 'What is justified belief?', in *Justification and Knowledge: New Studies in Epistemology*, Boston, Dordrecht, London: Reidel, 1–23.

Goodman, N. 1955: *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Greene, J. D., R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen 2001: 'An fMRI investigation of emotional engagement in moral judgment', *Science* 293, 2105–2108.

Kappel, K. 2006: 'The meta-justification of reflective equilibrium', *Ethical Theory and Practice* 9, 131–147.

Kelly, T., and S. McGrath 2010, 'Is reflective equilibrium enough?', *Philosophical Perspectives Volume* 24, 325–359.

Lehrer, K. 1965: 'Knowledge, thruth, and evidence', *Analysis* 25, 168–175.

Lehrer, K., and T. Paxon 1969: 'Knowledge: Undefeated justified true belief', *Journal of Philosophy* 66, 225–237.

Rawls, J. 1951: 'Outline of a decision procedure for ethics', *The Philosophical Review* 60, 177–197.

— 1971: *A Theory of Justice*. Harvard: Cambridge University Press, 1971.

— 1974/1975: 'The independence of moral theory', *Proceedings and Addresses of the American Philosophical Association* XLVII, 5–22.

Sayre-McCord, G. 1996, 'Coherentist epistemology and moral theory', in *Moral Knowledge?*, Sinnott-Armstrong, W. , and M. Timmons (eds.), Oxford University Press, 137–189.

Singer, P. 2005, 'Ethics and intuitions', *The Journal of Ethics*, 331–352.

Sorensen, R. 1992: *Thought Experiments*. New York: Oxford University Press.

Sosa, E. 1989: 'Equilibrium in coherence?', in *The Current State of the Coherence Theory*, J. Bender (ed.), Kluwer: Academic Press, 242–250.

Stich, S. 1988: 'Reflective equilibrium, analytic epistemology and the problem of cognitive diversity', *Synthese* 74.

Weinberg, J. 2007: 'How to challenge intuitions empirically without risking skepticism', *Midwest Studies in Philosophy* 31, 318–343.

Williamson, T. 2007: *The Philosophy of Philosophy*. Blackwell Publishing Ltd.

# Why Know-how and Propositional Knowledge Are Mutually Irreducible

David Löwenstein

The distinction between knowing how to do something and knowing that something is the case is a piece of common sense. Still, it has been suggested that one of these concepts can be reduced to the other one. Intellectualists like Jason Stanley (2011) try to reduce know-how to propositional knowledge, while practicalists like Stephen Hetherington (2011) try to reduce propositional knowledge to know-how. I argue that both reductionist programs fail because they make the manifestations of the knowledge to be reduced unintelligible. Contrary to both, I suggest that know-how and propositional knowledge are distinct, but conceptually interdependent.

## 1. Introduction

The distinction between knowing how to do something and knowing that something is the case is a piece of common sense. Still, it has been suggested that one of these concepts can be reduced to the other one. Intellectualists like Jason Stanley try to reduce know-how to propositional knowledge (cf. Stanley and Williamson 2001, Stanley 2011a and 2011b), while practicalists like Stephen Hetherington try to reduce propositional knowledge to know-how (cf. Hetherington 2006, 2008, and 2011). Both views have been worked out in much detail on which I cannot comment here. But my arguments are independent from these issues.

I argue that both reductionist programs fail because they make the manifestations of the knowledge to be reduced unintelligible. Contrary to both, I suggest that know-how and propositional knowledge are distinct, but conceptually interdependent. Before substantiating these points, I start with some pre-theoretic remarks about know-how.

## 2. Know-how

The concept of know-how has its point in explaining what Gilbert Ryle, the grandfather of the current debate, calls 'intelligent practice'. He writes:

> What is involved in our descriptions of people as knowing how to make and appreciate jokes, to talk grammatically, to play chess, to fish, or to argue? Part of what is meant is that, when they perform these operations, they tend to perform them well, i.e. correctly or efficiently or successfully. Their performances come up to certain standards, or satisfy criteria. But this is not enough. [...] To be intelligent is not merely to satisfy criteria, but to apply them; to regulate one's actions and not merely to be well-regulated. (Ryle 1949: 29)

I take it that this expresses the common sense view: First, know-how is a capacity to perform an activity *well* – that is, a capacity to succeed in that activity, to meet its standards. But second, not every capacity to meet the standards of an activity amounts to know-how. Not every such capacity is intelligent. Know-how involves an *understanding* of what the activity in

question demands – an understanding of its standards. Without such an understanding, one could only possess a mere ability or a mere disposition.

Know-how, by contrast, requires being *guided* by these standards. It is a *skill*, an intelligent ability – that is, an ability to perform well in an activity in virtue of one's understanding of the standards which govern it.

Given this background, I shall now turn to my criticisms of intellectualism and practicalism.

## 3. Against Intellectualism

Intellectualism is the view that know-how is a species of propositional knowledge. Roughly, it holds that knowledge how to A is knowledge that one can engage in A-ing in certain ways with which one is practically acquainted. But intellectualism fails because it leads to a vicious regress in the explanation of the manifestation of know-how.

Ryle has proposed to argue along these lines (1945; 1949), but I cannot adequately discuss the different possible interpretations of his texts here (cf. Löwenstein 2011). Instead, I shall present what I take to be the best version of the Rylean regress argument.

Suppose that a manifestation of know-how – like fishing successfully or drawing a correct inference – just is a manifestation of propositional knowledge. Then, the intelligence of these performances stems from the *application* of this propositional knowledge to the case at hand. After all, *having* propositional knowledge does not necessarily entail that it always bears on practice. But applying propositional knowledge is *itself* something one may do intelligently or not. Thus, it must itself be understood as an exercise of know-how.

Unfortunately, intellectualism requires us to also reduce *these* instances of know-how to propositional knowledge. This leads to an infinite chain of instances of propositional knowledge and leaves us with an inadequate account of the manifestation of know-how. Thus, intellectualism is false.

In order to make this argument as clear as possible, I present an explicit reconstruction:

1. The explanation of S's intelligently A-ing must involve appeal to S's employing their knowledge how to A such that S intelligently As.

2. RA: Intellectualism: Knowledge how to A is, for certain ways of acting $\varphi$ , knowledge that $\varphi(A)$.

---

3. The explanation of S's intelligently A-ing must involve appeal to S's employing their knowledge that $\varphi(A)$ such that S intelligently As.

4. If S employs knowledge that p such that S acts intelligently, then S intelligently applies the proposition that p to the case at hand.

---

5. The explanation of S's intelligently A-ing must involve appeal to an infinite number of instances of S's intelligently applying propositions to cases – namely that $\varphi_1(A_1)$, that $\varphi_2(A_2(\varphi_1(A_1)))$, and so on ad finitum – where $A_{n+1}$ refers to the activity of intelligently applying the knowledge that $\varphi_n(A_n)$.

6. No explanation of a subjet's acts may involve appeal to their execution of an infinite number of other acts.

---

7. Intellectualism is false.

Of course, premise 4 is the most crucial element of this argument. Everything depends on the question what 'acting intelligently' and 'intelligently applying propositions' come down to.

Stanley has rightly stressed that the commonplace understanding of the objection only attacks a straw man – that it over-intellectualizes intellectualism. This is because premise 4 is understood to hold that the application of a proposition to a case involves separate mental acts of considering and applying propositions, and that these acts are intentional actions. But there is clear phenomenal support for the view that the application of a proposition to a case often proceeds automatically and unintentionally (cf. Ginet 1975: 6-7). Thus, any prima facie plausible view – intellectualist or not – will deny that applying propositions to cases *always* involves separate intentional acts (cf. Stanley 2011: 14).

But unfortunately, this is a red herring: Stanley's objection only attacks premise 4 if intelligent practice is understood as intentional practice. But I have already pointed out that what Ryle calls 'intelligent practice' are activities which are regulated by standards such as efficiency, success, and correctness. Intelligence, in short, is being guided by norms.

Is applying propositions to cases 'intelligent' in this sense? Yes. It is possible to make mistakes in applying propositions to a case and it is possible to do so better or worse. Thus, this activity is clearly governed by norms. Ryle provides paradigm cases of people who exhibit such failures and shortcomings – e.g. the chess player and his maxims and the dull student of reasoning and his logical rules (cf. Ryle 1945: 5-7).

Stanley agrees about these points but disagrees about their consequences. He understands applying a proposition to a case as an automatic triggering of a representation of the proposition in question:

> Triggering a representation can certainly be done poorly or well. But this does not show that it can be done intelligently or stupidly. [...] Since triggering a representation is something we do automatically, [...] [premise 4 in the above reconstruction (D.L.)] results in a manifest implausibility. (Stanley 2011: 16)

Thus, Stanley thinks that an activity which is performed well, but automatically, does not qualify as intelligent. This is a puzzling view, since automaticity and intelligence certainly go together in many important cases.

Take, for instance, my knowledge how to read. I often read intentionally, but I also often read unintentionally and automatically – say, when I happen to see a sign in the street. But both are genuine exercises of my know-how. Both are governed by the same norms. Also, reading is not unique at all: We sometimes draw inferences or calculate sums automatically and unintentionally – according to internalized logical or mathematical principles. Thus, Stanley's view that automaticity excludes intelligence and thereby blocks the regress is mistaken.

However, this might be just another red herring. Stanley could simply bracket the question of automaticity and intentional action and hold that, in my terms, the application of propositions is a case of *mere ability* as opposed to intelligent know-how. But this last option also fails. As Ellen Fridland (2012) has beautifully shown, the capacities on which intellectualism must rely are clear cases of intelligent skills. For they must somehow make distinctions within all the available information and determine which piece of propositional knowledge would be the best guide in the current situation. And they must ensure that the application of this piece of knowledge actually results in an intelligent performance.

To illustrate, the propositional knowledge to which know-how is allegedly reducible can be individuated in a coarse-grained way or in a fine-grained way. But Fridland shows that either option causes serious trouble for intellectualism.

On the coarse-grained reading, different people can have the same know-how in virtue of knowing the same propositions, and one can put the same know-how, the same propositions, into practice on different occasions. But then, it becomes an open question how exactly such coarse-grained knowledge can guide a person through the endless particularities of any given situation. And whatever does this work must be intelligent.

On the fine-grained reading, know-how is reduced to great numbers of pieces of propositional knowledge, each specifying how something can be achieved for an individual person in a particular situation. But then, it becomes an open question how exactly the application-process selects one proposition from this vast number of ever so slighty different pieces of knowledge. Again, whatever does this work must be intelligent.

The intellectualist reply under consideration would have it that competent people merely happen to do these things well without being guided by an understanding of what it takes to do them well. But this is absurd. Competences to adjust to the specificities of cases are at the heart of intelligent practice.

I conclude that the regress argument stands to scrutiny. Intellectualism fails.

## 4. Against Practicalism

Let me now turn from intellectualism to practicalism – the view that propositional knowledge is a species of know-how. Roughly, it holds that knowledge that p is knowledge how to engage in the activities in what Hetherington calls "p's epistemic diaspora" (2011: 37) – a loose and open-ended list of activities including accurately asserting that p, basing decisions upon the truth of p, and so forth. But practicalism is bound for a complementary infinite regress in the explanation of the manifestation of propositional knowledge.

As seen above, propositional knowledge can be intelligently applied. But, more broadly, propositional knowledge manifests itself when a subject is in some way or other informed by her knowledge – that is, when she intelligently acts in the light of this knowledge.

Suppose that such a manifestation of propositional knowledge just is a manifestation of know-how. However, a manifestation of know-how must be understood as a *reflective exercise* of know-how. That is, the subject must employ their understanding of the standards governing the activity.

I have already introduced this pre-theoretic idea above. While I cannot offer a full account of the understanding of an activity's standards here, I shall nevertheless make one a more substantive claim: To understand the standards which govern an activity involves *at least a minimum* of knowledge of the sufficient and necessary conditions for meeting those standards – that is, propositional knowledge of the form 'Ceteris paribus, X suffices for A-ing well' or 'Ceteris paribus, good A-ing is possible only given Y'. Without any such propositional knowledge, it is impossible to understand the standards of A-ing.

Thus, know-how is not *exhausted* by propositional knowledge – as intellectualism would have it. However, it entails at least *some* propositional knowledge.

Unfortunately, practicalism requires us to also reduce *these* instances of propositional knowledge to know-how. This leads to an infinite chain of instances of know-how and leaves us with an inadequate account of the manifestation of propositional knowledge. Thus, practicalism is false.

As before, I shall now present an explicit reconstruction of my argument.

1. The explanation of S's acting intelligently with regard to the fact that p must involve appeal to S's intelligently acting in the light of their knowledge that p.
2. RA: Practicalism: Knowledge that p is, for certain activities φ, knowledge how to φ(p).

---

3. The explanation of S's acting intelligently with regard to the fact that p must involve appeal to S's intelligently acting in the light of their knowledge how to φ(p).

4. If S intelligently acts in the light of their knowledge how to φ(p), then S exercises their knowledge how to φ(p).

5. If S exercises their knowledge how to A, then S intelligently acts in the light of their knowledge that C(A), for at least some sufficient and at least some necessary conditions C on meeting the standards of A-ing.

---

6. The explanation of S's acting intelligently with regard to the fact that p must involve appeal to an infinite number of instances of S's exercising know-how – knowledge how to $\varphi_1(p)$, how to $\varphi_2(C1(\varphi_1(p)))$, and so on ad finitum – where $\varphi_{n+1}$ refers to those activities know-how of which is allegedly identical with the propositional knowledge that $C_n(\varphi_n(...(p)))$.

7. No explanation of a subjet's acts may involve appeal to their execution of an infinite number of other acts.

---

8. Practicalism is false.

Of course, the crux of this argument lies in premises 4 and 5.

Premise 4 may sound strange. Intelligently acting in the light of one's knowledge is perfectly intelligible when it concerns propositional knowledge. Then, it covers basing decisions upon the truth of the proposition known, asserting it, and so forth. However, what could it mean to intelligently 'act in the light of' *know-how*? But premise 4 is independent from this general problem. Practicalism maintains that all of the examples just mentioned are activities in 'p's epistemic diaspora'. Intelligently acting in the light of p is therefore understood as exercising the know-how to engage in those very activities. Premise 4 is an integral part of practicalism.

This shifts the burden of the argument to premise 5. Practicalists will probably reply that the intelligent exercise of knowledge how to A does not require any propositional knowledge about the sufficient and necessary conditions of meeting the standards of A-ing.

But how could this be true? Know-how is more than a mere disposition or a mere ability. It is a skill, an intelligent ability – an ability to achieve something in virtue of one's understanding of what it takes. Thus, rejecting premise 5 requires an account of this understanding which does not entail any propositional knowledge. But this is impossible.

To illustrate, consider the otherwise plausible idea that an understanding of some activity A consists in a meta-disposition to correct shortcomings in A-ing. However, if this meta-disposition is not accompanied by *any* propositional knowledge about A-ing, then it is only a blind regulatory mechanism rather than an *understanding* of A-ing.

Compare the following case: I have the ability to digest food and I certainly possess several mechanisms which correct shortcomings in my digestive system. Still, I do not know how to digest food. After all, these regulatory mechanisms are blind. They do not constitute my understanding of my digestive ability. And even those who *have* such an understanding do not digest in virtue of their understanding of digestion, but independently of it.

I conclude that the regress argument stands to scrutiny. Practicalism fails.

## 5. Equal Fundamentality

I have argued that both intellectualism and practicalism fall prey to vicious regresses. From this, we should draw two lessons.

First, the only option to escape from both regresses is to maintain the distinction between know-how and propositional knowledge.

How does this stop the anti-intellectualist regress? To intelligently perform an activity and thereby to manifest one's know-how does not require what intellectualism makes it require – the intelligent application of one's knowledge. Unlike propositional knowledge, know-how is a species of ability. And *qua* ability, it can be executed *directly*, without being applied.

What about the anti-practicalist regress? To intelligently act in the light of propositional knowledge does not require what practicalism makes it require – the reflective exercise of one's knowledge. Unlike know-how, propositional knowledge can inform a performance without being *activated in* the performance, but simply as part of its background reasons.

Thus, the distinction between know-how and propositional knowledge stops both regresses.

The second lesson I would like to draw from my findings is that know-how and propositional knowledge are distinct, but still interdependent.

Ryle famously held that know-how is conceptually prior to propositional knowledge since one cannot know that p without knowing how to find out whether p and without knowing how to use the concepts which are part of the content that p (cf. Ryle 1945: 15-16). I agree. But we should also appreciate a complementary insight: One cannot know how to do something without having at least a minimum amount of propositional knowledge about the sufficient and necessary conditions of meeting the standards of doing so. In this sense, propositional knowledge is conceptually prior to know-how.

Thus, both kinds of knowledge presuppose each other. To possess knowledge *at all* always means to possess *two kinds* of knowledge states.

This view retains parts of the respective motivations for intellectualism and practicalism: One cannot understand one of these concepts without understanding the other one, too. But the dependence runs in *both* directions. They are equally fundamental.

One might object that this proposal *also* leads to a regress problem. For any piece of knowledge still triggers an infinite chain of *other* pieces of knowledge. Knowledge how to A requires some propositional knowledge about the standards of A-ing, which in turn requires knowledge how to employ certain concepts, and so on. If such an infinite chain of knowledge is a problem for intellectualism and practicalism, how can it be all right *now*?

I should start by replying that I happily bite the bullet on offer. If we count pieces of knowledge, the number of pieces we will find is infinite. This is not an uncommon view – holism. It is pointless to try to capture the holistic web of knowledge in terms of a list.

However, the problem with intellectualism and practicalism is *not* that they imply this view – holism. What these theories were shown to imply is the much more problematic view that intelligently performing activities and acting intelligently with regard to facts both require the execution of an infinite number of *further acts*. *This* is why we must reject them.

The holistic interdependence of two distinct kinds of knowledge states does *not* entail that a manifestation of knowledge requires the execution of infinite numbers of further acts:

True, exercising know-how also requires acting in the light of propositional knowledge about the standards of the activity in question. But this is where the regress stops. Exercising *one* piece of know-how does not require exercising any *further* piece of know-how. *Having* propositional knowledge always requires *having* know-how. But acting in the light of the former does not require exercising the latter. When I act in the light of my knowledge that p, I do not exercise my knowledge how to find out whether p.

Thus, both vicious regresses are blocked.

## 6. Conclusion

I conclude that the distinction between know-how and propositional knowledge is crucial in understanding what Ryle calls 'intelligent practice', but that the interconnections of these concepts are an important topic which should be explored further.[1]

**David Löwenstein**

Freie Universität Berlin
david.loewenstein@fu-berlin.de

## References

Fridland, E. 2012: 'Problems with intellectualism', *Philosophical Studies*, Online First.

Ginet, C. 1975: *Knowledge, Perception, and Memory*. Dodrecht: Reidel.

Hetherington, S. 2006: 'How to Know (that Knowledge-that is Knowledge-how)', in *Epistemology Futures*, Oxford: Oxford University Press, 71–94.

— 2008: 'Knowing-That, Knowing-How, and Knowing Philosophically', *Grazer Philosophische Studien* 77, 307–24.

— 2011: *How to Know. A Practicalist Conception of Knowledge*. Chichester: Wiley-Blackwell.

Löwenstein, D. 2011: 'Knowledge-How, Linguistic Intellectualism, and Ryle's Return', in S. Tolksdorf (ed.): *Conceptions of Knowledge*, Berlin and New York: Walter de Gruyter. 269–304.

Ryle, G. 1945: 'Knowing How and Knowing That', *Proceedings of the Aristotelian Society* 46, 1–16.

— 1949: *The Concept of Mind*. Edition of 1963. Harmondsworth: Penguin Books.

Stanley, J. 2011a: *Know How*. Oxford: Oxford University Press.

— 2011b: 'Knowing (How)', *Noûs* 45, 207–38.

Stanley, J. and T. Williamson 2001: 'Knowing How', *The Journal of Philosophy* 98, 411–44

[1] I would like to thank my audience at GAP.8 for a lively and helpful discussion. Lars Dänzer, Ellen Fridland, Romy Jaster, Nadja El Kassar, David Lauer, David Ludwig, and Holm Tetens have earned even deeper gratefulness for their insightful critical and constructive comments on various earlier versions of these ideas.

# Interrogative Formen des Wissens und reduktiver Intellektualismus

Pedro Schmechtig

Für gewöhnlich liegt das Hauptaugenmerk in der Erkenntnistheorie auf Wissens-Zuschreibungen mit deklarativen Satzkomplementen (Wissen-dass). Im Alltag sind stattdessen interrogative Formen des Wissens (Wissen-*wh*), die nicht mit derartigen Satzkomplementen zum Ausdruck gebracht werden, sehr viel verbreiteter. Intellektualisten wie Jason Stanley (2011) zeigen sich von dieser Tatsache relativ unbeeindruckt, da ihrer Ansicht nach alle Formen des Wissens eine einheitliche propositionale Basisform besitzen und folglich auf Wissen-dass zurückführbar sind. Die vorliegende Arbeit stellt diese Sichtweise in Frage. Ich beginne mit einigen Vorbemerkungen zur Rolle von Fragen und Antworten im Rahmen der Standardsemantik (Abschnitt 2). Anschließend wende ich mich der zentralen Frage zu, was hinter der reduktiven Analyse von responsiven Interrogativen steckt (Abschnitt 3) bzw. warum die propositionale Standardauffassung nicht funktioniert (Abschnitt 4). Im letzten Teil wird dann eine alternative Sichtweise skizziert, die im Grundsatz besagt, dass sich die genannten Probleme vermeiden lassen, wenn man eine nicht-reduktive Analyse der Wissens-Relation um eine normative Erklärung des „Stehens in dieser Relation“ ergänzt (Abschnitt 5).

## 1. Fragestellung

Es gehört zur gängigen Auffassung in der Erkenntnistheorie, dass Wissens-Zuschreibungen, die ein deklaratives Satzkomplement beinhalten (Wissen-dass), im Mittelpunkt des allgemeinen Forschungsinteresses stehen. Im täglichen Leben sind wir jedoch sehr viel häufiger mit praktischen (Wissen-wie) bzw. interrogativen Formen des Wissens (Wissen-*wh*) konfrontiert. Zuschreibungen dieser Art beruhen nicht auf einer Verwendung von deklarativen Satzkomplementen. Intellektualisten wie Jason Stanley (2011) versuchen dieser Tatsache Rechnung zu tragen, indem sie eine in zweifacher Hinsicht reduktive Sichtweise propagieren:

(i) Wissen-*wie* ist eine *spezielle Unterart* von Wissen-*wh*

(ii) Wissen-*wh* kann aufgrund der *einheitlichen propositionalen Basisform* auf Wissen-*dass* zurückgeführt werden.

Das Problem mit dieser Sichtweise ist Folgendes: Wenn sich Annahme (ii) als falsch erweisen sollte, kann es uns eigentlich egal sein, ob (i) zutreffend ist oder nicht, der Intellektualismus würde schon daran scheitern, dass interrogative Wissens-Zuschreibungen *nicht in der angenommenen Weise reduktiv* sind. Die entscheidende Frage ist daher, ob eine reduktive Auffassung der Bedeutung interrogativer Wissens-Zuschreibungen – auf die sich Intellektualisten wie Stanley in ihrer Analyse von Wissen-*wie (Knowledge-how)* stützen – wirklich plausibel ist? Meine Antwort lautet „Nein“! Weder ist praktisches Wissen eine bloße Spezies von Wissen-*wh*, noch lassen sich interrogative Formen des Wissens auf Wissen-*dass* reduzieren. In der vorliegenden Arbeit geht es mir ausschließlich um den zweiten Teil der Behauptung. Ich beginne mit einigen Vorbemerkungen zur Rolle von Fragen und Antworten im Rahmen der Standardsemantik (Abschnitt 2). Anschließend wende ich mich der zentralen Frage zu, was hinter der reduktiven Analyse von responsiven Prädikaten steckt (Abschnitt 3) bzw. warum die Standardauffassung aus meiner Sicht nicht funktioniert (Abschnitt 4). Im

letzten Teil wird dann eine alternative Sichtweise skizziert, die im Grundsatz besagt, dass sich zentrale Probleme des reduktionistischen Ansatzes vermeiden lassen, wenn man eine nicht-reduktionistische (semantische) Analyse der Wissens-Relation um den Aspekt einer normativen Erklärung des „Stehens in dieser Wissens-Relation" ergänzt (Abschnitt 5).

## 2. Fragen und Antworten – einige Vorbemerkungen

Um beurteilen zu können, ob die reduktive Standardanalyse von Wissen-*wh* zutreffend ist oder nicht, sind zunächst einmal einige Vorbemerkungen erforderlich. Syntaktisch gesehen sind Fragen Instanzen eines bestimmten linguistischen Typs, d.h. sie besitzen eine sog. interrogative Form. Traditionell werden Interrogative in (1) *Stammformen (direkte Fragen)* und (2) *eingebettete Formen (indirekte Fragen)* unterteilt:[1]

(1) Wo befindet sich Peter? (*direkt)*

(2) Ich weiß, [wo sich Peter befindet]. (*indirekt)*

Die vorliegende Arbeit beschäftigt sich vorrangig mit einer bestimmten Klasse von (2), nämlich mit sog. responsiven eingebetteten Fragen. Was zeichnet responsive Interrogative aus?

Nach Lahiri (2002) müssen zwei Formen der Einbettung von Fragen unterschieden werden. Einerseits gibt es Einstellungsverben, die nur mit einem deklarativen Satzkomplement (kurz SK) verwendet werden können (wie z.B. *überzeugen*). Auf der anderen Seite finden wir Einstellungsverben, die nur mit einem interrogativen SK vereinbar zu sein scheinen (wie z.B. *fragen*, *untersuchen etc.*). Derartige *reine* Interrogative grenzen sich wiederum von *responsiven* Interrogativen ab. Responsive Interrogative – zu denen insbesondere epistemische Einstellungsverben gehören – zeichnen sich dadurch aus, dass sie mit *beiden Arten* von SKen verträglich sind. Diese Unterscheidung lässt sich folgendermaßen präzisieren:

**Rein rogative ‚question embedders' (fragen, untersuchen etc.)**

(3) Hans hat gefragt, wo seine Frau ist.

(4) ?? Hans hat gefragt, dass seine Frau in Berlin steckt.

Bei reinen Rogativen ist es unmöglich, eine Antwort auf die eingebettete Frage von (3) in Form einer propositionalen Einstellung wie (4) zum Ausdruck zu bringen. Derartige „question embedders" können nicht mit einem deklarativen SK verwendet werden.

Anders verhält es sich bei responsiven Interrogativen. Diese können in Form der Zuschreibung einer propositionalen Einstellung mit einem deklarativen SK verwendet werden. So kann beispielsweise mit (6) eine kongruente Antwort auf die indirekte Frage von (5) zum Ausdruck gebracht werden:

**Responsive ‚question embedders' (wissen, erinnern, entdecken etc.)**

(5) Hans weiß, wo seine Frau ist.

(6) Hans weiß, dass seine Frau in Konstanz ist.

Bei responsiven Prädikaten scheint der interrogative Gebrauch mit der propositionalen Verwendung eng verbunden zu sein: Die Wahrheit von (6) impliziert die Wahrheit von (5).

[1] Vgl. Higginbotham (1996, 361).

Doch um welche *Art von Antworten* handelt es sich bei responsiven Interrogativen? In der Regel wird die Bedeutung responsiver Interrogative in Relation zur Bedeutung *kongruenter* Antworten bestimmt:[2]

Frage: (F) Wo studiert Barbara?

Antwort: ($A_1$) Sie studiert in Berlin.

($A_2$) Sag Du es mir.

Im Vergleich zu ($A_1$) stellt ($A_2$) keine kongruente Antwort auf (F) dar. Obgleich natürlich ($A_2$) eine völlig legitime kommunikative Reaktion ist.

Wie genau kongruente Antworten zu bestimmen sind, ist in der Semantik für responsive Interrogative umstritten. Im Wesentlichen werden drei unterschiedliche Kandidaten diskutiert (vgl. Beck und Rullmann 1999):

– strenge vollständige Antworten *(strongly exhaustive)*

– weite vollständige Antworten *(weakly exhaustive)*

– das Anführen irgendeiner Antwort *(mention-some)*

Betrachten wir zur Erläuterung dieser Varianten ein kurzes Beispiel: Peter hält einen Vortrag vor genau drei Personen, nämlich A, B und C. Unsere Frage lautet anschließend: Wer hat Peters Vortrag gehört? Folgende Antwortmöglichkeiten stehen offen:

**Strenge vollständige Antwort**

A, B und C waren die einzigen Personen, die den Vortrag gehört haben. *Jedes* Element der Antwortmenge wird angeführt.

**Weite vollständige Antwort**

A, B und C haben den Vortrag gehört. Jedes Element der aktualen Antwortmenge wird angeführt, aber es könnte noch mehr Elemente geben.

**Anführen irgendeiner Antwort**

A hat den Vortrag gehört. *Irgendein* Element der Antwortmenge wird angeführt.

Wenn man das „Anführen irgendeiner Antwort“ als echte Möglichkeit zulässt, dann liegt es nahe, den Begriff der weiten vollständigen Antwort als eine Art Kombination aller „mention-some answers“ aufzufassen.

Intellektualisten wie Stanley haben ihre reduktionistische Sichtweise auf eine semantische Konzeption gestützt, die vor allem von Groenendijk & Stockhof (1982, 1984) verteidigt wurde.[3] Diese besagt, dass die Bedeutung responsiver Interrogative ausschließlich auf der Basis *strenger Vollständigkeit* zu bestimmen sei. Warum diese Festlegung auf vollständige Antwortmengen? Was spricht *für* eine solche Akzentuierung? Zwei Überlegungen sind an dieser Stelle zu nennen. Es scheint Fälle wie (7) zu geben, in denen das Anführen irgendeiner Antwort nicht hinreichend ist:

(7) Peter weiß, welche Tagungsteilnehmer in seinem Vortrag waren.

(8) Peter weiß, dass A ein Tagungsteilnehmer ist, der in seinem Vortrag war.

Aus der Wahrheit von (8) folgt *nicht* die Wahrheit von (7). Dies legt den Gedanken nahe, in Fällen wie diesen strenge Vollständigkeit zu fordern.

Die zweite Überlegung beruht auf der Beobachtung, dass insbesondere bei *epistemischen* Prädikaten ein Problem mit dem Begriff der *weiten Vollständigkeit* auftritt. Weite

[2] Vgl. hierzu Krifka (2001).

[3] Vgl. Stanley (2011, 60).

(vollständige) Antwortmengen scheinen keine Informationen über negative Abgrenzungen zu beinhalten. Nehmen wir an, die Zuschreibung von (7) ist korrekt und wir informieren Peter darüber, dass Heinz ein Tagungsteilnehmer war. In dieser Situation – so die Überlegung – sollte Peter wissen, sofern (7) wahr ist, ob Heinz in seinem Vortrag war. Doch diese Intuition lässt sich mit dem Begriff der weiten Vollständigkeit nicht erklären. Eine kongruente Antwort im Sinne der weiten Vollständigkeit verlangt nämlich nicht, dass Peters Wissen negative Abgrenzungen beinhaltet, d.h. die Möglichkeit ausschließt, dass Hans nicht in seinem Vortrag war.[4]

Ob die eben genannten Überlegungen hinreichend sind, die Behauptung zu stützen, dass strenge Vollständigkeit für *alle* Arten der Verwendung von responsiven Prädikaten grundlegend ist, darf jedoch – wie später noch deutlich wird – bezweifelt werden. Ich komme jetzt zur eigentlichen Frage dieser Arbeit, ob sich die eingangs angeführte Annahme (ii) im Rahmen einer reduktionistischen Sichtweise von responsiven Interrogativen rechtfertigen lässt.

## 3. Die reduktionistische Standardanalyse von Wissen-‚wh‘

Nach der reduktionistischen Standardanalyse ist das Verb „wissen“ ein prototypisches responsives Interrogativ. Die grundlegende (reduktionistische) Behauptung besagt: Bei Prädikaten dieser Art lässt sich der *interrogative Gebrauch auf die propositionale Verwendung* zurückführen. Dabei stehen zwei Hauptgedanken im Hintergrund: (a) Man versteht nur dann die Bedeutung einer Frage, wenn man weiß, wie eine kongruente Antwort auf diese Frage aussieht. (b) Wir antworten auf Fragen durch Äußerungen, die *Propositionen zum Ausdruck* bringen. Gemäß dieser zwei Behauptungen geht die Standardauffassung davon aus, dass der semantische Beitrag einer ‚wh‘-Frage die Menge der möglichen (vollständigen) Antwortpropositionen ist, die für eine solche Frage besteht. Fragen werden dann entweder direkt als Mengen von möglichen (alternativen) Propositionen definiert, die bezogen auf die aktuale Welt eine wahre (kongruente) Antwort zum Ausdruck bringen.[5] Oder sie werden abstrakter im Sinne einer Aufteilung des logischen Raums (*partitions*) als Funktionen von möglichen Welten in Propositionen bestimmt.[6] Darüber hinaus wird von vielen Reduktionisten angenommen, dass Antwortmengen durch existentielle Quantifizierung über die Werte jener Propositionen gebildet werden, welche die ‚wh‘-Phrase erfüllen.[7]

Im Rahmen dieser allgemeinen Überlegungen lassen sich Zuschreibungen der Form *[S-W-wh]* – wobei S für ein angemessenes Nomen, W für das zweistellige Einstellungsverb „Wissen“ [WSp] und ‚wh‘ für das interrogative SK (*why*, *wh*ere, *wh*en, *wh*ether usw. plus eine infinite oder finite Verbphrase) steht – auf Zuschreibungen der Form *[S-W-that]* in folgender Form reduzieren:

[4] In dieser Form hat insbesondere George (2011) für die Beibehaltung des Begriffs der strengen Vollständigkeit argumentiert.

[5] Vgl. hierzu Hamblin (1958, 1973).

[6] Vgl. Groenendijk & Stockhof (1984).

[7] Gemäß diesem Schema beinhaltet das Definiens eine existenzielle Quantifizierung von Propositionen. Eine solche Bedingung ist jedoch keineswegs unumstritten. Es gibt semantische Konzeptionen (wie die von Groenendijk & Stockhof 1984 oder Heim 1994), die darauf verzichten.

**Das reduktionistische Schema der Analyse von Wissen-‚wh'** [8]

„S weiß-*wh*" ist wahr gdw. es gibt die/eine Proposition p, sodass p eine Antwort auf die im *wh*-Satzkomplement eingebettete Frage ist & S weiß, dass p.[9]

Wie sich gezeigt hat, steht ein derartiger Ansatz vor zahlreichen Problemen. Im Folgenden soll es jedoch nur um solche Einwände gehen, die in einem direkten Zusammenhang damit stehen, dass „wissen" ein prototypisches (responsives) Interrogativ ist.

## 4. Weshalb der reduktionistische Ansatz nicht funktioniert

Die von mir diskutierten Einwände bewegen sich auf drei unterschiedlichen Ebenen: Unter Punkt (4.1) wird zunächst ein Einwand formuliert, der vorrangig gegen solche Ansätze gerichtet ist, die sich (wie im Fall von Stanley 2011) einer aus meiner Sicht unkritischen Verwendung der Semantik von Groenendijk & Stockhof bedienen. Danach wird unter Punkt (4.2) ein allgemeines Problem aufgeworfen – nämlich die Unvereinbarkeit mit „mention-some questions" –, das sich für alle reduktionistischen Programme (der semantischen Erklärung responsiver Prädikate) gleichermaßen stellt. Abschließend werde ich unter Punkt (4.3) speziell in erkenntnistheoretischer Hinsicht dafür argumentieren, dass reduktionistische Ansätze prinzipiell ungeeignet sind, ein angemessenes „setting" für Wissens-Zuschreibungen zu liefern.

### *4.1 Strenge versus weite Vollständigkeit*

Vielfach wurde betont, dass es Fälle der *Negation* von Wissens-Zuschreibungen gibt, die mit dem Begriff der *strengen Vollständigkeit* – so wie er von reduktiven Intellektualisten im Anschluss an die semantische Konzeption von Groenendijk & Stockhof unterstellt wird – nicht vereinbar sind. Hier ein solches Beispiel:[10]

(9) Peter weiß, wer die Prüfung bestanden hat, aber er weiß nicht, wer die Prüfung nicht bestanden hat.

Unserer Intuition nach sind Wissens-Zuschreibungen in der Form von (9) *nicht* widersprüchlich. G&S gehen jedoch in ihrer Semantik davon aus, dass die nachstehenden Sätze (10) und (11) als *äquivalent* zu betrachten sind, d.h. jeder Satz eines solchen Paares den anderen impliziert (*entails*):[11]

(10) Peter weiß, wer bestanden hat.

(11) Peter weiß, wer nicht bestanden hat.

Doch wenn das der Fall ist, dann sollte eine Zuschreibung wie (9) generell widersprüchlich sein. Mit anderen Worten: Wenn man unterstellt, dass (10) und (11) äquivalent sind, ist eine

---

[8] Das vorliegende Schema wird entweder allgemein (Higginbotham 1996) oder im Hinblick auf die Analyse spezifischer Interrogative vertreten: Bei Hintikka (1975) bzw. Böer & Lycan (1986) in Bezug auf *Knowledge-who*, bei Lewis (1982) in Bezug auf *Knowledge-whether* und bei reduktiven Intellektualisten wie Williamson & Stanley (2001) bzw. Stanley (2011) in Bezug auf *Knowledge-how*. Berit Broogard (2009, 440) hat eine leicht geänderte Formulierung vorgeschlagen, da sie die reduktionistische Position als *meta-linguistische* Aussage bezüglich der Wahrheitsbedingungen von Äußerungen der Art „S weiß-*wh*" rekonstruiert, d.h. nicht als eine Aussage über die Form von Propositionen, die mit derartigen Äußerungen zum Ausdruck gebracht werden.

[9] Wie bereits erwähnt, berufen sich reduktive Intellektualisten im Rahmen der Standardanalyse auf den Begriff der strengen Vollständigkeit, d.h. sie gehen davon aus, dass eine Zuschreibung der Form „S weiß-*wh*" nur dann in semantischer Hinsicht adäquat ist, wenn p im Sinne einer (strengen) vollständigen Antwortmenge von S gewusst wird.

[10] Vgl. zu dieser Art von Beispielen Sharvit (2002).

[11] Groenendijk & Stockhof (1984, 87).

strenge Lesart der Vollständigkeitsbedingung (für kongruente Antworten) nicht mit unserer Intuition bezüglich der Zuschreibung von (9) vereinbar.

Zur Rechtfertigung der angeführten Äquivalenzbehauptung lässt sich aus Sicht von G&S Folgendes sagen: Wenn man annimmt, dass Peter nur aufgrund *vollständiger Informiertheit* Wissen hat, d.h. Peter nur dann im Fall von (9) Wissen zugeschrieben werden sollte, wenn er eine vollständige Liste kennt, die enthält, wer bestanden und wer alles teilgenommen hat, dann scheint (9) unter Annahme der Bivalenz von Wahrheitswerten tatsächlich widersprüchlich zu sein – nämlich sowohl wahr als auch falsch. G&S gestehen jedoch zu, dass (9) eine andere Lesart bekommen würde, wenn man *nicht* von vollständiger Informiertheit ausgehen könnte. Demnach ist (9) nur dann widersprüchlich, wenn man Peters Wissen – was die Vollständigkeit der Antwortmenge anbetrifft – durch den Bereich der ‚wh'-Phrase als hinreichend fixiert ansehen kann. In gewöhnlichen (konversationalen) Äußerungskontexten ist diese Bedingung ihrer Meinung stets erfüllt.

Ich halte diese Behauptung für wenig überzeugend. Wenn vollständige Informiertheit unter normalen (konversationalen) Äußerungsumständen immer vorausgesetzt werden muss, hätten wir wahrscheinlich nur sehr selten Gelegenheit, jemandem Wissen zuzuschreiben. Mir scheint, hier wird das Pferd von hinten aufgezäumt. In vielen Situationen ist eine Zuschreibung von Wissen-*wh* angemessen, obgleich es keinerlei Evidenz gibt, aufseiten des Wissenden vollständige Informiertheit zu unterstellen. Auf Situationen, die diese Sichtweise nahelegen, komme ich gleich noch genauer zu sprechen.

So weit ich sehen kann, gibt es zwei Möglichkeiten, auf genannte Schwierigkeit zu reagieren: Einige Autoren haben vorgeschlagen, ein alternatives (semantisches) Programm zu entwickeln, wonach strenge Vollständigkeit zugunsten des Begriffs der *weiten Vollständigkeit* aufgegeben wird (Heim 1994, Spektor 2007). Andere sind hingegen der Ansicht, dass ein Festhalten an strenger Vollständigkeit – aufgrund des angesprochenen epistemischen Problems mit dem Begriff der weiten Vollständigkeit – unabdingbar ist, dafür aber die *Äquivalenzbehauptung* zurückgewiesen werden muss (George 2011). Das Problem für reduktive Intellektualisten ist jedoch, dass *beide* Alternativen unter den von ihnen gemachten Annahmen nicht gangbar sind: Einerseits wollen sie (aus gutem Grund) am Begriff der strengen Vollständigkeit festhalten. Andererseits ist es aber so, dass eine Zurückweisung der Äquivalenzbehauptung unmöglich erscheint, denn das würde bedeuten, Intellektualisten müssten ihr eigentliches Ziel – nämlich die reduktionistische Sichtweise – aufgeben.[12]

### *4.2 Das Problem der „mention-some"-Lesart von responsiven Interrogativen*

Eines der bedeutendsten Probleme der Standardtheorie, so wie sie von G&S verteidigt wird, beruht auf einer bestimmten Form der Ambiguität von *wh*-Fragen. In Fällen wie (12), (13) oder (14) bedarf es allem Anschein nach keiner vollständigen Antwortmenge, sondern lediglich des *Anführens irgendeiner Antwort*:

(12) Wer hat das Licht angemacht?

(13) Wo kann man eine Tageszeitung kaufen?

(14) Wie kommt man zum Bahnhof?

Darüber hinaus gibt es Fragen wie (15), die Ausdrücke enthalten, welche *explizit* zu erkennen geben, dass strenge Vollständigkeit von vornherein ausgeschlossen ist:[13]

(15) Wer *zum Beispiel* hat den Vortrag von Peter gehört?

[12] Eine ausführliche Begründung dieser Behauptung findet sich bei George (2011).

[13] Vgl. Beck & Rullmann (1999).

Dass eine Frage wie (15) eine kongruente Antwort besitzt, obgleich es dafür nicht der Kenntnis irgendeiner (vollständigen) Liste von Teilnehmern bedarf, ist sicherlich unstrittig.

Die Möglichkeit einer solchen ‚mention-some'-Lesart von Interrogativen stellt den reduktionistischen Ansatz vor zwei grundlegende Schwierigkeiten: *Erstens* scheint die Erklärung dieser Lesart – im Rahmen der semantischen Konzeption, auf die sich reduktive Intellektualisten wie Stanley beziehen – wenig überzeugend zu sein. *Zweitens* lässt sich im Zusammenhang mit einem responsiven Prädikat wie „wissen" ein Gettier-artiger Einwand formulieren, der das reduktionistische Programm unter erkenntnistheoretischer Perspektive in große Verlegenheit bringt. Ich werde beide Schwierigkeiten der Reihe nach kurz erläutern.

Nach G&S stellt die Tatsache, dass responsive Interrogative eine „mention-some"-Lesart besitzen können, kein ernsthaftes Problem dar, da es eine einfache Erklärung für dieses Phänomen gibt. Obwohl das Anführen aller relevanten Elemente einer Antwortmenge für eine vollständige Antwort notwendig ist, reicht es oft in vielen Situationen aus, wenn man nur eine *Teilantwort* (*partial answer*) gibt.

Diese Erklärung ist jedoch nicht überzeugend.[14] Der Grund ist Folgender: Im Gegensatz zu einer bloßen Teilantwort ist das „Anführen irgendeiner Antwort" – sofern sich diese als kongruent erweist – damit verbunden, dass der *rogative Impuls zum Stillstand* kommt. Wenn das „Anführen irgendeiner Antwort" erfolgreich ist, löst sich die betreffende Frage in aller Regel auf. Demgegenüber scheint das Auflösen der betreffenden Frage keine Erfolgsbedingung einer (kongruenten) Teilantwort zu sein. In Bezug auf paradigmatische Fälle wie (12) gibt es Teilantworten wie (TA), die, obgleich sie sich als kongruent erweisen, nicht zu einer vollständigen Auflösung der Frage führen:

(12) Wer hat das Licht angemacht?

(TA) Babara war es nicht.

Eine Antwort wie (TA) würde auch dann nicht den rogativen Impuls zum Stillstand bringen, wenn sie im Sinne strenger Vollständigkeit (z.B. durch eine Liste aller Personen, die nicht das Licht angemacht haben) komplettiert wird. Selbst wenn man in einer Situation wie (12) weiß, wer alles nicht das Licht angemacht hat, stellt sich weiterhin die Frage, wer es war, der den Lichtschalter betätigt hat. Entsprechend kann nicht davon ausgegangen werden, dass das Anführen einer Teilantwort *denselben responsiven Status* besitzt, den das Anführen irgendeiner (kongruenten) Antwort im Rahmen einer „mention-some"-Lesart besitzt.

Die zweite Schwierigkeit besagt, dass sich Gettier-artige Szenarien denken lassen, die den reduktionistischen Ansatz in Frage stellen. Angenommen es gibt Situationen, in denen zwei Personen A und B in Relation zu ein und derselben (wahren) Proposition (als Antwort auf die entsprechende ‚wh'-Frage) stehen, aber die Zuschreibung von Wissen-*wh* im Fall von A korrekt und im Fall von B nicht korrekt ist. Dann zeigen solche Fälle, dass die Zuschreibung von Wissen-*wh* weiterer Faktoren bedarf, weshalb eine reduktionistische Sichtweise – zumindest in der Gestalt, in der sie von führenden Intellektualisten in Anspruch genommen wird – nicht plausibel erscheint. Betrachten wir dazu das folgende Szenario:

> Hans und Peter wissen, wer die ‚Tour de France 2012' gewonnen hat. Beide stehen in einer Wissens-Relation zu ein und derselben Antwortproposition (=die Proposition, dass Bradley Wiggins gewonnen hat). Darüber hinaus glaubt Peter – im Gegensatz zu Hans, der dazu keinerlei Meinung hat –, dass der Gewinner der ‚Tour de France 2012' ein Eskimo ist, der im ‚Team Telekom' fährt. Peters Überzeugung, dass der Mann, der die ‚Tour de France 2012' gewonnen hat, ein Eskimo ist, der im ‚Team Telekom' fährt, ist aber offenkundig falsch. Weder Eskimos noch ein ‚Team Telekom' haben an der ‚Tour de France 2012' teilgenommen.

[14] Van Rooij (2004) kommt an dieser Stelle zu einer ganz ähnlichen Schlussfolgerung.

Im vorliegenden Szenario gibt es zwei Zuschreibungen von Wissen-*wh*, deren Basis ein und dieselbe (wahre) Antwortproposition (=Bradley Wiggins) ist. Dennoch haben wir die Intuition, dass (16) in *epistemischer Hinsicht* korrekt ist, während es (17) nicht ist:

(16) Hans weiß, wer die ‚Tour de France 2012' gewonnen hat. (*korrekt)*

(17) Peter weiß, wer die ‚Tour de France 2012' gewonnen hat. (*nicht korrekt)*

Wie lässt sich diese Intuition erklären? Offenbar damit, dass Peter zwar eine wahre Antwortproposition (=Bradley Wiggins) kennt, die Bekanntschaft mit dieser Proposition aber auf einer *Art von Zufälligkeit* basiert, die wir im Fall von Wissen ausschließen wollen. Denn Peter hat zusätzlich zur gewussten (wahren) Antwortproposition die problematische *Überzeugung*, dass es eine äquivalente Antwort gibt, die er hätte genauso anführen können, die aber offenkundig falsch ist. Anders gesagt, Zuschreibungen von Wissen-*wh* sind nur dann korrekt, wenn S eine kongruente (wahre) Antwort kennt *und S nicht die Überzeugung hat*, dass es eine äquivalente Antwortproposition gibt (=der Eskimo vom ‚Team Telekom'), die falsch ist. Doch wenn das so ist, dann lässt sich die Korrektheit von (16) *nicht allein auf propositionales Wissen* zurückführen. Zusätzlich zur gewussten (wahren) Antwortproposition muss nämlich noch eine andere Bedingung erfüllt sein: Das Wissen um eine wahre Antwort darf – bezüglich äquivalenter Antwortpropositionen – *nicht auf falschen Überzeugungen* beruhen, die in einem *relevanten* Zusammenhang zur Zuschreibung von Wissen-*wh* stehen.[15]

### *4.3 Der reduktionistische Ansatz liefert kein geeignetes „setting" für Wissens-Zuschreibungen*

Bislang wurden Schwierigkeiten formuliert, welche die verschiedenen Lesarten von (kongruenten) Antworten betreffen. Es wurde jedoch *nicht* in Frage gestellt, dass die Antwortmengen für responsive Einstellungsprädikate *propositionaler* Natur sind. Demgegenüber besagt ein weiterer (vielleicht noch radikalerer) Einwand, dass die reduktionistische Standardauffassung scheitert, da ein angemessenes „setting" für *epistemische* Interrogative nicht nur in semantischer, sondern auch in *ontologischer* Hinsicht anders zu bestimmen sei.

Insbesondere Jonathan Ginzburg (1995, 2011) hat dafür argumentiert, dass epistemische Interrogative keine strikte „proposition-denoting manifestation" besitzen. Was ist damit gemeint? Nach Ginzburg ist bei Einstellungsprädikaten eine gewisse Variabilität in Bezug auf die Frage zu beobachten, welche Entitäten vom jeweiligen SK denotiert werden. Einerseits scheint es Prädikate zu geben, die ganz im Sinne der Standardanalyse nur mit einem Typ von Entität (nämlich Propositionen) auskommen. Andererseits gibt es Klassen von Prädikaten – zu denen allen voran epistemische Einstellungsverben gehören –, bei denen das SK *mehr als nur einen Typ* von Entität prädizieren kann.

Im Hintergrund dieser Überlegung steht die Annahme, dass sich epistemische Einstellungsprädikate nicht direkt im Sinne einer *nominalen* Verwendung des SKs auf Fragen beziehen. Generell ist festzuhalten, dass Einstellungsprädikate nicht nur mit deklarativen oder interrogativen, sondern auch mit nominalen SKen verwendet werden können. Ein

[15] Falsche Überzeugungen stehen genau dann in einem *relevanten* Zusammenhang zur Zuschreibung von Wissen-*wh*, wenn sie (i) eine äquivalente Antwortproposition generieren und (ii) gleichzeitig einen guten Grund liefern, weshalb die Zuschreibung annulliert (defeated) werden sollte. Damit ist ganz bewusst keine substantielle Festlegung getroffen, was es in semantischer Hinsicht heißt, eine falsche Überzeugung als „relevant" zu bezeichnen. Ich werde im letzten Teil mehr dazu sagen. Vorab nur so viel: Die Relevanz einer falschen Überzeugung ist nicht in semantischer, sondern in *normativer* Hinsicht zu bestimmen. Sie hängt von der Art der Wertschätzung der epistemischen Position ab (ob es zutreffend ist, dass von S „in einer Relation des Wissens-*wh*" steht), wobei die Angemessenheit einer solchen Wertschätzung mit Ziel der erkenntnistheoretischen Untersuchung variieren kann.

Einstellungsprädikat kann aber nur dann *direkt* im Sinne eines nominalen SKs mit einer Frage verbunden sein, wenn es bezüglich dieses SKs allgemein bekannte *Tests der direkten Bezugnahme* besteht. Ginzburg führt dazu zwei der bekanntesten Tests an:

**Test I: Substituierbarkeit**

Klaus untersucht/diskutiert eine wichtige Frage.

Diese Frage lautet: Wer war gestern auf der Party?

Klaus untersucht/diskutiert, wer gestern auf der Party war.

**Test II: Existenzielle Verallgemeinerung**

Klaus untersucht/diskutiert, wer gestern auf der Party war.

*Daraus lässt sich folgern*: Es gibt eine Frage, die Klaus untersucht/diskutiert.

Diese Frage lautet: Wer war gestern auf der Party?

Demgegenüber scheinen epistemische Einstellungsprädikate nicht die Eigenschaft zu besitzen, diese Tests bestehen zu können:

**Test I: Substituierbarkeit**

Klaus weiß/erkennt eine wichtige Frage.

Diese Frage lautet: Wer war gestern auf der Party?

*Daraus folgt nicht*: Klaus weiß/erkennt, wer gestern auf der Party war.

**Test II: Existenzielle Verallgemeinerung**

Hans weiß, wer gestern im Spielkasino war.

*Daraus folgt nicht:* Es gibt eine Frage, die Hans weiß.

Sofern es nun aber zutreffend ist, dass epistemische Interrogative zu einer Klasse von Einstellungsprädikaten gehören, die nicht im Sinne eines nominalisierten SKs direkt auf Fragen Bezug nehmen, lässt sich anschließend zeigen, warum derartige Prädikate *keine* echte „proposition-denoting manifestation“ besitzen. Hier das betreffende Argument:

($P_1$) Es gibt Prädikate wie „überzeugen“, die mit einer propositionalen Einstellungsanalyse vereinbar sind, da sie den Test der direkten Referenz im Zusammenhang mit nominalisierten Frage-SKen (die Propositionen denotieren) bestehen.

($P_2$) Epistemische Einstellungsprädikate bestehen diese Tests nicht, d.h. sie können nicht mit nominalisierten Frage-SKen (die Propositionen denotieren) verwendet werden.

($P_3$) Vertreter der propositionalen Analyse gestehen zu, dass Einstellungsprädikate, die wie „überzeugen“ mit nominalisierten SKen vereinbar sind, *nicht mit interrogativen* SKen (insbesondere nicht mit sog. *whether*-SKen) verwendet werden können.

(K) Aufgrund von ($P_1$) - ($P_3$) ist zu bezweifeln, dass epistemische Einstellungsprädikate, die mit interrogativen SKen kombinierbar sind, eine echte „proposition-denoting manifestation“ im Sinne der propositionalen Standardanalyse besitzen.

Wenn dieses Argument durchgeht, muss die zentrale Annahme reduktiver Intellektualisten – wonach Wissen-*wh* eine einheitliche propositionale Basisform besitzt – schon deshalb

zurückgewiesen werden, weil eine solche *propositionale* Analyse nicht imstande ist, ein *geeignetes „setting" für epistemische Interrogative* zu liefern.[16]

Der Verdacht, dass ein rein propositionaler Ansatz ungeeignet ist, bestätigt sich auch in einer weiteren Hinsicht: Die Standardanalyse spezifiziert Fragen in Form von *invarianten Antwortbedingungen*. Doch bekanntermaßen ist der klassische Invariantismus in der Erkenntnistheorie äußerst umstritten. Wissens-Zuschreibungen mit interrogativen SKen liefern dafür einen weiteren Anhaltspunkt. Ob eine bestimmte Information eine kongruente Antwort auf die eingebettete Frage des ‚wh'-SK ist, scheint – *relativ zum Ziel der Erkenntnisbemühungen* – von Kontext zu Kontext zu variieren. Hierzu findet sich bei Ginzburg (1995) ebenfalls ein instruktives Beispiel. Betrachten wir die beiden folgenden Situationen:

**Situation (A): Peter befindet sich im Landeanflug auf Helsinki.**

Flugbegleiter: Wissen Sie, wo Sie sind?

Peter: Helsinki.

Flugbegleiter: Ok. Peter weiß, wo er ist.

**Situation (B): Peter steigt gerade aus einem Taxi in Helsinki.**

Taxifahrer: Wissen Sie, wo Sie sind?

Peter: Helsinki.

Taxifahrer: Oh mein Gott, Peter weiß nicht, wo er ist!

Dagegen ließe sich einwenden, dass eine derartige Variabilität von Antwortmengen *kein spezifisches* Problem des Standardansatzes ist.[17] Ziele und Perspektiven scheinen keine zusätzlichen Parameter von Antwortbedingungen zu sein, sondern lediglich *bereichsabhängige* Einschränkungen der „wh"-Phrase, die mit strenger Vollständigkeit vereinbar sind. Dieselbe Variation tritt beispielsweise auch bei Quantifizierungen auf, die sich nicht an (eingebetteten) Fragen orientieren.

Dieser Einwand setzt jedoch bereits eine ganz bestimmte Konzeption der Unterscheidung von semantischen und pragmatischen Bestandteilen der Analyse bereichsspezifischer Einschränkungen der *wh*-Phrase voraus. Demgegenüber ließe sich mit Ginzburg argumentieren, dass Festlegungen im Bereich quantifikationaler Ausdrücke möglicherweise rein semantischer Natur sind; das im Argument angesprochene Phänomen der Relativität von Wissens-Zuschreibungen bringt jedoch einen *zusätzlichen pragmatischen Aspekt* ins Spiel, der sich nicht auf einer (rein) semantischen Ebene erklären lässt. Ich werde im letzten Abschnitt einen Vorschlag unterbreiten, der deutlich werden lässt, worauf diese Trennung basiert bzw. warum im Hinblick auf epistemische Interrogative nur diese Herangehensweise adäquat zu sein scheint.[18]

Unter pragmatischen Gesichtspunkten lässt sich zudem die folgende Schwierigkeit ausmachen: Der reduktionistische Ansatz liefert kein geeignetes „setting" für epistemische

[16] Stanley (2011, 64ff.) hat eingewandt, dass sich dieses Argument an einer allgemeinen Kritik an der *relationalen Analyse* propositionaler Einstellungen festmacht, die sich seiner Meinung nach entkräften lässt. Analog zur Nominalisierung von propositionalen SKen (der Form „die Proposition, dass p") wird dafür argumentiert, dass nominalisierte Frage-SKe im Vergleich zu den meisten interrogativen SKen zu einer *anderen syntaktischen Kategorie* gehören. Ich bin nicht der Ansicht, dass diese auf (rein) syntaktischer Ebene angesiedelte Zurückweisung des obigen Arguments wirklich sinnvoll ist.

[17] Dieser Einwand findet sich unter anderem bei Stanley (2011, 66ff.), Lahiri (2002, 58) und George (2011, 100ff.).

[18] Darüber hinaus hat Aloni (2002) gezeigt, dass, selbst wenn man die semantischen Intuitionen der angesprochenen Erwiderung teilt, die Bereichsabhängigkeit der *wh*-Prase eine weitreichende Revision der Standardsemantik im Hinblick auf die Kontextabhängigkeit solcher Ausdrücke nach sich zieht.

Interrogative, da strenge Vollständigkeit nicht (zumindest nicht immer) mit den *Grice'schen Konversationsmaximen* – insbesondere der Maxime der Quantität („sei so informativ wie möglich") – vereinbar ist:[19]

**Beispiel für eine solche Verletzung der Quantitätsmaxime**

Peter weiß, dass Hans und Barbara zusammen auf der Party gewesen sind, da er gesehen hat, wie sie miteinander geflirtet haben. Er weiß aber nicht, wer sonst noch alles da war. Vor diesem Hintergrund führt Peter mit Klaus (Barbaras eifersüchtigem Freund) den folgenden Dialog:

Klaus: Weißt Du, wer auf der Party war?

Peter: Nein, ich weiß nicht, wer auf der Party war.

Legt man das Prinzip der *strengen Vollständigkeit* zugrunde, ist Peter in der vorliegenden Situation völlig aufrichtig gewesen, denn er hat nichts anderes als die Wahrheit gesagt.

Demgegenüber besagt jedoch die Grice'sche Intuition, dass Peter im vorliegenden Fall nicht aufrichtig war. Wenn Peter im Sinne der Quantitätsmaxime kooperativ gewesen wäre, hätte er gesagt, was er weiß, selbst wenn seine Antwort dabei unvollständig ist. Noch gravierender ist vielleicht ein zweiter Aspekt: Wenn Klaus voraussetzt, dass sich Peter im Sinne der Quantitätsmaxime kooperativ verhält, wird er annehmen, dass Peter keine Information hat, die er als Antwort auf die gestellte Frage anführen kann. Doch das ist dann ganz sicher falsch.

Man könnte an dieser Stelle einwenden, dass es sich hierbei „nur" um einen pragmatischen Effekt handelt. Streng genommen sagt Peter im vorliegenden Dialog die Wahrheit.[20] Diese Erwiderung ist jedoch aus zwei Gründen nicht zulässig: *Erstens* hatte ich bereits erwähnt, dass Befürworter des Begriffs der strengen Vollständigkeit im Zusammenhang mit dem ersten Einwand (Negation von Wissens-Zuschreibungen) dafür argumentiert haben, dass ihre Sichtweise zumindest unter normalen konversationalen Umständen zutreffend ist. Entsprechend kann jetzt nicht behauptet werden, dass die pragmatischen Bedingungen für ihre Position keine Rolle spielen. *Zweitens* ist es einfach falsch, dass Peter streng genommen die Wahrheit sagt. Auch wenn man strenge Vollständigkeit zugesteht, sind zwei Fälle zu unterscheiden: (a) Peter weiß nicht, dass es eine Proposition gibt, die eine Antwort auf die Frage liefert. (b) Peter weiß nicht, dass es eine Proposition gibt, die als Teil einer vollständigen Menge eine Antwort auf die Frage liefert. Gemäß dieser Unterscheidung lässt sich wie folgt argumentieren: Im vorliegenden Dialog sagt Peter streng genommen *nicht* die Wahrheit, denn dazu hätte er etwas in der Art von (b) behaupten müssen. Peter sagt aber etwas in der Art von (a) – und das ist definitiv falsch.

Ein weiteres Grundproblem des reduktionistischen Ansatzes stellen kongruente Antworten dar, die sich in *explanatorischer Hinsicht als irrelevant* erweisen. Ausgangspunkt ist die allgemeine Beobachtung, dass „Warum-Fragen" in der Semantik für responsive Interrogative eine Sonderstellung einnehmen, da sich im Zusammenhang mit solchen Fragen oft überhaupt keine vollständigen Antwortmengen konstruieren lassen (z.B. „Warum ist die Erde rund?"). Diese Sonderstellung trägt im Fall von Wissen dazu bei, dass es Fälle gibt, in denen S eine kongruente Antwort weiß, die Zuschreibung von Wissen aber nicht korrekt ist, da die Antwort im Kontext von S keine explanatorische Relevanz besitzt. Das allgemeine Schema für die Bildung derartiger Problemfälle sieht wie folgt aus:

**Schema für Gegenbeispiele mit eingebetteten „Warum"-Fragen**

S weiß, dass p, p ist eine wahre Antwort, aber S weiß p *nicht als Antwort* auf die indirekte „wh"-Frage.

---

[19] Vgl. zu dieser Art von Einwand: George (2011).

[20] Ein solcher Einwand wurde beispielsweise von Wolfgang Freitag (in mündlicher Konversation) erwähnt.

*Beispiel*: Hans weiß, dass Peter gern lügt. Peters ständiges Lügen ist der Grund, weshalb er von seiner Frau Barbara verlassen wurde. Hans weiß aber *nicht,* dass Peter *aufgrund* seiner Lügen von Barbara verlassen wurde. Nach der Standardauffassung gilt jedoch: Hans weiß, *warum* Peter von Barbara verlassen wurde, da Hans mit einer Proposition bekannt ist („Peter lügt gern"), die eine wahre Antwort auf die eingebettete Frage des ‚wh'-SKs liefert.

Es wäre sicherlich falsch, in diesem Fall zu behaupten, dass Hans weiß, *warum* Peter von Barbara verlassen wurde, obgleich es eine Proposition gibt, die von Hans gewusst wird und die eine Antwort auf die indirekte Frage des ‚wh'-SKs liefert.[21]

Damit komme ich zu einem letzten und – wie sich herausstellen wird – in besonderer Hinsicht richtungsweisenden Problem der reduktionistischen Sichtweise. Im Rahmen der Standardanalyse von Wissen-*wh* werden Antwortbedingungen in Form einer Quantifizierung über entsprechende Propositionen konstruiert. Es gibt jedoch eindeutige Fälle der Zuschreibung von Wissen-*wh*, in denen es unmöglich ist, von der *zuschreibenden* Person zu behaupten, sie würde mit der Äußerung ihrer Zuschreibung eine *existentielle Quantifikation über (wahre) Antwortpropositionen* zum Ausdruck bringen:

**Beispiel für einen solchen Zuschreibungskontext**

Hans: Weißt Du, wann der nächste Zug nach Berlin geht?

Peter: Nein, aber frag doch Barbara.

Hans: Wieso?

Peter: Sie ist Zugbegleiterin. *Barbara weiß, wann der nächste Zug nach Berlin geht.*

Ich denke, es sollte relativ unstrittig sein, dass ein angemessenes „setting" der Zuschreibung von Wissen-*wh* nicht nur aus der Sicht des Wissenden selbst, sondern auch aus der Perspektive der *zuschreibenden Person* gerechtfertigt sein sollte. Diese scheinbar triviale Anforderung wird jedoch im Rahmen der Standardanalyse nicht immer erfüllt.

Dagegen ließe sich Folgendes einwenden: Peters Zuschreibung ist im vorliegenden Beispiel nur dann korrekt, wenn es eine Proposition gibt, die eine (wahre) Antwort auf die eingebettete Frage liefert „Wann geht der nächste Zug nach Berlin?" und *Barbara* diese Proposition kennt. Wenn es das ist, was die Standardanalyse besagt, wo liegt dann das Problem?

Das Problem besteht darin, dass Peter – selbst wenn er unterstellt, dass Barbara irgendeine (wahre) Antwortproposition kennt – mit seiner Äußerung nicht das zum Ausdruck bringt, was nach der Standardauffassung mit einer solchen Zuschreibung von Wissen-*wh* angeblich zum Ausdruck gebracht werden soll. Denn wie das Beispiel zeigt, ist es Peter unmöglich, mittels existenzieller Quantifizierung diejenige Proposition herauszugreifen, die eine (wahre)

[21] Darüber hinaus hat insbesondere Jonathan Schaffer (2007, 2009) zu zeigen versucht, dass sich Probleme mit irrelevanten (kongruenten) Antworten nicht nur in Bezug auf „Warum-Fragen" ergeben, sondern auch in solchen Situationen auftreten, in denen *konvergierende* Fragen dieselben (wahren) Antworten besitzen. Schaffer hat daraufhin einen alternativen (reduktionistischen) Ansatz unterbreitet. Ich denke jedoch, dass dieser Vorschlag aus zwei Gründen problematisch ist: *Erstens* geht sein Ansatz wie die Standardanalyse davon aus, dass es für eine korrekte Zuschreibung von Wissen hinreichend sei, wenn es eine (wahre) Antwortproposition gibt, die von S gewusst wird (gleichwohl Antwortpropositionen nach Schaffer anders strukturiert sind). Entsprechend fällt dieser Vorschlag unter dieselbe Kategorie von Gettier-artigen Gegenbeispielen, die ich bereits in Verbindung mit dem ursprünglichen (reduktionistischen) Ansatz diskutiert habe. *Zweitens* ist keineswegs klar, ob konvergierende Fragen für die Standardanalyse ein ernsthaftes Problem darstellen. Und selbst wenn: Sofern es eine Erklärung dieser Schwierigkeit gibt, die dasselbe leistet und im Gegensatz zu Schaffer dabei bleibt, dass Wissen eine zweistellige Relation ist, würde sich sein Vorschlag als unnötig revisionistisch erweisen. Vgl. zur Kritik an Schaffers Einwand gegenüber der Standardanalyse: Brogaard (2009), Kallestrup (2009), Stanley (2011).

Antwort auf die eingebettete Frage des SKs liefert. Nichtsdestotrotz würden wir im vorliegenden Fall sagen, dass Peters Äußerung wahr ist. Entsprechend muss der Grund für diese Annahme ein anderer sein. Darüber hinaus gibt es zweifelsohne Fälle, in denen eine derartige Äußerung in epistemischer Hinsicht rational erscheint, *weil* die zuschreibende Person glaubt, dass sie wahr ist. Wenn der vorliegende Fall ein solcher ist, Peter aber keine (wahre) Antwortproposition auf die eingebettete Frage kennt, dann lässt sich mit der Standardauffassung nicht erklären, warum wir im vorliegenden Fall die Intuition haben, dass Peters Zuschreibung von Wissen-*wh* kein Akt subjektiver Willkür ist, sondern selbst als rational zu betrachten ist.

Wie lässt sich dieses Problem lösen? Ein wichtiger Schritt besteht darin, die beiden folgenden Aspekte voneinander zu trennen:[22]

(A) Analyse der (logischen) *Form der Wissens-Relation*

(B) Charakterisierung der Umstände, unter denen jemand *in dieser Wissens-Relation steht*

Für gewöhnlich werden diese beiden Aspekte nicht getrennt. Das scheint jedoch ein Fehler zu sein. Meines Wissens ist Kent Bach (Bach 2005) der Erste gewesen, der in einem bislang unveröffentlichten Aufsatz auf dieses Problem hingewiesen hat. Beispiele der eben angeführten Art zeigen nämlich Folgendes: Offenkundig ist es möglich, jemandem in einer angemessenen Form Wissen-*wh* zuzuschreiben, ohne dass diejenige Person, von der die Zuschreibung ausgeht, selbst *in einer Relation des Wissens-wh* stehen muss. Das heißt nun aber nicht, dass die Äußerung, in der die betreffende Zuschreibung zum Ausdruck gebracht wird, nicht die (logische) Form einer Wissens-Relation hat. Vielmehr müssen an dieser Stelle zwei Aspekte der allgemeinen Analyse von Wissen-*wh* unterschieden werden: Der erste Aspekt betrifft die *semantische* Bedingung der Zuschreibung. Diese muss erfüllt sein, sofern das, was in der Zuschreibung ausgedrückt werden soll, die (logische) Form einer Wissens-Relation besitzt. Der zweite Aspekt betrifft hingegen die *normativen* Umstände der Zuschreibung von Wissen. Bedingungen dieser Art legen unter anderem fest, wann es gerechtfertigt ist anzunehmen, dass diejenige Person, der Wissen zugeschrieben wird, „in einer Relation des Wissens-*wh*“ steht. Eine Trennung beider Aspekte ist durch die Tatsache begründet, dass es Fälle der Zuschreibung von Wissen-*wh* gibt, in denen die erste Bedingung zwar erfüllt ist – d.h. es handelt sich bei der zum Ausdruck gebrachten Zuschreibung der Form nach um eine Relation des Wissens zwischen S und der im interrogativen SK eingebetteten Frage Q –, gleichwohl liefert die Erfüllung dieser semantischen Bedingung keine hinreichende Basis für die Erklärung, warum eine solche Zuschreibung in epistemischer Hinsicht *korrekt* ist. Denn das Stehen in einer Relation des Wissens-*wh* – d.h. die Bekanntheit mit der Menge von Tatsachen, die eine kongruente Antwort auf die im interrogativen SK eingebettete Frage Q liefern – ist keine Voraussetzung für diejenige Person, die Wissen zuschreibt. Sofern wir nicht unterstellen wollen, dass Wissens-Zuschreibungen, die von solchen Personen ausgehen, willkürlich erfolgen, bedarf es einer zusätzlichen (normativen) Erklärung jener Umstände, unter denen Zuschreibungen von Wissen aus der Perspektive der zuschreibenden Person angemessen sind. Oder anders gesagt: Es bedarf einer Erklärung, wann man unterstellen darf, dass derjenige, dem mithilfe von Äußerungen der (logischen) Form „S weiß-*wh*“ eine epistemische Einstellung zugeschrieben wird, in der betreffenden Wissens-Relation steht.

Aufbauend auf dieser Differenzierung werde ich im abschließenden Teil zu skizzieren versuchen, wie ein „setting“ der Zuschreibung von Wissen-*wh* – das in semantischer Hinsicht nicht reduktiv und in normativer Hinsicht geeignet erscheint – aus meiner Sicht zu entwickeln ist.

[22] Vgl. Bach (2005), Masto (2010).

## 5. Ausblick – Verstehen als eine normative Quelle der Zuschreibung von Wissen-*wh*

Im Folgenden gehe ich von drei zentralen Annahmen aus: *Erstens* scheint es mir aufgrund der vorgebrachten Einwände angebracht zu sein, einen *nicht-reduktionistischen* Erklärungsansatz der Bedeutung von Sätzen der Form „S weiß-*wh*“ zu favorisieren. *Zweitens* plädiere ich im Anschluss an Bach für die *Trennung zweier Aspekte* der Analyse von Wissen-*wh*: einerseits die (semantische) Erklärung der logischen Form der Wissens-Relation und andererseits die Erklärung der normativen Bedingungen, unter denen jemand in einer solchen „Wissen-*wh*“-Relation steht. *Drittens* behaupte ich, dass im Rahmen einer Erklärung des zweiten (normativen) Aspekts nicht nur die *Variabilität kongruenter Antworten* (im Hinblick auf die Ziele unserer Erkenntnisbemühungen), sondern auch die spezifische *Perspektive der zuschreibenden Person* berücksichtigt werden muss.

Aufbauend auf diesen drei Grundannahmen lässt sich der von mir favorisierte Ansatz nun etwas genauer präzisieren. Die erste Präzisierung betrifft die *Form der Wissensrelation*. Meiner Ansicht nach können wir es dabei belassen, Wissen als eine binäre Relation zu betrachten. Diese Relation ist jedoch nicht – wie von Reduktionisten behauptet – rein propositionaler Natur. Vielmehr handelt es sich um eine *objektbezogene Relation der Bekanntschaft*, die zwischen einem *Subjekt S*, nämlich dem Wissenden, und dem jeweiligen *Objekt des Wissens* besteht. Im Fall der Zuschreibung von Wissen-*dass* sind Propositionen diejenigen Objekte, mit denen S bekannt sein muss, wohingegen es sich bei Wissen-*wh* bei den betreffenden Objekten um eingebettete Fragen handelt. Die (binäre) Wissens-Relation besitzt hier die allgemeine (logische) Form *[S-Q]*.

Diese Überlegung deckt sich mit der eingangs erläuterten Beobachtung, dass Wissen als ein prototypisches responsives Prädikat (im Gegensatz zu rogativen „question embedders“) sowohl mit der Verwendung von deklarativen als auch interrogativen SKen vereinbar ist. Dennoch fallen beide Arten der Zuschreibung von Wissen nicht einfach zusammen. Vielmehr unterscheiden sie sich in zweierlei Hinsicht: im Hinblick auf die *Art der Objekte*, mit denen der Wissende bekannt sein muss, und im Hinblick auf die *Art der normativen Bedingungen*, gemäß denen es angemessen ist, eine Zuschreibung mit der entsprechenden (logischen) Form zum Ausdruck zu bringen.

Nach dem vorliegenden Ansatz ist Wissen nicht (lexikalisch) ambig. Da der Unterschied von Wissen-*dass* und Wissen-*wh* nicht an der (logischen) Form der Wissens-Relation festgemacht wird, sondern lediglich die Art der Objekte betrifft (d.h. ihre kategoriale Beschaffenheit), handelt es sich in beiden Fällen um ein und dieselbe binäre Relation, nämlich um die Bekanntschaft von S mit dem jeweiligen Objekt des Wissens (Propositionen bzw. Fragen). Darüber hinaus besteht keine grundlegende Asymmetrie im semantischen Gebrauch interrogativer Prädikate: Responsive und (rein) rogative „questions embedders“ bringen dieselbe Relation [S-Q] zum Ausdruck. Verschieden sind jedoch die normativen Umstände, unter denen es angemessen ist, S zu unterstellen, dass S in der jeweiligen Relation des Wissens steht. Im Gegensatz zu (rein) rogativen Prädikaten haben responsive Interrogative so etwas wie ein erkenntnistheoretisches Design: *Bei ihnen basiert die Relation [S-Q] auf einer Bekanntschaft mit Tatsachen in der Welt, die als Antwort auf die (eingebettete) Frage den rogativen Impuls (notwendig) zum Stillstand bringen.*

Die zweite wichtige Präzisierung betrifft den *normativen Aspekt* der Zuschreibung von Wissen. Wann sind wir berechtigt, anzunehmen, dass jemand *in einer Relation* des Wissens-*wh* steht? Nach der vorangegangenen Diskussion muss diese Erklärung zwei Dinge leisten: Sie muss einerseits mit einem nicht-reduktionistischen Programm der semantischen Erklärung der (logischen) Form der Zuschreibung Wissen-*wh* vereinbar sein. Und sie muss andererseits die Variabilität der Antwortbedingungen auf eine Weise berücksichtigen, dass

nicht der Eindruck entsteht, die Angemessenheit der Zuschreibung von Wissen-*wh* unterläge der bloßen (subjektiven) Willkür der zuschreibenden Person.

Die grundlegende Idee, die ich in diesem Zusammenhang verfolge, setzt ein *expressivistisches ‚framework'* voraus. Grob gesagt ist damit Folgendes gemeint: Wenn eine Person A eine Zuschreibung der Form „S weiß-*wh*" äußert, dann bringt A mit dieser Äußerung einen mentalen Zustand zum Ausdruck, der mit einer *Proeinstellung (Billigung)* gegenüber der epistemischen Position von S verbunden ist. Derartige Proeinstellungen entspringen jedoch nicht der Willkür subjektiver Interessen; sie sind vielmehr an normative Standards gebunden, welche die *objektive Angemessenheit* solcher Einstellungen regulieren. Nur relativ zu diesen Standards ist A darin gerechtfertigt, anzunehmen, dass S in einer epistemischen Position ist, gemäß der unterstellt werden darf, dass S in einer Relation des Wissen-wh zur eingebetteten Frage des SKs steht.

Wie lässt sich die Bezugnahme auf einen normativen Standard unter den vorliegenden Bedingungen charakterisieren? Ich gehe an dieser Stelle von der Annahme aus, dass ein basaler Zusammenhang zwischen der Zuschreibung von Wissen-*wh* und derjenigen Hinsicht existiert, unter der eine bestimmte Art von Antwort den interrogativen Impuls der eingebetteten Frage des SKs auflöst. Um in einer Relation des Wissens-*wh* stehen zu können, muss S die *Hinsicht verstanden* haben, unter der eine (kongruente) Antwort dem *Ziel der erkenntnistheoretischen Untersuchung* dient, d.h. den damit verbundenen Erkenntnisbemühungen Nachdruck verleiht. Nur wenn S wirklich verstanden hat, inwiefern eine Beantwortung der eingebetteten Frage dazu führt, dem Ziel der epistemischen Untersuchung näherzukommen, ist es für S möglich, diejenige Menge von Antworten (Tatsachen in der Welt) zu selegieren, die den interrogativen Impuls *unter epistemischen Gesichtspunkten* zum Stillstand bringen. Kongruente Antworten, deren responsiver Status nicht darin begründet ist, dass S die Hinsicht verstanden hat, unter der die Beantwortung der Frage dem Ziel der erkenntnistheoretischen Untersuchung dient, mögen unter praktischen, politischen, ästhetischen oder anderen Gesichtspunkten korrekt sein – sie rechtfertigen jedoch nicht die Annahme, dass S in einer Relation [S-Q] steht, deren *Bekanntschaft genuin epistemischer Natur* ist. Entsprechend gilt für die zuschreibende Person: Wenn nicht klar ist, ob S die Hinsicht verstanden hat, unter der eine Beantwortung der im SK eingebetteten Frage dem Ziel der epistemischen Untersuchung dient, mag S noch so viele (kongruente) Antworten kennen, die zuschreibende Person ist in diesem Fall nicht gerechtfertigt zu unterstellen, dass S in einer Wissens-Relation [S-Q] steht. Entsprechend lässt sich der normative Standard wie folgt formulieren:

> (V) Man sollte nur dann unterstellen, dass S in einer Relation des Wissens-*wh* steht, wenn S verstanden hat, in welcher Hinsicht eine (kongruente) Antwort auf die eingebettete Frage des *wh*-SKs dem (genuinen) Ziel der erkenntnistheoretischen Untersuchung dient.[23]

Ausgehend von dieser Charakterisierung vertrete ich bezüglich der Erklärung des *normativen Aspekts* der Zuschreibung von Wissen-*wh* den folgenden allgemeinen Ansatz:

---

[23] Es ist an dieser Stelle jedoch vor Folgendem zu warnen: Mit der Behauptung von (V) ist *keine* inhaltliche Festlegung getroffen, was genau es heißt, dass etwas ein (genuines) Ziel der erkenntnistheoretischen Untersuchung ist. Insbesondere geht der vorliegende Ansatz nicht die Verpflichtung ein, behaupten zu müssen, dass die Hinsicht, in der etwas dem Ziel der Erkenntnisbemühungen dient, immer darin besteht, das Erlangen wahrer Überzeugungen zu vermehren respektive das Erlangen falscher Überzeugungen zu vermeiden. Ein normativer Standard wie (V) impliziert nicht, dass Wahrheit das einzige oder primäre Ziel der epistemischen Praxis ist. Vielmehr habe ich an anderer Stelle (Schmechtig 2010, *forthcoming*) dafür argumentiert, dass es relativ zur Art der epistemischen Praxis und den damit verbundenen Wertschätzungen variieren kann, in welcher Hinsicht eine kognitive Aktivität dem Ziel der erkenntnistheoretischen Untersuchung dient.

**Normativ-expressivistischer Erklärungsansatz**

Mit der Äußerung eines Tokens der Form „S weiß-*wh*“ bringt die zuschreibende Person A zum Ausdruck, dass S in einer epistemischen Position ist, im *Hinblick auf einen Korrektheitsstandard (V)* eine kongruente Antwort auf die eingebettete Frage des *wh*-SKs geben zu können.

Nach diesem Ansatz ist eine Zuschreibung der Form „S weiß-*wh*“ nicht *per se* aufgrund der semantischen Bedingungen korrekt oder inkorrekt, sondern nur *relativ* zu einem Standard wie (V). Dieser Standard ist eine Art *normative Quelle*, relativ zu der die zuschreibende Person A entscheiden kann, ob Ss Bekanntschaft mit der eingebetteten Frage des *wh*-SKs in epistemischer Hinsicht korrekt ist, d.h. ob S in einer Relation des Wissens-*wh* steht.

So weit in aller Kürze zu einigen Eckpunkten des von mir in Anspruch genommenen Ansatzes. Abschließend möchte ich noch die Vorzüge dieser Herangehensweise benennen: Mit der vorliegenden Erklärung des normativen Aspekts verbindet sich ein *nicht-reduktionistisches* Programm der semantischen Analyse von Wissen-*wh*. Aus diesem Grund stellen Fälle der *Negation von Wissens-Zuschreibungen* – die sich für reduktive Intellektualisten wie Stanley als problematisch erwiesen haben – keine ernsthafte Schwierigkeit mehr dar. Im Rahmen eines solchen nicht-reduktionistischen Programms wird die problematische *Äquivalenzbehauptung* aufgegeben. Das heißt nun aber nicht, dass der Begriff der weiten Vollständigkeit besser dasteht. Im Gegenteil: Es scheint Situationen zu geben, in denen die Forderung nach strenger Vollständigkeit durchaus berechtigt und kohärent ist. Eine nicht-reduktionistische Erklärung der Bedeutung von Wissen-*wh*, bei der auf die problemerzeugende Äquivalenzbehauptung verzichtet wird, trägt dieser Tatsache Rechnung; sie ist damit vereinbar, dass in manchen Situationen strenge Antwortbedingungen gefordert sind.

Auf der anderen Seite entgeht der vorliegende Ansatz jenen Einwänden, die innerhalb der reduktionistischen Standardauffassung zu einer zweifelhaften Erklärung der „mention-some“-Lesart von responsiven Interrogativen geführt haben. Insbesondere scheint es so zu sein, dass das angeführte Gettier-artige Szenario unter Hinzunahme der vorgeschlagenen Erklärung des *normativen* Aspekts von Wissen-*wh* nicht mehr greift. Wenn wir annehmen – wie es in der Regel getan wird – , dass im Rahmen einer Gettier-artigen Untersuchung das bloße zufällige Erlangen von wahren Überzeugungen nicht ausreicht, sondern vielmehr Wissen das eigentliche Ziel der Untersuchung ist, und wir zusätzlich unterstellen, dass eine Zuschreibung in der Art von (16) unter anderem deshalb gerechtfertigt ist, weil sie mit (V) im Einklang steht, dann wird schnell klar, warum (17) in epistemischer Hinsicht nicht korrekt ist, obwohl diese Art von Zuschreibung unter rein semantischer Perspektive mit (16) äquivalent zu sein scheint (d.h. dieselbe wahre Antwortproposition besitzt). Unter Anwendung von (V) lässt sich nur dann behaupten, dass S in einer Relation des Wissens-*wh* steht, wenn S eine (wahre) Antwort auf die eingebettete Frage des betreffenden SK kennt („Wer hat die Tour de France 2012 gewonnen?“) *und* S gleichzeitig die *Hinsicht verstanden* hat, unter der diese Antwort dem Ziel der Untersuchung dient. Unstrittig ist, dass S im Fall der Äußerung von (17) eine in semantischer Hinsicht (kongruente) Antwort auf die eingebettete Frage des betreffenden SKs kennt (die Proposition, dass Bradley Wiggins gewonnen hat). Fraglich ist jedoch, ob die zuschreibende Person in diesem Fall unterstellen darf, dass S *in einer Relation* des Wissens-*wh* steht. Angesichts der seltsamen Hintergrundüberzeugungen, die S besitzt (S glaubt, dass der Gewinner der „Tour de France 2012“ ein Eskimo ist und im „Team Telekom“ fährt), hat die zuschreibende Person A guten Grund, daran zu zweifeln. Wiewohl S mit einer Proposition bekannt ist, die eine (kongruente) Antwort auf die eingebettete Frage des SKs liefert, hat S offenkundig *nicht* die Hinsicht verstanden, unter der eine (kongruente) Antwort auf die eingebettete Frage dem Ziel der Erkenntnisbemühungen dient – zumindest dann nicht, wenn die zugrunde liegenden Praxis der Wissens-Zuschreibung eine Gettier-artige Untersuchung ist, bei der das Erlangen von

Wissen das Ziel der Erkenntnisbemühungen ist. A hat im Fall von (17) guten Grund, daran zu zweifeln, dass die semantische Adäquatheit der von S gewussten Antwortproposition auf einer Bekanntschaft mit Tatsachen beruht, die im Rahmen der vorliegenden Untersuchung dazu berechtigen anzunehmen, dass S tatsächlich in einer Relation des Wissens-*wh* zur eingebetteten Frage des SKs steht. Da S nicht die Hinsicht verstanden hat, unter der einer in semantischer Perspektive äquivalente Antwortproposition dem eigentlichen Ziel der erkenntnistheoretischen Untersuchung dient, kann A nicht unterstellen, dass S in einer Relation [S-Q] steht, die nicht nur den semantischen, sondern auch den *normativen* Aspekt der Zuschreibung erfüllt, d.h. eine in epistemischer Hinsicht *korrekte* Zuschreibung von Wissen-*wh* ist.

Darüber hinaus liefert die vorgeschlagene Herangehensweise ein geeignetes „setting" der Zuschreibung von Wissen-*wh*. Sie führt nicht mehr zu einer gravierenden Verletzung von Konversationsmaximen und behandelt das Problem der explanatorisch irrelevanten Antworten auf eine Weise, bei der keine Abstriche bezüglich der Einheitlichkeit der Wissens-Zuschreibung gemacht werden müssen. Weder ist die (logische) Form der Wissens-Relation uneinheitlich, noch ist der Gebrauch von interrogativen Prädikaten – wie es bei alternativen (reduktionistischen) Ansätzen der Fall zu sein scheint – als grundlegend asymmetrisch zu betrachten.[24] Ebenso wenig wird eine revisionistische Position in Bezug auf die aus meiner Sicht völlig plausible Annahme vertreten, der zur Folge Wissen-*wh* ein prototypisches responsives Interrogativ ist, das eingebettete Fragen zum Gegenstand hat.[25]

Eine zentrale Eigenschaft des vorgeschlagenen Ansatzes besteht zudem darin, die hervorgehobene Kontextsensitivität (kongruenter) Antworten nicht als ein Problem der (logischen) Form der Wissens-Relation zu betrachten. Stattdessen wird die Variabilität der Antwortbedingungen als Ausdruck der *Relativität des normativen Aspekts* der Analyse von Wissen-*wh* verstanden. Es ist nicht die (logische) Form der Zuschreibung, die innerhalb der verschiedenen Kontexte variiert. Vielmehr sind es die Bedingungen, unter denen die zuschreibende Person *gerechtfertigt* ist anzunehmen, dass jemand in einer Relation des Wissens-*wh* steht. Damit ist Folgendes gemeint: Man ist nur dann gerechtfertigt anzunehmen, dass S in einer Relation des Wissens-*wh* steht, wenn aus der Menge der Antwortbedingungen, die semantisch adäquat sind, diejenigen selegiert werden, die im Hinblick auf das eigentliche Ziel der erkenntnistheoretischen Untersuchung den interrogativen Impuls der im SK eingebetteten Frage auflösen. Da jedoch die Ziele der Untersuchung in Abhängigkeit von der jeweiligen Erkenntnispraxis veränderlich sind, ist klar, dass die Menge der (kongruenten) Antwortpropositionen, die in epistemischer Hinsicht adäquat erscheinen, mit der Art des Erkenntnisziels variieren. Der vorliegende Ansatz wird dieser Überlegung gerecht, indem er den normativen Aspekt der Zuschreibung von Wissen-*wh* als *Hinsicht* charakterisiert, unter der eine Äußerung der Form „S weiß-*wh*" als korrekt bzw. inkorrekt zu bezeichnen ist. Derartige Hinsichten lassen sich nicht absolut, sondern nur relativ zu Zielen der Erkenntnisbemühungen bestimmen.

---

[24] Eine solche Kritik wird beispielsweise von Masto (2010) an Schaffers Ansatz geübt.

[25] Berit Brogaard (2009) hat beispielsweise vorgeschlagen, die Zuschreibung von Wissen-*wh* als eine Art *de re Wissen* zu betrachten, das S von demjenigen Objekt hat, auf das ein ‚wh'-Prädikat angewandt wird. Ihrer Ansicht nach besteht eine unmittelbare Verbindung zwischen ‚wh'-Klauseln, die durch ein Einstellungsverb wie Wissen eingebettet werden, und ‚wh'-Klauseln in sog. Pseudo-Spaltsätzen (*pseudo-clefts*) wie z.B. „Was Hans gestern zerbrochen hat, war seine Vase". Die in Pseudo-Spaltsätzen auftretenden ‚wh'-Klauseln lassen sich diesem Ansatz nach wie gewöhnliche Prädikate (ähnlich definiter bzw. indefiniter Beschreibungen) behandeln. Ein Satz der Form „S weiß, warum A verliebt ist" ist nach diesem Vorschlag wie folgt zu lesen: Es existiert ein Grund G, von dem S weiß, dass G die Eigenschaft hat, zu erklären, warum A verliebt ist. Eine solche *de re* Analyse ist jedoch extrem revisionistisch; sie geht davon aus, dass ‚wh'-Klauseln *keine indirekten Fragen* beinhalten, sondern gewöhnliche Prädikate sind. Das ist intuitiv wenig einleuchtend.

Auf derselben Grundlage ist es möglich, diejenigen Fälle zu rekonstruieren, durch die der reduktionistischen Standardanalyse der Einwand gemacht werden kann, dass sie die Zugänglichkeit der Antwortbedingungen nicht aus der Perspektive der zuschreibenden Person erklären kann. Der vorliegende Ansatz vermeidet ganz bewusst, von der zuschreibenden Person A zu verlangen, sie müsse mit einer Äußerung der Form „S weiß-*wh*" eine existenzielle Quantifizierung über eine (wahre) Antwortproposition zum Ausdruck bringen. Zwar muss für eine adäquate Äußerung in semantischer Hinsicht unterstellt werden, dass S eine (wahre) Antwort auf die eingebettete Frage des betreffenden SK kennt; aber nach dem vorliegenden Ansatz ist das, was A zum Ausdruck zu bringen versucht, etwas anderes. Andererseits ist es so, dass die Art der Rechtfertigung einer solchen Zuschreibung nicht auf bloßer subjektiver Willkür beruht. A ist nur dann gerechtfertigt zu unterstellen, dass S in einer Relation des Wissens-*wh* steht, wenn A davon ausgehen kann, dass sich S in einer epistemischen Position befindet, die im Einklang mit einem Standard wie (V) steht. A muss davon überzeugt sein, dass S die Hinsicht verstanden hat, unter der eine Beantwortung der eingebetteten Frage (des betreffenden SK) dem Ziel der erkenntnistheoretischen Untersuchung dient. Ist dieser Standard verletzt, ist A nicht gerechtfertigt, von S zu behaupten, dass S in einer Relation des Wissens-*wh* steht – selbst dann nicht, wenn S eine kongruente Antwort kennt, aufgrund der eine Zuschreibung der Form „S weiß-*wh*" in semantischer Hinsicht adäquat erscheint. Mit anderen Worten, was A mit einer Äußerung der Form „S weiß-*wh*" zum Ausdruck bringen will, ist die Tatsache, dass S in einer Relation des Wissens-*wh* steht. Dies ist nur dann der Fall, wenn (i) S eine (kongruente) Antwort auf die eingebettete Frage des betreffenden SK kennt und (ii) S in einer Position ist, wonach die betreffende Antwort in epistemischer Hinsicht korrekt ist, d.h. im Einklang mit dem oben angeführten Standard (V) steht.[26]

Damit erfüllt der vorgeschlagene Ansatz wesentliche Punkte einer auch unter epistemischen Gesichtspunkten adäquate Analyse von Wissen-*wh*. Ich komme daher zu dem Ergebnis, dass eine kombinierte Analyse – bestehend aus einer nicht-reduktionistischen Semantik der Wissens-Relation plus einer normativ-expressivistischen Erklärung des „Stehens in dieser Relation" – prinzipiell in der Lage sein sollte, ein geeignetes „setting" der Zuschreibung von Wissen-*wh* zu liefern, das Einwänden, die sich gegenüber der reduktionistischen Sichtweise erheben lassen, auf Dauer standhält.

**Pedro Schmechtig**

Technische Universität Dresden
Institut für Philosophie
Lehrstuhl für theoretische Philosophie
D-01069 Dresden
Pedro.Schmechtig@gmx.de

## Literatur

Aloni, M. 2002. "Questions under cover", in Words, Proofs, and Diagrams. California: CSLI Stanford.

Bach, K. 2005. "Questions and answers, comments on Jonathan Schaffer's 'Knowing the Answer'", Bellingham Summer Philosophy Conference, August 2005.

[26] Ein weiterer Vorteil des vorliegenden Ansatzes besteht darin – ohne dies an dieser Stelle weiter erläutern zu können –, dass er sich nahtlos in eine Erklärung für praktisches Wissen (*knowledge-how*) einfügt.

Beck, S. & Rullmann, H. 1999. “A flexible approach to exhaustivity in questions”, Natural Language Semantics 7, 249-298.

Böer, S. & Lycan, W. 1986. Knowing who, Cambridge: Cambridge University Press.

Braun, D. 2006. “Now you know who Hong Oak Yun is”, in E. Sosa and E. Villanueva, (Hrg.): Philosophical Issues 16, 24–42.

Broogard, B. 2009. “What Mary did yesterday: Reflections on knowledge-wh”, Philosophy and Phenomenological Research 78 (2), 439-467.

Égré, P. & Spector, B. 2007. “Embedded questions revisited: an answer, not necessarily the answer”, Presentation at MIT Ling-Lunch Seminar and Journées Sémantique et Modélisation.

George, B. 2011. Question Embedding and the Semantics of Answers, PhD thesis, UCLA.

Ginzburg, J. 1995. “Resolving questions I & II’”, Linguistics and Philosophy 18, 459–527 & 567–609.

Ginzburg, J. 2011. “How to Resolve *How to*”, in J. Bengson und M. Moffett (Hrg.): Knowing How, Oxford: oxford University Press, 215-243.

Groenendijk, J. & Stockhof, M. 1982. “Semantic Analysis of wh-Complements”, *Linguistics and Philosophy* 5: 175-233.

Groenendijk, J. & Stockhof, M. 1984. *Studies in the Semantics of Questions and the Pragmatics of Answers*, Ph.D. thesis, University of Amsterdam.

Hamblin, C. 1958. “Questions”, *Australasian Journal of Philosophy* 36, 159-168.

Hamblin, C. 1973. “Questions in Montague English”, *Foundations of Language* 10, 41-53.

Heim, I. 1994. “Interrogative semantics and Karttunen's semantics for know”, in Rhonna Buchalla and Anita Mittwoch (Hgg.), IATL 1. Hebrew University of Jerusalem, 128-144.

Higginbotham, J. 1996. “The semantics of questions”, in: Shalom Lappin (ed.): *The Handbook of Contemporary Semantic Theory*, Blackwell, 361-383.

Higginbotham, J. & May, R. 1981. “Questions, quantifiers and crossing”, *The Linguistic Review* 1, 41–80.

Hintikka, J. 1976. *The Semantics of Questions and the Questions of Semantics*. North Holland Publishing Company, Amsterdam, 1976.

Kallestrup, J. 2009. “Knowledge-wh and the Problem of Convergent Knowledge”, *Philosophy and Phenomenological Research* 78, 468-477.

Karttunen, L. 1977. “Syntax and semantics of questions”, *Linguistics and Philosophy* 1, 3-44.

Krifka, M. 2001 “Quantifying into question acts”, *Natural Language Semantics* 9, 1-40.

Lahiri, U. 2002. *Questions and Answers in Embedded Contexts*. Oxford: Oxford University Press.

Lewis, D. 1982. “Whether report”, in *Philosophical Essays*, dedicated to Lennart Aqvist on his Fiftieth Birthday, 194-206.

Masto, M. 2010. “Questions, answers, and knowledge-wh”, *Philosphical Studies* 147, 395–413.

Schaffer, J. 2007. “Knowing the answer”, *Philosophy and Phenomenological Research* 75,383-403.

Schaffer, J. 2009. „Knowing the Answer Redux: Replies to Broogard and Kallestrup”, *Philosophy and Phenomenological Research* 78, 477-500.

Schmechtig, P. 2009. „Epistemische Ziele und die Angemessenheit von Wert-Einstellungen“, in G. Schönrich (Hrsg.): Wissen und Werte, Paderborn: mentis 2009, S. 253-292.

Schmechtig, P. 2011. “Expressivismus und der (relative) Wert des Wissens“, in Beiträge des 34. Internationalen Wittgenstein Symposiums, hgg. v. Ch. Jäger/ W. Löffler, Kirchberg am Wechsel, 263-265.

Schmechtig, P. (*forthcoming*). “Expressivism and the (relativ) Value of Knowledge”.

Sharvit, Y. 2002). “Emedded questions and `de dicto' readings'”, *Natural Language Semantics* 10, 97-123.

Spector, B. 2007. “Modalized questions and exhaustivity”, *Proceedings of Semantics and Linguistic Theory* 17. CLC publications.

Stanley, J. 2011. *Know How*, Oxford University Press, 2011.

van Rooij, R. 2004. “The utility of mention-some questions”, *Research on Language and Computation* 2, 401-416.

Williamson, T. & Stanley, J. 2001. “Knowing how”, *Journal of Philosophy* 98, 411-444.

# Practical Knowledge

Michael Schmitz

The contribution deals with knowledge of what to do, and how, where, when and why to do it, as it is found in a multitude of rules, procedures, maxims, plans, and other instructions. It is argued that while this knowledge is conceptual and propositional, it is still irreducible to theoretical knowledge of what is the case and why it is the case. It is knowledge of goals, of ends and means, rather than of facts. It is knowledge-to that is irreducibly practical in having world to mind direction of fit and the essential function of guiding as yet uncompleted action. While practical knowledge is fundamentally different from theoretical knowledge in terms of mind-world relations, the practical and theoretical domains are still parallel in terms of justificatory and inferential relations, they are like mirror images of one another. It is shown that if this view of practical knowledge is accepted, convincing Gettier cases for practical knowledge can be constructed. An extensive analysis of these cases demonstrates the usefulness of the notions of practical deduction, abduction, and induction.

## 1. Introduction: Knowledge and the Theory Bias

There is a ubiquitous bias for the theoretical over the practical in contemporary philosophy that I will call the "theory bias." It is manifest, for example, in the idea of a general truth-conditional semantics; of deductive reasoning as being theoretical and logical consequence as being entirely accountable in terms of the preservation of truth; in cognitivism in metaethics; in accounts of actional experience that treat it as a kind of perceptual experience, and accounts of intention that treat it as a kind of belief. But the theory bias is perhaps nowhere more evident and more pervasive than in philosophical accounts of knowledge. Whereas theoretical knowledge has always been a central topic, even an obsession, of Western philosophy, there is comparatively very little work on practical knowledge. Indeed knowledge is traditionally defined as justified true belief, and even those who reject the traditional definition tend to think of knowledge as a form of belief. But it is, to say the least, not obvious that knowledge of what to do, and where, when and how to do it are forms of belief.

Moreover, when practical knowledge is discussed, the notion of practical knowledge in play tends to be rather restricted. So for Elizabeth Anscombe (1957) and her followers practical knowledge seems to be just a special kind of knowledge of what is the case, based on special, practical sources of evidence like one's own actions or plans. The prime example would be knowledge of what one is currently doing based on the experience of doing it rather than on observation. But it is hard to see how, for example, knowledge of what to do in the event of a fire could be assimilated to this way of thinking about practical knowledge. The other topic that has been discussed under the heading of "practical knowledge" is of course know-how in the sense of skill. This also plays some role in Anscombe, but it is most closely associated with Gilbert Ryle (1949), who famously and influentially argued that know-how had been disregarded in favour of knowing that. Ryle took know-how to exclusively consist in non-discursive skills and capacities and did not discuss, at least not under this heading, the vast discursive, conceptual practical knowledge we have in the form of maxims, rules, recipes, and other instructions, ranging from how to make pie and atom bombs to what to do in the event of a fire or earthquake.

Recently Stanley & Williamson (2001; see also Stanley 2011) have challenged the established Rylean view of know-how. They argue that all know-how (or "knowledge-how") is reducible to propositional knowledge-that of ways of performing actions. There are no Rylean irreducible skills – or at least they should not be called "knowledge-how." For example, on their view to know how to ride a bicycle is to know, with regard to some x, that x is a way of riding a bicycle, with the further condition that this way of riding a bicycle be represented under "a practical mode of representation." I believe that Stanley & Williamson's account is an improvement over Ryle's in so far as – but only insofar as – it emphasizes that knowledge-how can be conceptually articulated. But the problem with their account is not only that it falls into the opposite extreme of Ryle's by denying that "know-how" ever refers to mere skill and by thus reducing all know-how to conceptual, propositional knowledge. Naturally interpreted, Stanley & Williamson further propose to reduce practical conceptual knowledge of how to do things to theoretical conceptual knowledge of what is the case. To put the same point differently, they reduce knowledge of means and ends to knowledge of facts. Strikingly, Stanley & Williamson seem to take this aspect of their position for granted and show no sign of even being aware that they are engaged in this second reductionist project. They only discuss the first project of reducing all know-how to conceptual knowledge, and the sizable literature that their article has inspired, while often critical of their reduction of skill, has followed them in this regard. There thus appears to be a blind spot in the current philosophical outlook for knowledge that is conceptual but yet irreducibly practical in the sense of being knowledge of goals, of means and ends, rather than of facts. I think this pattern supports the diagnosis of a deep-seated theory bias in contemporary philosophy. The mind-set behind this might be glossed as follows: if it is thought, if it is conceptual, intellectual, if it essentially involves reasoning, it surely must be theoretical. The practical, if it exists at all, is just the lower, non-intellectual level of non-conceptual, non-discursive skills.

Against this, I want to argue in this paper that practical knowledge can neither be reduced to mere skill nor to a special way of knowing what is the case. Rather, states of practical knowledge are conceptually structured attitudes that are irreducibly practical in that they have world-to-mind direction of fit (Anscombe 1957, Searle 1983). That is, fit between mind and world is achieved by adapting the world to the contents of the mind rather than vice versa. These states are prescriptive rather than descriptive, and they constitute knowledge of goals, of ends and means, rather than of facts.

If this is right, practical knowledge is diametrically opposed to theoretical knowledge in terms of world-mind relations. But at the same time I want to show that there are deep structural parallels between practical and theoretical knowledge: they are mirror images of one another. For example, theoretical knowledge is a well-justified and successful kind of theoretical attitude, and we are often looking for knowledge of the causes of given effects, while practical knowledge is a well-justified and successful kind of practical attitude, and we often want to know how to achieve given ends, that is, we want to know the means for these ends, want to know how to cause them. It is therefore no accident at all that both kinds of states are called "knowledge": they are irreducibly different kinds of knowledge, but they are still both equally knowledge. To use a once much-abused phrase: they are separate, but equal.

In what follows, I will now first give a more extensive characterization of practical knowledge and then argue for each of the features just listed, that practical knowledge indeed has it. I will then enter into a discussion of practical knowledge Gettier cases. Some of those who resist Stanley & Williamson's first reduction of skill to knowledge-that have argued that if they were right, there should be Gettier cases for knowledge-how, but that there aren't any (see Stanley 2011: 216ff for discussion and references). And the examples that have been suggested are indeed unconvincing. However, I will show that the deep reason for this lies in the second rather than the first reduction: once we accept that practical knowledge is prescriptive, has world-to-mind direction of fit, and so on, we can construct practical

knowledge Gettier cases that are mirror images of the standard theoretical knowledge Gettier cases. The comparison and analysis of both kinds of cases will also reveal that the deep structural parallels between both kinds of knowledge extend to justification and reasoning. This will support the view that theoretical and practical knowledge are structurally parallel mirror images of one another throughout, and will provide the beginnings of an argument that this is also true for the practical and theoretical domains in general. In this way, the paper will also be a sustained argument to treat the practical as a separate and equal domain and to stop ignoring it or assimilating it to the theoretical in the way the theory bias perennially tempts us to.

## 2. Practical Knowledge as Knowledge-to

We need to get clearer about the scope and nature of practical knowledge. I have already emphasized that by "practical knowledge" I here mean knowledge that is discursive and conceptual. Its primary manifestations are speech acts and thoughts rather than bodily actions. And even when we restrict ourselves to conceptual knowledge, practical knowledge is not the same as know-how, in spite of their customary association. Practical knowledge is both more and less than know-how. It is less than know-how because there are many knowledge ascriptions using this term or its cognates which clearly do not ascribe practical knowledge, for example, when we say that Peter knows how Napoleon lost at Waterloo – a straightforward instance of theoretical knowledge. At the same time, practical knowledge is much more than know-how, because it also includes knowing what to do, and when, where and why to do it. Note that for all these varieties of practical knowledge ascriptions, there are corresponding varieties of ascriptions of theoretical knowledge once we leave out the "to": knowledge of what people do, and when, where and why they do it, is clearly theoretical rather than practical knowledge. Practical knowledge is thus much more adequately glossed as knowledge-to rather than as knowledge-how. It's the "to" that's the mark of the practical in the context of knowledge ascriptions and we will soon see why.

What about Anscombe's paradigm case of practical knowledge, knowledge of what one is currently doing? I believe that part of the reason that Anscombe focussed on this variety of knowledge is that despite her stated opposition to the theory bias with regard to knowledge – which she referred to as the "incorrigibly contemplative modern conception of knowledge" (1957: 57) – she did not completely overcome it. In particular she could not quite break free of the idea that like theoretical knowledge practical knowledge would need to be factive at least in the sense of providing a guarantee that the relevant action comes to pass. Knowledge of what one is currently doing seems to provide at least a better chance of meeting this constraint than practical knowledge that is prior to the relevant action. At the same time one can claim, as Anscombe does, that this knowledge is irreducibly practical because it is (primarily or solely) based on actional rather than perceptual experience. However, since this knowledge is essentially of a still on-going action it cannot really be factive since the action may yet fail to be completed. Part of what makes the action what it is, is what the agent is still about to do. For example, part of what makes what I am currently doing the writing of a paper is how I plan to continue and finish this action. And of course I may not succeed in finishing it for one reason or another.

While this prevents this variety of knowledge from being factive it is at the same time what makes it practical. Practical knowledge is essentially forward-looking, directive and prescriptive, and can therefore only be directed at yet unfinished actions. That is why ascriptions of practical knowledge use non-finite verb forms and why the "to" in particular is characteristic for practical knowledge. And that is also why practical knowledge cannot be knowledge of facts, but must be knowledge of goals, of ends and means. Further evidence in support of the claim that knowledge-to is an irreducibly practical attitude is provided by the

fact that just like verbs ascribing practical knowledge, verbs designating intentions always take the to-infinitive. The same is also true of verbs ascribing other practical postures – speech acts and attitudes – like promises, orders, and obligations. And then there is another class of verbs designating practical attitudes like wanting and desiring, whose clausal complements – when they have clausal complements – take either the to-infinitive or other non-finite verb forms. All this supports the contention that in the ascription of postures the presence of non-finite verb forms in general and the to-infinitive in particular indicates that a practical posture is being attributed.

## 3. Practical Knowledge as Providing Answers to Practical Questions

It is an interesting fact that whereas theoretical knowledge can be ascribed both by means of questions pronouns, indicating that the ascribee knows the answer to a theoretical question, and more directly by means of that-clauses, practical knowledge can only be ascribed in the former, more indirect way. Compare (1) to (2):

(1) a. I know what I did last night.

b. I know what I did last night: I went to see a movie.

c. I know that I went to see a movie last night.

(2) a. I know what to do tonight.

b. I know what to do tonight: I will go see a movie.

b. I know what to do tonight: Let's go see a movie.

c. *I know to see a movie tonight.

(2c) and its equivalents are unacceptable in English, German, French, Italian and probably other Indo-European languages – though there may of course be languages where one can say things like that. In either case, I think we should take this to be primarily a fact about linguistic ascriptions of practical knowledge states rather than about the states so ascribed. There do seem to be states of knowing what to do and how to do it with specific contents, even if these contents cannot be specified in the same breath.

But what are these states exactly? The examples in (2b) reinforce the point that they are irreducibly practical. A practical question is a question about what to do, and the answer to such a question can be provided by an intention – as the expression of which, rather than of a belief "I will go see a movie" is naturally interpreted – or by an order, and so by a prescriptive, self- or other-directive posture with world-to-mind direction of fit. However, at the same time the fact that both these answers can be given points to an apparent further disanalogy between theoretical and practical knowledge and to a difficulty in more precisely determining the nature of the state that bears the status of being practical knowledge.

As was mentioned already, most epistemologists nowadays, even if they have given up on the project of specifying the x in the equation "(theoretical) knowledge = justified true belief + x" apparently made necessary by Gettier cases, still accept that theoretical knowledge is a form of belief. Recently some have begun challenging even that view though on the basis of experimental data showing that in some cases subjects are inclined to ascribe knowledge, but not belief. Whether these data really question the traditional view of theoretical knowledge depends – among other things such as what weight should be assigned to such results – on whether the subjects employ the same notion of belief as in the traditional view. If they do, one possible conclusion would be that theoretical knowledge is not only an irreducible status of beliefs, but an irreducible state in its own right. The corresponding view for practical

knowledge is also possible, but perhaps a more pressing question is what should be the counterpart of the traditional view. What practical state could be the bearer of the status of practical knowledge in the way that belief is the bearer of the status of theoretical knowledge on the traditional view? The difficulty is that if we answer "intention", it's not clear how we should deal with the other-directive case, since it's questionable that we can intend other people's actions. This difficulty may not be insurmountable though. While it does seem odd to say that we intend other people's actions, on the other hand it seems right that when we direct somebody to do something, we do intend to get them to do what we directed them to do. Another possibility is that the underlying attitude might be one of willing rather than intending. Or practical knowledge might turn out to be an irreducible state in its own right after all. I don't think we need to decide this in the present context. The crucial point for now is that, whichever of these alternatives turns out to be the right one, the relevant state is irreducibly practical. However, in what follows I will assume, for ease of reference, that intention can be the state that, under certain conditions to be determined, has the status of being practical knowledge.

But doesn't this difficulty suggest deeper disanalogies between theoretical and practical knowledge and the two domains more broadly? I don't think so, for two reasons. First, I think we need to take into account that because much more attention has been given to the theoretical domain, there has also been much more unifying conceptual work in this area. Notably, a generic notion of belief has been established. If we had a corresponding practical notion, we would probably not feel the difficulties outlined above anymore. Second, the source of the apparent disanalogy is that practical knowledge can be self- as well as other-directive, but should there really be no analogous feature in the theoretical domain? It is tempting to think that the analogous feature is that I can enjoy my theoretical knowledge for myself, for my own consumption as it were, or pass it on to others. However, it seems to me an even better analogy is provided by the fact that my knowledge can be based on my own experience as well as on the testimony of others. The reason this analogy works even better is that it provides a mirror image of the practical case with regard to how other people mediate between the knowledge state and its satisfaction condition. In the testimony case the testimony of the other mediates between my knowledge state and the state of affairs, the fact in the world, that I know of, its satisfaction condition. And the direction of causation in this case is world-to-mind as always for cases of mind-to-world direction of fit. The witness is receptive to the world – perceptually receptive, let us assume, or else it would just be another link in a chain of testimonies which ultimately must bottom out in a perceptual episode – and I am receptive to her and thus succeed in gaining knowledge. In the practical case, the other applies my knowledge and thereby brings about its satisfaction condition if she applies it successfully – or else if she just passes it on provides another link in a chain of practical testimony which ultimately must bottom out in action. So the other's action mediates between my knowledge state and the state of affairs that I knew to be the goal to achieve or how to bring about. And the direction of causation in this chain is mind-to-world as always for cases of world-to-mind direction of fit. My directive caused the action of the other and that action caused my knowledge to be satisfied, that is, to be applied successfully.

As this example already illustrates and as we shall see shortly in the discussion of practical Gettier cases it is essential to take into account the differences in direction of fit and direction of causation in relation to satisfaction conditions when trying to construct practical analogues to theoretical cases (or vice versa). Otherwise one can easily be misled into finding disanalogies between the domains which are really just artefacts of a failure to understand their true differences as well as their commonalities.

## 4. The Satisfaction Condition of Practical Knowledge

I have already taken a stance on what the satisfaction conditions of states of practical knowledge are, but I want to make that stance more explicit and spent some more time discussing and defending its consequences and once again comparing my analysis to those of Stanley & Williamson and Anscombe. The satisfaction conditions of states of practical knowledge are the satisfaction conditions characteristic for forward-looking, prescriptive, world-to-mind-direction of fit states. Intentions and orders are satisfied when they are properly executed or realized; practical knowledge is satisfied when it is applied successfully. The parallel also holds with regard to the causal role of the relevant states or acts. Intentions and orders need to cause the intended or ordered action in order to count as executed and thus as satisfied (Searle 1983); practical knowledge likewise needs to cause the relevant action in order to count as having been applied successfully. Let us compare this with Stanley and Williamson's analysis of know-how. This analysis ascribes to practical knowledge the satisfaction conditions characteristic of theoretical knowledge. The satisfaction condition of the knowledge that some x is a way of riding a bicycle can be an already completed action: having observed somebody bicycling in a certain way, I now know that this way is a way of bicycling. Nor need this satisfaction condition be caused by the knowledge state. It will rather conversely have caused this knowledge state. In other words, the causal relations will be those characteristic for a state with a mind-to-world direction of fit.

What is wrong with Stanley and Williamson's analysis can also be brought out by reflecting on the fact that this analysis could be satisfied by somebody who had the perceptual skill to recognize something as a way of riding a bicycle, while lacking both the actional skill to ride it himself and conceptual practical knowledge on how to do it. It might be objected though that this supposed counterexample does not take into account their additional condition that the relevant state of affairs be represented under a practical mode of representation. But while, as many commentators have pointed out, it remains unclear what precisely Stanley and Williamson mean by a "practical mode of representation", it seems at least safe to say that they think of it on the model of Frege's concept of "Sinn" or sense. That is, they have in mind a property of what is commonly called "propositional content", a special aspect under which the relevant state of affairs is represented. But I think it has been established that this solution won't do. Knowledge-how and practical knowledge more broadly cannot be accounted for in terms of a morning star / evening star kind of difference. This is because the difference between theoretical knowledge and practical knowledge is a matter of psychological mode – or, when we look at their linguistic manifestations, a matter of the force of the corresponding speech acts – rather than of propositional content. This is true no matter whether we think of theoretical knowledge as a form of belief and practical knowledge as a form of intention or willing, or of either or both as irreducible kinds of states in their own right. It is the mode that determines direction of fit and direction of causation and thus the – practical or theoretical – relation between a state and its satisfaction condition. It is, for example, because a state is an intention that it must cause the intended action in order to count as satisfied, not because of what is being intended, not because of the content of the intention, and therefore one cannot satisfactorily account for practical knowledge solely in terms of aspect or sense.

The picture I am proposing has a consequence already alluded to that can seem counterintuitive at first sight: it is possible for states of practical knowledge to fail to be satisfied. I may know what to do, and even where, when, and how to do it, and yet I or others may fail to apply this knowledge successfully. From the point of view of a way of thinking about knowledge that is primarily inspired by theoretical knowledge that may seem intolerable. But again, knowing what to do and how to do it is not knowledge of what is the case. It is not knowledge of facts, but knowledge of goals, of ends and means, and it is in the

nature of such knowledge that these goals may fail to be realized. It is a consequence of the prescriptive, action-guiding, world-to-mind direction of fit nature of practical knowledge that it may fail to be satisfied – through no fault of its own, as it were, but through the failure or absence of action applying it. This last condition is essential though. The mistake must be in the performance or lack thereof, not in the knowledge state itself. Therefore, a counterfactual success guarantee does hold. It must have been possible to apply the knowledge successfully. If I know how to cook Spaghetti Bolognese, this knowledge may not be applied successfully in a given instance, but it must be possible for it to be so applied. Otherwise it cannot be claimed that it was practical knowledge rather than just some idea or plan on what to do. If practical knowledge fails to be applied successfully "the mistake is in the performance, not in the judgment" as Anscombe (1957: 82) puts it, quoting Theophrastus. The knowledge state must still have been well-justified in the sense of being properly based on practical reasoning and / or practical experience. It still must have been a success as far as its purely intellectual credentials are concerned. Incidentally this result also highlights the failure of Stanley and Williamson's first reductionism. Intellectual success is not everything: non-intellectual skills are still required for the successful application of practical knowledge in action and not reducible to it.

## 5. Practical Knowledge as Based on Practical Experience and Practical Reasoning

Practical knowledge is also irreducibly practical in how it is guided by and essentially assessable according to practical criteria of adequacy. For example, if I claim to know how to open the door, I do not just claim the (theoretical) knowledge that if I do x, the door will open. Rather, it will be understood that I claim to know a means that is also acceptable according to practical criteria of adequacy. Smashing the door with an axe will only count as acceptable if it is really urgent to open the door, if there is no less damaging way of opening it, or if the door is not very valuable, we wanted to get rid of it in any case, or smashing it just seemed an outrageously cool thing to do, etc. Likewise, when we say that somebody always knows what to do, we do not merely mean that that person always has some idea or plan, however crazy, on what to do. We mean that these plans are well-justified, that they are based on adequate practical reasoning, that they are good plans according to practical criteria. Finally, when we consider general practical knowledge of the kind we find in recipes and knowledge of various kinds of procedures, for example, technological procedures, such knowledge is based on practical experience. That is, this knowledge specifies means and ends that have been tested for their practical adequacy, that have been found to be worthwhile and appropriate through experience that evaluated them according to practical criteria of adequacy. Such a process can be called one of "practical induction", where particular successful applications of a general rule, procedure, recipe, or plan support or confirm this rule, procedure, recipe, or plan.

## 6. Practical Knowledge and Gettier-cases

If knowing-how is a species of knowing-that as proposed by Stanley & Williamson, one would expect there to be Gettier cases for knowing-how also. Consider a classic example of a Gettier case:

> Bill sees his colleague Fred driving in a Porsche. Believing on this basis that Fred owns a Porsche, Bill then infers that a colleague of his owns a Porsche. Unbeknownst to Bill, his other colleague Hannah owns a Porsche, which she had lent to Fred for the day. So Bill has a justified true belief that a colleague of his owns a Porsche. But intuitively,

> Bill's belief that a colleague of his owns a Porsche is not a case of knowledge. (Stanley 2011: 216)

Stanley & Williamson suggest the following as an example for a knowing-how Gettier case:

> Bob wants to learn how to fly in a flight simulator. He is instructed by Henry. Unknown to Bob, Henry is a malicious imposter who has inserted a randomizing device in the simulator's controls and intends to give all kinds of incorrect advice. Fortunately, by sheer chance the randomizing device causes exactly the same results in the simulator as would have occurred without it, and by incompetence Henry gives exactly the same advice as a proper instructor would have done. Bob passes the course with flying colors. He has still not flown a real plane. Bob has a justified true belief about how to fly. But there is a good sense in which he does not know how to fly. (op. cit.: 435)

However, many commentators have found this example unconvincing. On the basis of the account of practical knowledge developed so far we can now diagnose in which respects it is unconvincing and disanalogous to genuine Gettier cases and construct a case that is actually analogous. This will further support that account.

The crucial point is that the satisfaction condition for claims of practical knowledge is the successful application of that (putative) knowledge in action. The case invented by Stanley & Williamson feels disanalogous to standard Gettier cases because the glitch does not compromise the relation between the knowledge state and its satisfaction condition. It is tempting to think the cases must be analogous because in both cases the aberration affects the way the knowledge state was acquired. But only in the theoretical knowledge case does the aberration also affect the relation between the knowledge state and its putative object or satisfaction conditions. Let us look at this more closely.

In the theoretical knowledge case Bill's false belief that Fred owns a Porsche justifies his belief that a colleague owns a Porsche. This belief is true, but only accidentally so, because it is only true in virtue of Hannah's owning a Porsche, which Bill does not know. So the justification relation does not connect the knowledge state to the state of affairs in virtue of which Bill's belief that a colleague of his owns a Porsche is true. By contrast, in the flight simulator case Bob's relevant mental states are properly connected to their objects, their conditions of satisfaction. There is nothing accidental about the satisfaction of the states. If Bob skilfully applies the instructions he was given, they will be satisfied, that is, executed. Of course, Bob still acquired these states in an accidental way. But since this does not affect the relation between them and their satisfaction conditions, it does not threaten the claim of these states to be knowledge in the same way in which the claim to theoretical knowledge in the Porsche case is undermined. Also note there is a further disanalogy with regard to justification: whereas in the knowledge-that case the justification consists in an (apparently) observation-based belief regarding a particular matter of fact, in the knowledge-how it consists in practical testimony, in advice, which in the story is not clearly specified as being general or particular.

We can thus make sense of and indeed vindicate the difference in intuitions with regard to these cases. The cases are disanalogous because in the theoretical, but not the practical case, the aberration in the way in which the relevant state is acquired also affects the relation to its object, and there is a further disanalogy concerning justification. So what could an analogous case look like? On the basis of our reflections so far, we can describe it in the abstract. The aberration would need to affect the relation between the putative practical knowledge state and its application in action. This state would need to guide action that would turn out to be successful, but in such an aberrant, accidental manner that we would still not want to count it as an instance of practical knowledge. Further, for the analogy to be complete, its justification would need to involve a particular goal in a way analogous to how a particular (putative) fact is involved in the theoretical Gettier case. And such a case can indeed be constructed.

Consider the following case (somewhat artificially) designed to mirror the theoretical Gettier case:

> Michael wants to do something good for his nice colleagues before leaving the department. His means are limited, but he thinks he knows what to do: he will give just one present to one of them, in an exemplary fashion, and anonymously: his old Porsche. But how should he do it? He thinks he knows: he will give it to his colleague Fred by means of parking it in a driveway just around the corner, which he takes to be Fred's, with a note saying "Please accept this car as a gift!" However, unbeknownst to Michael, Fred has moved out of his house recently, but – you guessed it – Fred and Michael's colleague Anna has moved in, so that she ends up getting the car.

So Michael succeeds in giving a present to one of his colleagues, but did he know what to do? I think in this case we will be just as reluctant to ascribe this practical knowledge to him as we are reluctant to ascribe theoretical knowledge in the theoretical Gettier case. The reason is that while what Michael does brings about the satisfaction of his general intention to give a present to one of his colleagues, the particular means he chooses only accidentally leads to the realization of the state of affairs that is the satisfaction condition of his general intention. It is only a coincidence that he brings about this state of affairs in his failed attempt to execute his intention to park his Porsche in Fred's driveway, just like in the theoretical Gettier case it is only a coincidence that Bill's general belief that a colleague owns a Porsche is true, though his particular belief that Fred owns one, on which the general belief is based, is false. That is why we are reluctant to ascribe practical knowledge to him, even though he not only succeeds in doing what he set out to do, but can point to a justification for his putative knowledge, namely that he has a plan on how to reach his goal. That is why his intention, even though it is both satisfied and justified, is not an instance of practical knowledge, just like the beliefs in the standard Gettier cases are not instances of knowledge even though they are also both satisfied and justified.

So in this way we can construct a practical analogue to the theoretical Gettier case and this supports the general hypotheses that practical knowledge is both essentially different from theoretical knowledge in its direction of fit and the way representational success is achieved, and that its domain is a mirror image of that of theoretical knowledge. To further support these hypotheses I will now engage in a more detailed analysis of what is going on these examples with regard to inferential and justificatory relations. In so doing, I will make use of the notions of practical deduction, abduction and induction which I won't be able to fully explain and discuss in the confines of the present paper. But they will still help us to better understand the domain of practical inference and practical justification and how it is a mirror image of theoretical inference and justification.

## 7. Knowledge, Deduction, and Abduction: Theoretical and Practical

To get clearer about the practical and theoretical Gettier cases, let me be maximally explicit about the states involved in both cases and the inferential and justificatory relations between them, beginning with the deductive relations. The singular belief / intention deductively entails the existential, general one. Just like believing that his colleague Fred owns a Porsche logically commits Bill to believing that a colleague of his owns a Porsche, intending to give a Porsche to colleague Fred logically commits Michael to intending to give a Porsche to a colleague.

We further need to include the belief that Fred drove the Porsche and its relation to the belief that he owns it and the plan to park the Porsche on the driveway and its relation to the intention to give it to Fred. I believe that both cases involve *abductive inference*. In the

familiar theoretical case of abduction that Fred owns the Porsche is inferred as the best explanation for the given fact that he drove one. I propose to think of the inference from the goal of giving Fred the Porsche to the means of giving it to him by parking it on the driveway as being likewise an instance of abduction, of practical abduction or inference to the best means.

It seems appropriate to group inference to the best explanation and inference to the best means under the same heading of abduction because they are mirror images of one another and both kinds of non-demonstrative causal reasoning – taking "cause" in a wide sense here. In the theoretical case, a subject asks *why something happened*, looking for a cause that best explains a given effect. In the practical case a subject wonders *how to make something happen*, looking for a cause that is the best means to bring about a given end. What is best is, respectively, determined by theoretical or practical criteria of adequacy: crudely, the best explanation is the one most likely to be true, and the best means the one that maximizes utility. (These slogans here are not meant to suggest that what is best can always or even usually be figured out by using decision theory or related formal systems which presuppose that numerical values can be assigned to likelihood and utility.) So Bill reasons that Fred and thus a colleague owning the Porsche is the best, most likely, explanation of him driving around in it, and Michael concludes that parking the Porsche in the driveway is the best means of giving it to him and thus to a colleague. Bill reasons to the cause of a given effect and Michael to a means for a given end, and if we took the trouble we could easily fill in some of their further beliefs and theories, respectively preferences and plans, that lead them to make the choices they do. Note how our two cases are mirror images of one another in so far as while Michael's intention to give the Porsche to Fred is the starting point of his abductive reasoning and represents an end and thus an effect relative to the means that he is looking for, Bill's belief that Fred owns the Porsche is the end point of his abductive reasoning and represents the cause relative to the effect that is his starting point.

## 8. Knowledge and Justification

Let us now consider the justificatory relations between the relevant attitudes, respectively their objects, the states of affairs, facts and goals, they are directed at, and the reasons for them. (I will leave open here whether reasons themselves are attitudes or state of affairs.) These relations are arguably the crux of the whole matter, because the claim of a state to be knowledge essentially rests on its justification. So we need to get clear about in which sense Michael's intention to give a colleague a Porsche can be justified by his intention to give it to Fred or how that intention in turn can be justified by his intention to do this by means of parking the Porsche on the driveway, and why these justificatory relations should be considered to be the practical analogue of how Bill's belief that a colleague owns a Porsche is justified by his belief that Fred owns one and how that belief in turn is justified by the belief that Fred drove a Porsche. This task is made especially urgent by the fact that we are more accustomed to thinking of ends as justifying means rather than conversely of means justifying ends as I am suggesting here.

Let me hone in on the kind of justification we are interested in by distinguishing it from other possible ways of justifying the general intention. One could elucidate the way in which the intended states of affairs would be desirable or good, one could point to its positive consequences, or one could deductively derive it from a yet more general maxim or rule to the effect that one should always give a Porsche to a colleague when leaving a department, which in turn might be supported by practical induction. But the most fundamental challenge to an intention is to claim that it can't be executed – because if it can't all other reasons for or against it are moot – and so to defend against or pre-empt such a challenge is to support or justify the intention. For example, if someone challenges my intention to climb a Dutch

mountain by asking “A Dutch mountain – which Dutch mountain do you intend to climb?”, I could support my general intention by expressing a particular intention to climb Deneederlandseberg – the artificial mountain some people want to build in Flevoland province. Accordingly Michael can support his general intention to give his Porsche to a colleague by specifying the particular intention by means of which he wants to realize it. Of course generally the more interesting question is whether such a particular intention can be executed in a way that is consistent with other plans and not too costly, and so Michael supports and justifies his particular intention to give the Porsche to Fred (and thus also the general intention that it entails) when he finds a means to give it to Fred anonymously, as he planned to. If no such means could be found, this would be a reason to abandon the intention to give it to Fred, or even, if there were no suitable alternative, to abandon the general intention. So there is not only a sense in which ends justify and provide reasons for means, but means can also justify and provide reasons for ends, and the absence of means can provide reasons against ends.

It is easy to see that there is again an exact parallel in the theoretical domain, where we also find justificatory relations in both directions. The belief that Fred drove the Porsche supports and justifies the belief that Fred owns it which is inferred as the best explanation for it. But that this explanation is available also supports the belief in the (putative) fact that is being explained. Having a good explanation for a putative state of affairs makes it more likely that that state of affairs does indeed obtain and thus also that the corresponding belief is an instance of knowledge. If in our scenario the suggested explanation for Fred driving a Porsche were not available – say because Bill knew that Fred’s finances did not allow him to own a Porsche or that he was strongly opposed to expensive sport cars – this could also undermine Bill’s conviction that it had actually been Fred who was driving the car. It could even force him to abandon his belief if he thought no explanation for it could be possible, but it could of course also make him search for and find other explanations. In any case, the tighter a belief is integrated into a web of justification, of evidence and explanation for it, and the larger that web is, the more likely it is to be an instance of theoretical knowledge.

In the practical case analogously the tighter an intention is integrated into a web of justification, of reasons and plans for realizing it, and the larger that web is, the more likely it is to be an instance of practical knowledge. We are much more likely to accept that somebody knows what to do if there are not only considerations that make his end desirable, but if he also has a plan, a means for realizing it. Otherwise we will merely consider it to be some idea on what to do. So the fact that Michael has a plan for realizing his intention, by means of parking it in a certain driveway, supports and justifies the claim of his intention to be an instance of practical knowledge.

In conclusion let me again emphasize that all these parallels hold in spite of the fact that the theoretical and the practical are diametrically opposed in terms of the direction of fit and the direction of causation. This is evident in the way that the failure of Bill’s belief to be an instance of theoretical knowledge is a mirror image of the failure of Michael’s intention to be an instance of practical knowledge. Michael’s general intention to give a colleague a Porsche is not an instance of knowing what to do even though it is supported through a corresponding singular intention to give the Porsche to Fred, which in turn is justified through a specific plan, a means of accomplishing this, and even though both the general intention and the means intention get satisfied, that is, executed. This is because the singular intention fails to get executed and the means intention only accidentally causes the satisfaction of the general intention. Bill’s situation is parallel, except that the direction of fit and of causation between mind and world are reversed. His general belief that a colleague owns a Porsche fails to constitute knowledge even though it is supported through the singular belief that colleague Fred owns it, which in turn is based on the belief that Fred drove the Porsche, and even though this latter belief and the general belief are satisfied, that is, true. This is because the

singular belief is false and the fact that Hannah owns the Porsche only accidentally causes the general belief that makes it true, by being the cause of Fred's driving the Porsche, which in turn causes Bill to infer the false belief that he owns it.

All of this supports the contention that there is a variety of knowledge, conceptual practical knowledge, which is irreducible to, but structurally parallel with, theoretical knowledge. I presented at least the beginning of an argument that there are practical analogues for all theoretical justificatory and inferential relations, including practical forms of deduction, abduction, and induction. Finally, since this practical knowledge tends to be disregarded in current debates, notably in discussions of know-how, this also supports the diagnosis of a deep-seated theory bias in contemporary philosophy.[1]

**Michael Schmitz**

Universität Wien
Michael.Schmitz@univie.ac.at

## References

Anscombe, E. 1957: *Intention*. Oxford: Blackwell.

Ryle, G. 1949: *The Concept of Mind*. The University of Chicago Press.

Searle, J. R. 1983: *Intentionality*. Cambridge: Cambridge University Press.

Stanley, J. 2011: 'Knowing (How)', *Noûs* 45 (2), 207-238.

Stanley, J. & Williamson, T. 2001: 'Knowing How', *Journal of Philosophy* 98 (8), 411-444.

[1] I would like to thank the audience at my talk in Konstanz at GAP.8 for a good discussion and Sebastian Kletzl for very helpful written comments on a draft. Since this talk took place right before I left the department in Konstanz after more than 10 years I would also like to use this opportunity to thank my colleagues for their support over the years. Special thanks go to Anna Kusser, to whom I dedicate this paper.

# 6. Ästhetik und Religionsphilosophie

# Combining Bayesian Theism with Pascal's Wager

Stamatios Gerogiorgakis

In the light of the newest criticisms against it, Bayesian Theism appears to be an endangered project. In order to give the Bayesian analysis of religious faith a chance independently of the question whether these criticisms are sound or not, I propose a Bayesian reading of Pascal's Wager which is exclusively based on subjective probabilities. I argue that following religious maxims for reasons which are independent from religious ethics is inconsistent. I argue, further, that one's following religious maxims only makes sense only under the condition that one bets on God's existence, i.e. has a strong faith.

## 1. Motivational Remarks

Swinburne's Bayesian argument for the existence of God consists in calculating by Bayes's theorem the likeliness of God's existence on the total available evidence, once we assume: 1) some plausible value for the likeliness of the fact that would be that the evidence is made available to us by God and 2) the ratio of the intrinsic probability of the existence of God alone to the intrinsic probability of the evidence (without assuming that God gave us the evidence, that is). I.e. (for h = the hypothesis that God exists; k = tautological evidence, e.g. laws of logic and physics; e & k = the total evidence for the existence of God):

(BT)→P (h | e & k) = P (e | h & k) · P (h | k) / P (e | k).

Richard Swinburne (2004) argues that the value for P (h | e & k), i.e. the likeliness of the hypothesis that God exists given the total evidence, approximates, nevertheless does not reach the value 1. In other words, God is very likely to exist. The basis for Swinburne's calculation is set by 1) the assumption by the principle of indifference (see below) that it is as likely for God to provide us with some evidence for His existence as it is likely that He would not provide us with this evidence (i.e. P (e | h & k) = ½) and 2) the assumption of a value approximating but not reaching 2 for the ratio P (h | k) / P (e | k).

As one easily sees, Swinburne's result depends largely on probabilities which are not calculated but assumed; in other words: on prior probabilities.[1] The principle of indifference allows for example for assigning *a priori* equal probability values to alternative events, for whose statistic frequency we are completely clueless.

Jeremy Gwiazda objected that the hypothesis that God exists ('h') given the necessities of nature and thought ('k') is very simple, whereas our available evidence ('e') for the existence of God given the same necessities is much more complex. The principle of simplicity allows to assign simple hypotheses which are explanatory of certain events, higher prior probabilities than more complex hypotheses which are explanatory of the same events. Gwiazda (2010, 360) feels justified to conclude from this that the ratio P (h | k) / P(e | k) must result in a

[1] By contrast to the largest part of the bibliography concerning the topic objective vs. subjective Bayesianism which uses the term "prior probability" to denote the assumed, non-calculated probability, Swinburne (2004), 67, uses "prior" and "intrinsic" probability interchangeably and offers no specific term for the assumed, non-calculated probability. This is a pity since assumed, non-calculated probabilities play a great role in his analysis. By the term "prior probabilities" I am using here the linguistic convention of the objective-vs.-subjective-Bayesianism-bibliography, not Swinburne's usage.

number which rather exceeds the value 4. And due to the principle of indifference Swinburne gives the value ½ for P (e | h & k) (≈ given that there is a God, it is as probable that we would possess our available evidence for His existence, as that we would not possess it). But then the value for the probability of God's existence given the evidence is absurdly high: P (h | e & k) = P (e | h & k) • P (h | k) / P (e | k) ≥ 2, which leads Bayesian Theism as a whole *ad absurdum*.

To fix the calculation, i.e. to avoid the absurdity of the likeliness of God's existence given the total evidence being assigned a value higher than one (i.e. P (h | e & k) ≥ 1), Gwiazda considers the option of abandoning the value ½ for P (e | h & k). So, if, by Gwiazda's calculations, P (h | k) / P (e | k) ≥ 4 then, in order for the likeliness of God's existence given the total evidence to approximate the value 1, the probability for our evidence for God's existence to be available given God's existence should be less than 25 per cent (i.e. P (e | h & k) < 0,25). This means that God would rather avoid to let us possess our available evidence for His existence. For one thing, the justification for such an assumption would be very debatable and suspect of being *ad hoc*. Why would God do something like this? For another, the aforementioned justification would not be based on any generally accepted principle of inductive logic. But what is worse is that an *ad hoc* assumption of an at-most-25-per-cent-probability by which God would let us have some evidence for His existence, would substitute Swinburne's initial one-half-probability for the same state of affairs, a probability which was at least based on the principle of indifference. Since we feel justified to assume *ad hoc* that it would be rather unlikely (less than 25 per cent) for God to let us possess some evidence for His existence, there is not any particular reason why the atheist should not expect this likeliness or rather unlikeliness to be extremely low, say P (e | h & k) ≤ 0,00000000025, to make P (h | e & k) very low as well. If you do not care if your assumptions are *ad hoc*, you cannot very reliably argue that your opponent should care!

Swinburne (2011) has argued against this criticism that the principle of simplicity is to be applied in comparing between explanatory hypotheses, not however between explanatory hypotheses and events. This means that P (h | k), despite its pertaining to a much simpler assumption, does not have to be much higher than P (e | k). Swinburne insists in the value ½ for P (e | h & k), and gives the value 0,01/0,005 (= 2) for the ratio: P (h | k) / P (e | k), which results in the value 1 for P (h | e & k).

Swinburne's counter-argument against Gwiazda leaves much room for criticism. It presupposes a selective application of the principle of simplicity which can be easily considered to be *ad hoc*. Moreover, Swinburne's insistence in very high values for P (h | e & k) fails to account for those facets of religious faith, which maintain that it would be rational to affirm God's existence independently of how probable or improbable His existence is. Pascal's Wager exemplifies this kind of religious faith. The history of religion has many examples of religious traditions to offer, which support Pascal's line of argument. I propose to affirm a Bayesian Theism which supports Pascal's line of argument. Among other things, Pascal's line of argument can be analysed in terms of Bayesian probabilities and shown to be immune against Gwiazda's criticism (cf. section 4 and 5).

## 2. Gambling Odds in Bayes's Theorem

Gambling odds express the payouts from a bet on the occurrence of an event. The higher the likeliness of the event, the lower are the gambling odds for its occurrence. The lower the likeliness of the event, the higher are the gambling odds. For example, ideally, when an event is as likely as it is unlikely to happen, i.e. P (e) = 0,5, then the gambling odds for this event amount to 1/0,5 = 2. I.e. if one bets that the event occurs and this happens, then one doubles her stake.

A very interesting feature of sports bet is that bookmakers, in order to attract gamblers to join the game, normally try to justify their gambling odds with some grade of likeliness which would seem the customers plausible and make them bet. By doing so, bookmakers take into consideration subjective expectancies based on the known condition and skills of teams, which, normally, are not based on frequencies.

If the gambling odds for the game Rot-Weiß Erfurt of the German third league vs FC Barcelona is 100 for a victory of the Erfurt team or a deuce (i.e. those who bet on Erfurt's victory over Barcelona or deuce will take hundred times their stake if Erfurt really wins or succeeds to keep the deuce) and 1,01 for Barcelona's victory, then the assumed likeliness which underlies the gambling odds for Erfurt's victory over Barcelona or a deuce will be: 1 / 100 = 0,01. The corresponding figure for Barcelona's victory over Erfurt will be the rest 0,99 (notice that the bookmaker offers 1 / 1,01 ≈ 0,99), which entitles for a 0,01 € gain for every € bet on Barcelona's victory if Barcelona wins.[2] The value 0,99 given for Barcelona's victory is, I repeat, anything but a frequency which resulted by use of empirical methods. It is rather a likeliness based on a guess with which the bookmaker hopes to meet the subjective estimations of as many potential customers as possible to make them feel justified to come and bet. For example, the gambling odds would take for granted that Erfurt's victory and a deuce would be extremely improbable and would allow for a very high revenue for bets on these outcomes, just because the public, the potential customers that is, would be intuitively inclined to risk some low stake on Erfurt's victory over Barcelona only if the revenue would be extremely high. Generally, the gambling odds which bookmakers issue for the outcome of soccer, basketball, baseball or rugby matches are normally not based on frequencies but are supposed to pre-empt the public's guesses of the degree of likeliness of these outcomes. Although often there is a general agreement on these guesses, they are estimations which reflect an utterly subjective stance.

The main reason for which the likeliness of events which underlies gambling odds reflects a subjectivist stance is the following: neither bookmakers nor anyone else have the means to calculate the exact degree of probability which corresponds to the gambling odds for these outcomes. If FC Barcelona, which is supposed to be one of the strongest soccer teams worldwide when these lines are written, would happen to play against Rot-Weiß Erfurt there would be no prehistory of games between the two teams to result in certain frequencies for Barcelona's victory, Erfurt's victory or a deuce. Therefore no gambling odds would be issued if the gambling odds would be based on frequencies.

Let me dwell for a moment on the motivation of RWE-supporters to bet on their team's victory over Barcelona.

Perhaps there are a few supporters of RWE with a distorted relation with reality who really believe that RWE has many chances to overcome in the match against Barcelona. Let us call them obtuse RWE-supporters. They would bet on behalf of their team in order to earn money. But most RWE-supporters are sober: their expectation to earn money for their bet on RWE's victory over Barcelona is fairly low. What they aim at with this bet is to demonstrate their identification with the team by "sacrificing" the money of the stake. The risk to lose their money is very high. However, they would bet small amounts which do not weigh much. For the small amount they enjoy, apart of the feeling of identification with the team, the anticipation of the – admittedly extremely improbable – case of a victory for Erfurt, which gives an extraordinary revenue. Their stake has a double function then: they buy something for it and it gives them the hope of earning the high revenue in case their team wins. But this is *a win-win-situation* which would make sober RWE-supporters (people that is, who

[2] The fictive example results in a sum of gambling odds which is above 1. As irritating as they can be, sums over 1 depict a usual tactic of bookmakers which cannot be the subject of this paper.

sympathize with RWE but acknowledge the very poor chances of their team to resist the international superstars of the Catalan team) bet anyway. I resume:

(A) In the very probable case that they lose their money, sober RWE-supporters will have enjoyed demonstrating their selfless identification with the team; which means that they do buy something for their stake even if they lose the bet.

(B) In the very improbable case that the Erfurt team wins Barcelona, the revenue is high.

Let me add to this informal presentation of the idea that sober supporters of RWE would bet that RWE beats Barcelona, a Bayesian argument for the same claim. The symbols are defined thus: e = a posteriori knowledge comprising among others Barcelona's usual abilities and RWE's constant decline since the reunification of Germany; h = the hypothesis that RWE will beat Barcelona; k = background knowledge about soccer and gambling containing the facts (A) and (B) above). P (h | k) expresses the likeliness of the fact which would be that sober RWE-supporters would (rationally) behave as if they expected their team to beat Barcelona. But this is more likely than unlikely to happen at least because of the fact (A) and, of course, because of the fact (B). I.e. P (h | k) ≥ 0,5. In (BT) the subformula P (e | h & k) says that FC Barcelona is expected by the sober RWE-supporter to be much stronger than Rot-Weiß Erfurt even if it is sensationally beaten by RWE. Since it is extremely unlikely that the expectation of the sober RWE-supporter that his team would win would change the perception of the strength of the two teams, one can take P (e | h & k) to have a value very close to P (e | k). I.e. P (e | h & k) / P (e | k) ≈ 1. But this, again, makes P (h | e & k) ≈ P (h | k) ≥ 0,5. QED.

## 3. Pascal's Wager

Giving an account for one's belief that a hypothesis is likely to be true is sometimes like describing the behaviour of fair dice in terms of the theory of probabilities. Swinburne's inductive, Bayesian argument on the existence of God is such an account. When the behaviour of fair dice is described in terms of the theory of probabilities an *a priori* assumption concerning what is a fair die is presupposed: any die which does not allow in the long run for the frequency of 1/6 for every number to appear is said *a priori* to be unfair. Like in the dice case, any state of affairs in which general principles (e.g. the principle of simplicity and the principle of ignorance) which express an underlying general homogeneity in terms of prior probabilities for certain events, would not apply, would be *a priori* inappropriate to calculate God's existence in Swinburne's sense. In throwing dice and in demonstrating God's existence à la Swinburne, *a priori* probabilities play a great role.

But unlike Swinburne's program and throwing dice, there are calculations of the likeliness of a hypothesis in which *a priori* probabilities do not play a role. Giving an account for Pascal's belief in the existence of God is more like analysing gambling odds for soccer matches than describing the behaviour of fair dice. Pascal's Wager is a betting on God's existence which resembles closely the previous example of betting on the outcome of the game Rot-Weiß Erfurt vs FC Barcelona. Like the sober RWE-supporter, Pascal thinks it rational to bet on God's existence even if the objective probability of winning should be admitted to be low. Pascal justifies this by showing that Pascal's Wager introduces a *win-win-situation* for bettors of a certain kind:

> Let us weigh the gain and the loss in wagering that God is... If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is. ... [Y]ou would be imprudent, when you are forced to play [i.e. you will play anyway – comment and italics by me, S.G], not to chance your life to gain three at a game where there is an

> equal risk of loss and gain. But there is an eternity of life and happiness. (Pascal 1670: §233)

Pascal's certainty that in betting on God's existence there is nothing to lose and an extraordinary high revenue to earn is based on a hidden assumption which closely resembles the basic attitude of the sober RWE-supporter towards the match of his quite mediocre team against the international powerhouse FC Barcelona: if he loses, he loses nothing due to motive (A) above, not to mention what he could earn if Erfurt scored first at some point of the match and managed to defend its goalposts for the rest of it – cf. motive (B).

By analogy, Pascal would have written some things in vain in case that God does not exist. But he would not have written them without having got some joy from writing them. Additionally to this not inconsiderable joy, Pascal had an expectation of oceans of joy in case that God exists to reward those who believe in Him.

Of course, it needs someone like an RWE-supporter to fail regretting a lost bet on the victory of Erfurt over Barcelona. Many others would feel unpleasant after a lost bet like this ("Stupid me! It said "Erfurt", not "Frankfurt"!). Likewise, it needs a Pascal to fail regretting a lost bet on the existence of God. Others, let us say Richard Dawkins, would very much regret to have worshipped God in their young years and would consider this a waste of time if nothing worse. Pascal's argument does not show that *everyone* should consider the bet for the hypothesis that God exists a win-win-situation. It rather shows that it is irrational not to bet on the existence of God if you *already (and gladly) conform* with what God is supposed to demand you to do.

## 4. A Bayesian Wager

In Mark 9:24 the father of a young epileptic who asks Jesus to cure his child from epilepsy cries for help. However it is not only help for the boy's health after which the father seeks. After being informed by Jesus (Mark 9:23) that the faith of the saved is the prime cause of salvation and estimating his own faith as low, the father of the young epileptic in Mark 9:24 seeks *help for his unbelief* also. Curiously enough, the father does nothing wrong to fear of being characterized as unfaithful. He asks Jesus for help and, in fact, he claims to believe in Jesus and in the cure of his boy. Nevertheless, at the same time he asks Jesus to help his unbelief. I think that this can be seen as a clear case of cognitive dissonance – and of one which is not very uncommon.

Some of the readers will have had acquaintances with persons similar to the father of Mark 9:24. People who *behave* like devote Christians but have to admit that their faith is weak: people who are good in the moral sense and vegan in terms of diet (fasting is very essential in the Roman-Catholic as well as in the Orthodox Church), unselfish and caring, monogamous and never frivolous but have to admit to be sceptics if they isolate the moral issues from their intellectual frame of mind.

The case of the father in Mark 9:24 and the moral vegan sceptic have many similarities with the sober supporter of Rot-Weiß Erfurt, of whom I have already spoken. Like the sober RWE-supporter would enjoy the victory of his team against Barcelona although he has to admit that Barcelona is the stronger team, the father in Mark 9:24 and the moral vegan sceptic would enjoy to see the theistic arguments victorious in their struggle against the superior scientific mainstream. In a way very similar to the sober RWE-supporter's way, the moral vegan sceptic makes the same sacrifices which religious persons would make. But these virtues are only because of her *moral* intuitions. By contrast to the religious, the moral vegan sceptic is *intellectually* not certain of God's existence neither is she certain whether worshipping God is a good idea.

In the lines to follow I shall show that a moral vegan sceptic would be irrational not to maintain to have a strong faith as a result of being moral and vegan. Drawing this conclusion is dependent on a combination of Bayesian Theism with Pascal's Wager.

In a way very similar to the RWE-supporter, by making a sacrifice in terms of diet or life style, the moral vegan sceptic engages in a make-believe game with a high revenue. Like the bookmaker offers the RWE-supporter one hundred times his stake, Christians believe that the revenue for a justified life is eternal life.

In order to explore what it would be consistent for the moral vegan sceptic and the father in Mark 9:24 to believe, I shall formulate the likeliness for the existence of God in terms of Pascal's Wager as P (h | e & g) whereby 'g' expresses the overall expectations of the religious, including natural necessities, laws of thought, i.e. logic, plus *the likeliness which corresponds to the gambling odds for believing in God*. The moral vegan sceptic would assume that faith in God returns a high revenue due to the low intrinsic likeliness of the existence of God. Let her assume that the existence of God if she bets on the existence of God pays 20 to 40 times more than if she bets on the existence of God in vain. I.e. $20 \leq P(h \mid g) / P(e \mid g) \leq 40$ because the intrinsic likeliness of the existence of God without taking the betting motivation and the gambling odds into consideration is $0{,}05 \geq y \geq 0{,}025$. Recall that Gwiazda assumed the value of the ratio P (h | g) / P (e | g) as higher than 4 to lead Bayesian Theism *ad absurdum* and to receive by Swinburne only a debatable objection. Let us see however what happens if we assume the ratio to have a value between twenty and forty in the context of my Bayesian Wager. For P (e | h & g) = 0,025 we get by Bayes's Theorem:

P (e | h & g) • P (h | g) / P (e | g) = P (h | e & g); consequently $0{,}025 \cdot 20 \leq P(h \mid e \& g) \leq 0{,}025 \cdot 40$; therefore: $0{,}5 \leq P(h \mid e \& g) \leq 1$. This shows that the father in Mark 9:24 and similar cases, e.g. the moral vegan sceptic, should consider the existence of God as a hypothesis which is fairly likely to hold on the total evidence – whereby the fact that the father in Mark 9:24 and the moral vegan sceptic cannot help themselves from behaving like if they were religious, is included in the evidence.

## 5. Conclusions

I conclude that for those who engage in some elements of the religious behaviour for reasons independent of religiosity, say for the father in Mark 9:24 or a moral vegan sceptic, it is rational to believe in the existence of God (i.e. $0{,}5 \leq P(h \mid e \& g) \leq 1$) already if they expect the world to be created by God by a certainty which amounts at 2 point five per cent (i.e. P (e | h & g) = 0,025 under the condition that they suspect the revenue from faith in God to be between 20 and 40 times their investment (i.e. $20 \leq P(h \mid g) / P(e \mid g) \leq 40$). That is, at the end of the day the father in Mark 9:24 and the moral vegan sceptic will have to realize that their motivation to bet on the existence of God presupposed a likeliness of God's existence on the total evidence much higher (from above 50 to somewhat below 100 per cent) than the intrinsic likeliness of the existence of God which they postulated without considering the betting situation (from two-and-a-half to five per cent).

As one sees, Bayesian Theism is a useful instrument to justify the rationality of a strong instead of a weak faith, even if Gwiazda's and Schurz's arguments against Swinburne's version of Bayesian Theism are supposed to be true. Gwiazda's and Schurz's arguments show at most that Bayesian Theism does not offer a justification to convert the unfaithful.

Converting the unfaithful is not the topic of the Bayesian Wager, like it was not the topic of Pascal's original version of the Wager. Instead of trying to convert atheists the Bayesian Wager only shows that, if you are already sympathetic towards religion and follow for independent reasons the rules which a devote religious person would also follow, nevertheless you have to admit that your faith is weak, then it makes no sense to continue admitting that

your faith is weak, since betting on God's existence is a win-win-situation from your subjective stance.

Technically speaking, the subjective stance is introduced in the calculation with the use of the factor 'g' instead of 'k'. This point can be utilized to formulate a counter-argument to the effect that the conditions e and g, i.e. the ones which make the conditional probability P (h | e & g) high, do not *genuinely* explain h. But this accusation has been repeatedly launched also against Swinburne's original argument (Schurz, 2009; 2011). Consequently, my Bayesian Wager would not worsen the reputation of Bayesian Theism. Further, the Bayesian Wager adopts Gwiazda's objections to Swinburne's calculation of the ratio P (h | g) / P (e | g) without leading to absurd results.

**Stamatios Gerogiorgakis**

University of Erfurt
stamatios.gerogiorgakis@uni-erfurt.de

## References

Gwiazda, J., 2010, "Richard Swinburne, *The Existence of God*, and Exact Numerical Values", *Philosophia* 38, 357-63.

Pascal, B., 1670, *Pensées*, transl. by W. F. Trotter, Mineola N.Y.: Dover Publications, 2003.

Schurz, G., 2009, "Kreationismus, Bayesianismus und das Abgrenzungsproblem", paper read at the 7th triennial conference (GAP.7) of the German Society of Analytical Philosophy (Gesellschaft für analytische Philosophie), Bremen 14-17/9/2009. Published online at: http://duepublico.uni-duisburg-essen.de/servlets/DerivateServlet/Derivate-29983/Proceeding_GAP7_Nachdenken_Vordenken.pdf, 169-182. Retrieved on 2/1/2013.

Schurz, G., 2011, "Bayesianische Bestätigung des Irrationalen? Zum Problem der genuinen Bestätigung ", paper read at the 22nd German Congress of Philosophy, Munich, 11-15/9/11. Published online at: http://epub.ub.uni-muenchen.de/12354/. Retrieved on 2/1/2013.

Swinburne, R., 2004², *The Existence of God*, Oxford: Clarendon Press.

Swinburne, R., 2011, "Gwiazda on the Bayesian Argument for God", *Philosophia* 39, 393–6.

# Zur Rechtfertigung religiöser Überzeugungen durch pragmatische Argumente

Christoph Kurt Mocker

In der jüngeren Diskussion über „Pascals Wette“ besteht ein weitgehender Konsens darin, dass ein schlüssiges pragmatisches Argument für den Glauben jedenfalls nicht unmittelbar auf den Glauben selbst gerichtet sein kann, sondern, wenn überhaupt, nur auf unsere Praxis, sofern sie den Glauben fördert. Hier soll gezeigt werden, warum im Gegenteil pragmatische Argumente *nicht* auf eine glaubensbildende Praxis beschränkt werden können, wenn sie gültig sein sollen. Die in diesem Aufsatz entwickelte *reductio ad absurdum* dieser Praxis-Beschränkung und ihre anschließende Auflösung ergeben: Wenn der Nutzen des Glaubens für die Begründung einer religiösen *Praxis* relevant ist, dann ist er auch *doxastisch* relevant, insofern nämlich, als dann kein entscheidender Grund für ein atheistisches oder agnostisches Urteil spricht. Dieses Ergebnis ist nicht etwa bloß ein Appendix zur vorherrschenden „praktischen“ Interpretation von Pascals Wette, sondern weist über diese hinaus: auf eine alternative Form pragmatischen Arguments, zur Rechtfertigung einer bereits *bestehenden* Glaubensüberzeugung.

„Was dürfen wir glauben“, diese Frage hat eine besondere Brisanz bis heute in der Religionsphilosophie. Die prominente Replik von Blaise Pascal lautet: Wir *dürfen* nicht bloß glauben, dass Gott existiert, sondern wir *sollen* es sogar. In seiner berühmten „Wette“ hat er dies bekanntlich pragmatisch zu begründen versucht: Der (mögliche) Nutzen des theistischen Glaubens ist ein entscheidender, pragmatischer Grund dafür, zu glauben.

Die jüngere religionsphilosophische Diskussion um Pascals „Wette“[1] hat u.a. gezeigt, dass es eine Vielfalt möglicher pragmatischer Argumente für den Glauben zu unterscheiden gilt, die in Pascals eigenem, dichtem Text enthalten oder zumindest angelegt sind. Was die *Art* des Nutzens betrifft, so muss er nicht eudämonistisch, sondern kann auch moralisch sein; und der eudämonistische Nutzen selbst muss kein transzendenter sein (göttliche Lohnung), sondern kann sich bereits immanent im irdischen Leben verwirklichen (Kontingenzbewältigung, Tröstung usw.).[2]

Egal welche Art von Nutzen man ansetzt,[3] so besteht aber doch in einem Punkt weitgehender Konsens (den auch Pascal schon berücksichtigte): Wenn es ein schlüssiges pragmatisches Argument geben soll, dann kann es sich eigentlich nicht unmittelbar auf den (theistischen) Glauben selbst richten: Dass wir *urteilen* sollen, „Gott existiert“; sondern höchstens mittelbar auf diesen Glauben, indem es begründet, dass wir auf eine Weise *handeln* sollen, welche die Glaubensbildung (längerfristig) fördert.[4]

Hier soll gezeigt werden, warum im Gegenteil pragmatische Argumente *nicht* auf eine glaubensbildende Praxis beschränkt werden können. Die in diesem Aufsatz entwickelte

[1] Wichtige Aufsätze zum Thema versammelt Jordan 1994. Eine ausführliche Monographie ist Jordan 2006. Im deutschsprachigen Raum referiert und kommentiert Weidemann 2007, Kapitel 4, Abschnitt 1, ausführlich die jüngere Diskussion.

[2] Vgl. dazu Jordan 2006 und Weidemann 2007, Kapitel 4.

[3] Neben der Art des Nutzens spielt für die Form und Schlüssigkeit eines pragmatischen Arguments die *Wahrscheinlichkeit* dieses Nutzens bzw. einzelner Nutzenwerte die entscheidende Rolle (vgl. Löffler 1996 & 1999).

[4] Zur Begründung dieser These im nächsten Abschnitt mehr.

*reductio ad absurdum* dieser Praxis-Beschränkung und ihre anschließende Auflösung ergeben: Wenn der Nutzen des Glaubens für die Begründung einer religiösen *Praxis* relevant ist, dann ist er auch *doxastisch* relevant, insofern nämlich, als dann kein entscheidender Grund für ein atheistisches oder agnostisches Urteil spricht.

Dieses Ergebnis ist nicht etwa bloß ein Appendix zur vorherrschenden „praktischen" Interpretation von Pascals Wette, wie am Ende dieses Aufsatzes kurz skizziert wird, sondern weist über diese hinaus: auf eine alternative Form pragmatischen Arguments, zur Rechtfertigung einer bereits bestehenden Glaubensüberzeugung, anstatt zur Begründung einer noch nicht vorhandenen religiösen Überzeugung. Damit erhalten pragmatische Argumente eine weit realistischere normative Funktion: nicht als letztes, verzweifeltes Mittel, um den Nicht-Gläubigen auf die eigene Seite zu ziehen, sondern zur Verteidigung, dass ein Gläubiger rechtmäßig glauben darf.

## 2. Das Argument für die Beschränkung pragmatischer Argumente auf die Praxis

Das Argument für die Beschränkung des argumentativen Werts des Glaubens-Nutzens auf eine glaubensförderliche Praxis ergibt sich aus folgender (plausiblen) Annahme über die (logischen oder psychologischen) Grenzen unserer Überzeugungsbildung: Nach dem Nutzen des theistischen Glaubens können wir zwar unser Handeln ausrichten, nicht aber unser Urteil. Folglich kann die Aussicht auf einen Nutzen unsere Überzeugungen höchstens mittelbar, nicht aber unmittelbar verändern.[5] Diese explikative Kraft des Nutzens ist aber eine notwendige Voraussetzung seiner normativen Kraft:[6] Kann der Nutzen einer Überzeugung lediglich zur Erklärung unseres Handelns, nicht auch unseres Urteilens beitragen, dann liefert der Nutzen auch keinen guten/normativen Urteilsgrund, sondern lediglich einen guten/normativen Handlungsgrund.

## 3. Der Widerspruch dieser Beschränkung

Allerdings scheint von hier aus nun folgender normative Widerspruch zwischen praktischer und theoretischer Normativität zu drohen.

**a) Praktische/Handlungsnormativität:**

Existiert ein schlüssiges praktisches Argument für den Glauben, dann gibt es einen Handlungsgrund, der einen Mangel an Evidenzen für oder stärkere Evidenzen gegen den Glauben[7] zugunsten des Glaubens normativ überwiegt und dadurch eine normative Begründung liefert für eine Praxis der religiösen Überzeugungsbildung; dafür, dass wir uns praktisch die Ausbildung einer Glaubensüberzeugung zum Zweck setzen sollen.

---

[5] Das sind die Thesen des *direkten* doxastischen *Involuntarismus* und des *indirekten* doxastischen *Voluntarismus*. Vgl. dazu Jordan 2006, Kapitel 2, Pojman 1986, Kapitel 6 und Weidemann 2007, Kapitel 4.

[6] Dieses Normativitätsprinzip ist eine Variante des bekannten Prinzips "Sollen impliziert Können".

[7] Pascal selbst geht in seiner "Wette" zunächst davon aus, dass es keine überwiegenden Evidenzen für oder gegen die Existenz Gottes gibt. Er selbst betont aber später die Gültigkeit seines Arguments auch für die Voraussetzung eines (beliebigen) evidentiellen Übergewichts zugunsten der *Nicht*-Existenz Gottes. - Hier kommt es darauf an, dass sowohl ein Mangel an Evidenzen als auch ein Übergewicht an Evidenzen gegen Glauben plausiblerweise einen normativen Grund *gegen* den Glauben darstellt (Das ist eine unkontroverse, weil schwache "evidentialistische" Annahme über die normative Relevanz von Evidenzen), denn natürlich ist nur unter dieser Voraussetzung überhaupt ein pragmatisches Argument für den Glauben erforderlich.

**b) Theoretische/Urteilsnormativität:**

Dass der Mangel an Evidenzen für bzw. das Übergewicht von Evidenzen gegen den Glauben *qua* Handlungsgrund (gegen eine religiöse Praxis) durch pragmatische Handlungsgründe zugunsten einer religiösen Praxis aufgewogen wird, impliziert nicht, dass der Mangel an Evidenzen bzw. das Übergewicht von Evidenzen gegen den Glauben *aufhört*, ein *theoretischer* Grund gegen ein theistisches Urteil und für ein agnostisches bzw. atheistisches Urteil zu sein. Und da ihnen in diesem Kontext, bezogen auf das Urteil, keine pragmatischen Gründe entgegenstehen, liefert der Mangel an Evidenzen für bzw. das Übergewicht von Evidenzen gegen den Glauben eine normative Begründung für ein agnostisches bzw. atheistisches Urteil.

Das ist ein normativer Widerspruch. Ein normativer Widerspruch kann aus mehreren Gründen auftreten. Erstens, weil etwas, z.B. eine Handlung, *widersprüchlich normiert* ist: $O(H) \wedge \neg O(H)$. Oder zweitens deshalb, weil etwas *Widersprüchliches normiert* ist, z.B. eine Handlung und ihre Unterlassung: $O(H) \wedge O(\neg H)$. Drittens, und das ist hier der Fall, weil zweierlei Optionen[8] geboten sind, die widersprüchliche Ziele/Zwecke haben. Das sind im vorliegenden Fall die Praxis der Überzeugungsbildung mit dem Zweck des Glaubens ($Z(G(p))$) auf der einen und ein agnostisches bzw. atheistisches Urteil mit dem Ziel[9] eines agnostischen bzw. atheistischen Überzeugungszustands auf der anderen Seite $Z(G(\neg p))$ bzw. $Z_j(G(\neg p))$: $O(Z(G(p))) \wedge O(Z(\neg G(p)))$ bzw. $O(Z(G(p))) \wedge O(Z(G(\neg p)))$.

## 4. Die Auflösung dieses Widerspruchs

Dieser Widerspruch löst sich auf, wenn man die *mehrteilige* Argumentationsstruktur beachtet, die pragmatische Argumente bereits in der Pascalschen Analyse zeigen, und im zweiten Schritt ein plausibles normatives Prinzip einführt.

**1. Schritt: konsequentialistische Analyse.**

Der erste Schritt ist eine konsequentialistische, „entscheidungstheoretische" Analyse der Überzeugungsalternativen sowie der Optionen, welche auf diese Glaubenszustände abzielen: Welche Option hat den größten Erwartungswert?[10] Dazu müssen sowohl der Wert des möglichen Erkenntnis-Erfolgs bzw. -Misserfolgs als auch der Nutzen oder Schaden der Glaubensalternativen sowie die Wahrscheinlichkeiten für diese Ergebnisse quantitativ „kalkuliert" werden. Ergibt diese (quantitative) konsequentialistischen Analyse einen insgesamt überwiegenden pragmatischen Nutzwert, dann handelt es sich bei der der entsprechenden Glaubensoption um die insgesamt beste Alternative (für eine Person), trotz fehlender Evidenzen für oder überwiegender Evidenzen gegen den Glauben.

**2. Schritt: Prinzipien der gründe- und normativitätstheoretischen Applikation.**

Erst ein *zweiter* Schritt ist dann die *normative* Applikation der konseqentialistischen Bewertung, für die folgende zwei Prinzipien intuitiv plausibel erscheinen:

---

[8] Unter einer „Option" soll hier einfach dasjenige verstanden werden, das ein möglicher Gegenstand guter/normativer Gründe ist. Handlungen und Urteile sind Optionen in diesem Sinne.

[9] Unter einem „Urteil" wird hier einfach der Gegenstand einer Begründung durch Evidenzen verstanden, dessen intendiertes und unmittelbares Resultat eine Überzeugung ist (das Intendiertsein eines Urteilsresultats legitimiert die Rede von einem „Ziel" des Urteils). Unter diesen Begriff soll hier, der Einfachheit halber, aber auch eine Urteils*enthaltung* fallen, deren Grund gerade das *Fehlen* von Evidenzen ist, so wenn von einem „agnostischen Urteil" die Rede ist.

[10] Pascal ist bekanntlich zu einem Ahnherren der Entscheidungstheorie geworden, die heute wiederum eine präzise entscheidungstheoretische Rekonstruktion seines Wett-Arguments ermöglicht (vgl. dazu z.B. Löffler 1996 & 1999 sowie Hacking 1994).

I) Man soll die insgesamt bessere/beste Option (mit dem größeren/größten Erwartungswert) realisieren.

I') Die normativen Gründe für die bessere/beste Option sind entscheidende Gründe dafür, dass man diese Option realisieren soll.[11]

Diese beiden Prinzipien stehen dann unter dem bereits diskutierten Vorbehalt, nach den beiden Prinzipien:

II) Man soll nur das (als Option realisieren), was man auch kann.

II') Guten Gründen dafür, dass man etwas (eine Option realisieren) soll, muss man auch folgen können.

Nach diesen Prinzipien löst sich der obige Widerspruch zwischen praktischer und theoretischer Normativität wie folgt elegant auf:

**a) Praktische/Handlungsnormativität:**

i) Der Nutzen des Glaubens ist ein guter Grund dafür, dass man die Option einer religiösen Praxis zur theistischen Überzeugungsbildung realisieren soll: $O(H_r)$. - Wegen I)/I').

**b) Theoretische/Urteilsnormativität:**

ii) Es ist nicht der Fall, dass man aus pragmatischen Gründen eine theistische Urteilsoption realisieren soll: $\neg O(U(p))$.[12] - Wegen II)/II').

Entscheidend ist aber die dritte Konsequenz: Dass man ebenfalls *nicht atheistisch* urteilen soll. Denn nach I)/I') gilt auch:

iii) Fehlende Evidenzen für oder Evidenzen gegen die Existenz Gottes sind keine entscheidenden normativen Gründe, derentwegen man agnostisch oder atheistisch urteilen soll, dass Gott nicht existiert: $\neg O(\neg U(p))$ bzw. $\neg O(U(\neg p))$[13].

(Mithin ist für keine der konkurrierenden Urteilsoptionen, weder die theistische noch die agnostische bzw. atheistische, ein entscheidender Urteilsgrund vorhanden, dass man sie realisieren soll, wenn auch jeweils aus unterschiedlichen Gründen:

Ad i) Der Nutzen des Glaubens ist kein entscheidender Grund dafür, dass man ein theistisches Urteil fällen soll, weil man das gar nicht *kann* (wegen II/II').

Ad ii) Evidentiell bedingte[14] Gründe sind keine entscheidenden Gründe für ein agnostisches bzw. atheistisches Urteil, weil der resultierende Nicht-Glaube bzw. Unglaube die insgesamt schlechtere Option ist (wegen I/I').)

Zur Analyse dieses Ergebnisses ist Folgendes zu sagen.

Das Prinzip I)/I') ist verantwortlich dafür, dass das evaluative Übergewicht zugunsten des Glaubens aus dem konsequentialistischen Analyseschritt verhindert, dass die fehlenden Evidenzen für bzw. Evidenzen gegen die Existenz Gottes im normativen Analyseschritt normativ entscheidende Bedeutung erlangen für ein agnostisches bzw. atheistisches Urteil, obwohl ihnen (hier) keine überwiegender normativer Grund entgegensteht. Der Fall pragmatischer Argumente für den Glauben ist also eine Ausnahme zur gewohnten Regel, dass

---

[11] Vgl. Broome 2004 zum Zusammenhang zwischen entscheidenden Gründen und einem Sollen.

[12] Zu unterscheiden von der Forderung „Man soll nicht urteilen, dass p“ im Sinne einer *Unterlassungs*forderung für das Urteil, dass p - formal: $O(\neg U(p))$.

[13] Wiederum ist das nicht zu verwechseln mit der *Unterlassungs*forderung für ein atheistisches Urteil - formal: $O(\neg U(\neg p))$.

[14] D.h. Urteilsgründe, die bedingt sind durch das Fehlen von Evidenzen für die Existenz Gottes bzw. durch überwiegende Evidenzen gegen die Existenz Gottes.

normative Gründe für eine Option, denen keine stärkeren Gründe entgegenstehen, auch entscheidend sind und man diese Option deshalb realisieren soll.

Damit ist im Ergebnis die These dieses Aufsatzes bestätigt, dass sich pragmatische Argumente für den Glauben insofern nicht auf eine Praxis der theistischen Überzeugungsbildung beschränken lassen, als die pragmatische Begründung dieser Praxis zugleich normativ impliziert, dass ein widersprechendes agnostisches bzw. atheistisches Urteil *nicht* begründet ist.

## 5. Jenseits einer „praktischen Glaubenswette“: eine „defensive“ Version des pragmatischen Arguments.

Satz iii) ergibt sich unmittelbar aus der konsequentialistischen Analyse, sofern diese zeigt, dass der religiöse Glaube (allgemein oder auch nur der einer bestimmten Person)[15] einen pragmatischen Mehrwert gegenüber den Alternativen des Nicht-Glaubens oder Unglaubens besitzt. Damit ist Satz iii) insbesondere unabhängig von Satz i), der die heute religionsphilosophisch favorisierte „praktische“ Form eines pragmatischen Arguments, zugunsten einer Praxis der Überzeugungsbildung, zum Ausdruck bringt. Mit Satz iii) kann man nun aber zu einer weiteren, eigenständigen Form von pragmatischem Argument gelangen: einem Argument nicht zur *Begründung* dafür, sich praktisch um den Glauben zu bemühen, sondern zur *Rechtfertigung* eines bereits (aus anderen Gründen oder Ursachen[16]) *bestehenden* Glaubens.

Dazu gelangt man relativ leicht, wenn man iii) ergänzt um folgendes plausible, minimale Rechtfertigungsprinzip:

> R) Eine Überzeugung ist (minimal) gerechtfertigt, wenn keine entscheidenden Gründe gegen sie sprechen.[17]

Ein religiöser Glaube ist dann nach iii) und R) dadurch rechtfertigbar, dass sein Nutzen evaluativ den epistemischen Unwert aufwiegt, der ihm dadurch zukommt, das Evidenzen für ihn fehlen oder überwiegende Evidenzen gegen ihn sprechen. Obwohl in diesem Fall evidentiell bedingt ein normativer Grund (ohne stärkeren Gegengrund) gegen den Glauben spricht, sind das wegen iii) *keine entscheidenden* Gründe gegen den Glauben, derentwegen man ihn *aufgeben* sollte. Auf Grund von R) kann ein solcher Glaube in einem substantiellen Sinn als gerechtfertigt gelten.

Ein solches „defensives“ pragmatisches Argument zur (bloßen) *Rechtfertigung* des (bereits) Gläubigen, dass er glauben *darf*, erscheint sowohl bescheidener als auch realistischer im Vergleich zur herkömmlichen praktischen Auslegung pragmatischer Argumente, die den Nicht- oder Ungläubigen zu einer religiösen Konvertierungspraxis „einladen“ will.

**Christoph Mocker**

christoph.mocker@googlemail.com

---

[15] „Einzelfallsensible“ pragmatische Argumente erscheinen insgesamt deutlich aussichtsreicher als pragmatische Argumente, die sich generell auf *den* Wert *des* Glaubens berufen.

[16] Normalerweise solche der religiösen Sozialisation.

[17] Damit wird keineswegs der starke Anspruch erhoben, dass dieses Prinzip als *das* zentrale (normative/rationale) Rechtfertigungsprinzip anzusehen sei. „Minimal“ ist dieses Prinzip vielmehr in dem Sinne, dass es *eine*, allerdings fundamentale Facette dessen erfasst, was unter Rechtfertigung sinnvollerweise zu verstehen ist.

## Literatur

Broome, J. 2004: „Reasons“, in P. Petit, S. Scheffler, M. Smith und R. J. Wallace (Hrg.): *Reason and Value. Themes from the Moral Philosophy of Joseph Raz*, 28-55.

Hacking, I. 1994: „The Logic of Pascal's Wager“, in J. Jordan (Hrg.) 1994, 21-30.

Jordan, J. (Hrg.) 1994: *Gambling on God. Essays on Pascal's Wager*. Lanham und London 1994.

— 2006: *Pascal's Wager. Pragmatic Arguments and Belief in God*. Oxford 2006.

Löffler, W. 1996: „Bemerkungen zur neueren Diskussion um ‘Pascals Wette‘“, in A. Schramm (Hrg.): *Philosophie in Österreich 1996*, Wien 1996, 389-404.

— 1999: „Die Logik der existentiellen Entscheidung. ‘Pascals Wette‘ in der Sicht der Analytischen Religionsphilosophie“, in C. Kanzian und R. Siebenrock (Hrg.): *Gottesentdeckungen*, Thaur, München und Wien 1999, 105-26.

Pojman, L. P. 1986: *Religious Belief and the Will*. London und New York 1986.

Weidemann, C. 2007: *Die Unverzichtbarkeit natürlicher Theologie*. Freiburg und München 2007.

# Kunst und Moral

Lisa Katharin Schmalzried

Kunstwerke wie Nabokovs *Lolita*, Riefenstahls *Triumpf des Willens* oder Shakespeares *Der Kaufmann von Venedig* werfen die Frage auf, ob sie moralisch bewertbar sind und ob und wie eine solche moralische Bewertung ihre Gesamtbewertung als Kunstwerke beeinflusst. Diese Frage ist verknüpft mit einer grundlegenderen, nämlich welche Werttheorie für Kunstwerke man akzeptieren möchte. Man kann zwischen monistischen und pluralistischen Werttheorien unterscheiden. Gemäß einer monistischen Theorie gibt es nur einen einzigen angemessenen Bewertungsmaßstab. Eine pluralistische Theorie umfasst verschiedene Bewertungskriterien, deren Beurteilungen gemeinsam den Gesamtwert eines Kunstwerkes ausmachen. Der radikale Autonomismus, wonach der moralische Wert eines Kunstwerkes irrelevant für dessen Gesamtwert ist, ist eng verbunden mit der monistischen Idee. Eine pluralistische Theorie stellt jedoch eine vielversprechendere Alternative dar. Indem eine konkrete pluralistische Werttheorie skizziert wird, und typische Bewertungen auf Basis dieser Theorie herausgegriffen werden, kann gezeigt werden, dass der moralische Wert eines Kunstwerkes dessen Gesamtwert kontextualistisch beeinflussen kann: Ein moralisches Defizit kann sich sowohl positiv als auch negativ auf den Gesamtwert auswirken. Gleiches gilt für einen positiven moralischen Wert. Da zugleich verneint wird, dass die pluralistische Werttheorie ein eigenständiges moralisches Bewertungskriterium beinhalten sollte, argumentiert dieser Artikel für einen sogenannten *indirekten Kontextualismus*.

## 1. Hinführung

Ein Werk wie Nabokovs *Lolita* ist aus kunstphilosophischer Sicht äußerst spannend. Der Roman gilt als einer der einflussreichsten und bekanntesten Romane des 20. Jahrhunderts. Die Pariser Zeitung *Le Monde* nimmt ihn beispielweise in ihrer Liste der *100 Bücher des Jahrhunderts* auf. Dies mag verwundern, bedenkt man, welche Geschichte der Roman erzählt. Der Ich-Erzähler Humbert schildert, wie er sich in die zwölfjährige Dolores, die Tochter seiner Vermieterin, verliebt, die Mutter heiratet, um in der Nähe der Tochter zu sein, nach dem Unfalltod der Mutter mit der Minderjährigen quer durch Amerika fährt und währenddessen eine sexuelle Beziehung mit dem Mädchen unterhält.[1] Seit seinem Erscheinen steht der Roman immer wieder im Verdacht, unmoralisch zu sein, weil die Pädophilie des Protagonisten nicht deutlich verurteilt wird. Aber ist es überhaupt möglich, *Lolita* moralisch zu bewerten, und woran macht eine solche moralische Bewertung gegeben falls fest? Und wenn eine moralische Bewertung möglich ist, ist der moralische Wert dann relevant für die eigentliche Bewertung von *Lolita* als Kunstwerk betrachtet? Wie beeinflusst der moralische Wert den Gesamtwert? Diese Fragen stellen sich nicht nur im Fall von *Lolita*. Es gibt eine Vielzahl von Kunstwerken, die die Frage nach der Angemessenheit und dem Einfluss einer moralischen Bewertung aufwerfen. Man denke beispielsweise an Shakespeares *Der Kaufmann von Venedig*, an de Sades *Die 120 Tage von Sodom und Gomorra* oder auch an Picassos *Guernica*.

In der zeitgenössischen Debatte haben sich unterschiedlichste Positionen im Hinblick auf diese Problemstellung herausgebildet. Der radikale Autonomismus sieht eine moralische Bewertung von Kunstwerken als grundsätzlich sinnlos an.[2] Gemäß dem moderaten

[1] Siehe Nabokov (1999).

[2] Es ist umstritten, ob und wer diese Position tatsächlich verteidigt, siehe Jacobson (2006: 344); Giovanelli (2007: 118 f.).

Autonomismus kann es durchaus sinnvoll sein, Kunstwerke moralisch zu bewerten. Der moralische Wert eines Werkes ist aber dennoch irrelevant für die eigentliche Bewertung eines Kunstwerkes.[3] Dem widersprechen die moralistischen Positionen. Sie gehen davon aus, dass der moralische Wert relevant für die Gesamtbewertung eines Kunstwerkes ist. Die verschiedenen Spielarten des Moralismus sind unterschiedlicher Meinung darüber, wie geartet die Einflussnahme des moralischen Werts auf den Gesamtwert eines Kunstwerkes ist. Laut dem moderaten Moralismus *kann* sich ein positiver (bzw. negativer) moralischer Wert eines Werkes positiv (bzw. negativ) auf dessen Gesamtwert auswirken.[4] Der Ethizismus geht einen Schritt weiter und postuliert, dass ein positiver moralischer Wert sich *immer* positiv und ein moralisches Defizit sich immer negativ auf den Gesamtwert auswirkt, soweit der moralische Wert ästhetisch relevant ist.[5] Der Kontextualismus bricht diese positive Korrelation zwischen moralischer Bewertung und Einflussnahme auf die Gesamtbewertung auf. Ein positiver moralischer Wert kann somit den Gesamtwert eines Kunstwerkes positiv oder auch negativ beeinflussen. Gleiches gilt für ein moralisches Defizit.[6]

Der vorliegende Essay greift in diese Debatte ein. Es soll aufgezeigt werden, dass die Frage, ob und wie der moralische Wert die Gesamtbewertung eines Kunstwerkes beeinflusst, eng verbunden ist mit einer grundlegenderen Frage: Was ist eine angemessene Bewertung eines Kunstwerkes? Anders formuliert, was für eine Art von Werttheorie für Kunstwerke ist angemessen? Man kann zwischen zwei Typen von Werttheorien unterscheiden, monistischen und pluralistischen Werttheorien. Eine monistische Werttheorie geht davon aus, es gäbe nur einen einzigen angemessenen Bewertungsmaßstab für alle Kunstwerke, wohingegen laut einer pluralistischen Werttheorie es mehrere verschiedene Bewertungskriterien gibt und der Gesamtwert eines Werkes sich aus den einzelnen Bewertungen ergibt. Der moderate Autonomismus ist eng mit dem monistischen Gedanken verknüpft. Indem eine grundsätzliche Kritik an dem monistischen Gedanken vorgebracht wird und die Idee einer pluralistischen Werttheorie als vielversprechendere Werttheorie ins Spiel gebracht wird, wird es denkbar, dass der moralische Wert relevant für die Gesamtbewertung eines Kunstwerkes ist. Auf Basis eines konkreten Vorschlags einer pluralistischen Werttheorie soll des Weiteren gezeigt werden, dass der moralische Wert eines Werkes dessen Gesamtwert kontextualistisch beeinflussen kann. Dies geschieht, indem typische Beurteilungen auf Basis der einzelnen Kriterien der pluralistischen Werttheorie herausgegriffen werden und aufgezeigt wird, wie ein moralischer Wert diese beeinflussen kann. Somit kann gezeigt werden, dass der moralische Wert eines Kunstwerkes relevant für dessen Gesamtbewertung ist, ohne dass ein moralisches Bewertungskriterium als eigenständiger Beurteilungsmaßstab in die pluralistische Werttheorie aufgenommen wird.

Der Artikel unterteilt sich in vier Teile. Zunächst wird eine Erwiderung auf den radikalen Autonomismus skizziert, indem zwei Möglichkeiten dargestellt werden, wie man Kunstwerke sinnvoll moralisch bewerten kann. Daran anschließend wird die Verbindung zwischen moderaten Autonomismus und monistischen Werttheorien aufgezeigt, ebenso wie solche Werttheorien kritisiert werden. In einem dritten Schritt wird eine pluralistische Werttheorie vorgestellt, um in einem vierten Schritt auf deren Basis zu argumentieren, dass der moralische Wert den Gesamtwert kontextualistisch beeinflussen kann.

[3] Siehe bsp. Anderson und Dean (1998).

[4] Siehe bsp. Carroll (1996), (1998).

[5] Siehe bsp. Gaut (2007).

[6] Siehe bsp. Jacobson (1997).

## 2. Radikaler Autonomismus und die moralische Bewertbarkeit von Kunstwerken

Ist es möglich, Kunstwerke moralisch zu bewerten? Der radikale Autonomismus verneint dies. Diese Position ist tatsächlich radikal. In unserem alltäglichen Umgang mit Kunstwerken beschreiben wir diese häufig mit moralisch aufgeladenem Vokabular und meist verstehen wir solche Aussagen auch problemlos. So werfen wir beispielsweise Tarantinos *Pulp Fiction* vor gewaltverherrlichend zu sein oder verurteilen Riefenstahls *Triumpf des Willens*, weil der Film Hitler glorifiziert. Warum sollte dies nicht möglich sein?

Zwei Argumente stützen die Position des radikalen Autonomismus. Zum einen erwähnt Gaut eine Überlegung, die man als *Argument des Kategorienfehlers* bezeichnen kann: Es ist nicht möglich, Kunstwerke moralisch zu bewerten, da nur Personen (sinnvoll) moralisch bewertet werden können. Die moralische Bewertung macht an den Handlungen und u.U. an den Überzeugungen und Emotionen der Person fest. Da Kunstwerke keine Personen sind, nicht handeln und denken können, macht es keinen Sinn, sie moralisch zu bewerten.[7]

Eine zweite Überlegung stammt aus der Debatte um den (ästhetischen) Kognitivismus, d.h. der Frage, ob Kunstwerke Wissen vermitteln können und ob diese Fähigkeit zur Wissensvermittlung gegeben falls relevant für die Gesamtbewertung des Werkes ist. Spricht man einem Kunstwerk einen moralischen Wert zu, mag man ihm damit die Fähigkeit attestieren, moralisch relevantes Wissen zu vermitteln. So kann man die Aussagen verstehen, ein Kunstwerk sei moralisch wertvoll oder auch gefährlich. Manche Antikognitivisten bestreiten nun, dass Kunstwerke Wissen vermitteln können. So wird beispielsweise argumentiert, Kunstwerke würden höchstens Trivialitäten beinhaltenen, so dass man von ihnen nichts Neues *lernen* könnte.[8] Oder aber es wird darauf verwiesen, dass Kunstwerke keine zuverlässigen Begründungen liefern, zumal wenn es sich um fiktionale Werke handelt.[9] Wenn Kunstwerke aber kein (moralisches) Wissen vermitteln können, können wir sie dahingehend auch nicht (moralisch) kritisieren.

Die Kritik am Argument des Kategorienfehlers setzt an dessen starken Annahme an, *ausschließlich* Personen wären sinnvoll moralisch bewertbar. So fällt erstens auf, dass wir uns über allgemeine Handlungsweisen Gedanken machen und fragen, ob diese moralisch richtig oder falsch sind. Dabei schreiben wir sie nicht zwangsläufig einer Person zu.

Zweitens begehen wir nicht den Fehler und verwechseln Kunstwerke mit Personen, noch behandeln sie wie x-beliebige Gegenstände, wenn wir sie moralisch kritisieren. Sprechen wir von „Kunstwerk“, meinen wir damit entweder den Gegenstand oder aber den Inhalt des Kunstwerkes.[10] Die moralische Kritik eines Werkes setzt an dessen inhaltlicher Ebene an, und daher wird es möglich sie moralisch zu kritisieren. Im Falle *Lolitas* kritisieren wir beispielsweise die Art und Weise wie Humberts Pädophilie dargestellt wird. Diese wird nicht klar verurteilt. Teilweise scheint der Roman den Leser dazu einzuladen, lustvoll an Humberts sexuellen Phantasien teilzuhaben. Als anderes Beispiel sei Elis Roman *American Psycho* erwähnt. Die Art und Weise wie die Morde und Vergewaltigungen geschildert werden, ist moralisch problematisch, gerade weil das Werk darin scheitert ironisch zu sein.

Verallgemeinern wir diese Beispiele: Kunstwerke drücken u.U. entweder explizit oder aber durch die Art und Weise ihrer Darstellung moralische Haltungen aus.[11] Unter einer

---

[7] Siehe Gaut (2007: 69).

[8] Siehe Stolnitz (1992).

[9] Siehe Carroll (2002: 4 ff.).

[10] Siehe Wollheim (1980).

[11] Dieser Gedanke klingt bei unterschiedlichen Autoren an. Teilweise sprechen sie nicht von der moralischen Haltung, sondern von einem moralischen Standpunkt oder einer moralischen Perspektive, siehe beispielsweise Pole (1962: 200); Walton (1994); Carroll (1996: 421); Devereaux (1998: 237); Booth

moralischen Haltung sei eine Positionierung hinsichtlich eines moralisch relevanten Sachverhalts verstanden. Teilt der Kritiker die moralische Beurteilung des Werkes, spricht er ihm einen moralischen Wert zu, anderenfalls ein moralisches Defizit.

Wie kann man dem antikognitivistischen Argument begegnen? Dieses Argument konzentriert sich primär auf propositionales Wissen. Spricht man von moralisch relevantem Wissen, muss man damit aber nicht nur propositionales moralisches Wissen meinen, also das Wissen, dass etwas moralisch richtig oder falsch ist. Es gibt auch praktisches moralisches Wissen. Als genuin moralisches praktisches Wissen sei die Fähigkeit bezeichnet, allgemeine moralische Grundsätze auf konkrete Situationen anzuwenden.[12] Kant spricht hier von moralischer Urteilskraft.[13] Daneben kann es praktisches Wissen geben, welches u.U. moralisch wertvoll ist. Man denke beispielsweise an eine gut geschulte Vorstellungskraft oder an Emphatiefähigkeit.[14] Des Weiteren gibt es das Wissen, wie es ist, sich in einer bestimmten Situation zu befinden. Dieses Wissen kann moralisch relevant sein.[15] Es kann helfen, andere Sichtweisen und die Komplexität von Situationen besser zu verstehen.

Indem man den Wissensbegriff so erweitert, kann man dem antikognitvistischen Argument begegnen.[16] Kunstwerke verfügen unterschiedlich gut ausgeprägt über die Fähigkeit, diese unterschiedlichen Arten von Wissen zu vermitteln. Man kann ihnen somit ein moralisches Potenzial zusprechen.[17] Stuft man das Wissen, das die Werke transportieren können, als moralisch problematisch ein, haben die Werke in dieser Hinsicht ein moralisches Defizit. Handelt es sich um moralisch wertvolles Wissen, ist das Werk moralisch positiv zu beurteilen.

Diese beiden Argumentationsstränge zusammengenommen, zeigen Möglichkeiten auf, wie Kunstwerke sinnvoll moralisch bewertbar sind. Der moralische Wert eines Kunstwerkes bestimmt sich zum einen daraus, ob und welche moralische Haltungen es ausdrückt, zum anderen und ergänzend, ob das Werk über das Potenzial verfügt, moralisch relevantes Wissen zu vermitteln.

## 3. Moderater Autonomismus auf Basis monistischer Werttheorien

Ist der moralische Wert, so definiert, relevant für die Gesamtbewertung eines Kunstwerks? Stellt man diese Frage, wirft man damit eine grundlegendere auf, nämlich wann eine Bewertung relevant oder, anderes formuliert, angemessen für die Bewertung eines Kunstwerkes ist. Und hiermit wird nach einer Werttheorie für Kunstwerke gefragt. Es ist zu erwarten, dass die Antwort auf die Frage nach der Relevanz des moralischen Werts entscheidend davon abhängt, welche Werttheorie man zu Grunde legt.

Diese Hypothese bestätigt sich im Falle des moderaten Autonomismus. Gemäß dem moderaten Autonomismus ist der moralische Wert irrelevant für die Gesamtbewertung. Diese Haltung steht im engen Zusammenhang mit der Idee einer monistischen Werttheorie, laut der es nur einen einzigen angemessenen Bewertungsmaßstab für Kunst gibt.

---

(1998: 376 ff.); Carroll (2000); Kieran (2002: 41), Devereaux (2004: 6); Stecker (2005[1]: 139 f.); Gaut (2007: 68).

[12] Siehe Nussbaum (1990: 156); Kieran (1996: 341).

[13] Siehe Kant (1977: AA VII 275).

[14] Siehe bsp. Putnam(1975-76); Currie (1998); Gaut (2006: 116).

[15] Siehe bsp. Harold (2008: 50 ff).

[16] Zur Vertiefung in die Debatte um den ästhetischen Kognitivismus siehe Kieran (2003); Gaut (2003), (2006); Carroll (2008).

[17] Siehe beispielsweise Beardsmore (1973); Novitz (1980: Kap. 6); Carroll (2002); Gaut (2003), (2006).

### *3.1 Radikaler Moralismus*

Zunächst ist diese Verbindung nicht offensichtlich, könnte man doch einen radikalen Moralismus vertreten:[18] Allein der moralische Wert eines Kunstwerkes ist relevant für dessen Gesamtbewertung.[19]

Der radikale Moralismus wird jedoch mit schwerwiegenden Einwänden konfrontiert. Erstens sind nicht alle Kunstwerke moralisch bewertbar, so dass man sie auf Basis des radikalen Moralismus nicht (angemessen) beurteilen könnte.[20] Zweitens kann der radikale Moralismus nicht erklären, wie es gute, aber moralisch zweifelhafte Werke geben kann oder umgekehrt moralisch einwandfreie, aber mittelmäßige oder schlechte Kunstwerke.[21] Will man nicht viele gut etablierte Kunstbewertungen revidieren, sollte eine Werttheorie dies erklären können. Drittens zeigt sich der radikale Moralismus insensibel gegenüber der Form von Kunstwerken.[22] Die Form und Struktur eines Werkes spielt nur insoweit eine Rolle, als sie die moralische Haltung des Werkes transportiert und hilft moralisch relevantes Wissen zu vermitteln. Ihr wird kein eigenständiger Wert zugesprochen. Dies kann im Hinblick auf Kunstkritik als kontraintuitiv betrachtet werden. Somit stellt der radikale Moralismus eine stark normativ aufgeladene Werttheorie dar, die viele gut etablierte Bewertungen von Kunstwerken nicht erfassen und erklären kann.

### *3.2 Radikaler Ästhetizismus*

Hält man an der monistischen Strategie fest, stellt sich die Frage, welcher andere Bewertungsmaßstab als einzig relevant für die Bewertung von Kunstwerken angesehen werden kann. Nahe liegend ist es auf den ästhetischen Wert eines Werkes zu verweisen und somit eine radikal ästhetizistische Position vorzuschlagen. Die Schwierigkeit hierbei ist zu erklären, was man genau unter dem ästhetischen Wert eines Werkes zu verstehen hat.

Zunächst sei eine wichtige Unterscheidung erwähnt. Teilweise wird „ästhetisch wertvoll“ als Synonym für „künstlerisch wertvoll“ verwandt. [23] So verstanden, würde es sich beim radikalen Ästhetizismus um eine tautologische Position handeln, und die Frage, welche Bewertung eines Kunstwerkes angemessen ist, würde unverändert Bestand haben. Damit die Position interessant wird, sollte man ein engeres Verständnis von ästhetischem Wert bevorzugen, bei welchem die Frage, ob der ästhetische Wert mit dem Wert eines Kunstwerk als Ganzes betrachtet zusammenfällt, (zunächst) offen bleibt.

Beardsley schlägt eine instrumentalistische Definition des ästhetischen Werts vor:[24]

> [...] the aesthetic value of anything is its capacity to impart–through cognition of it–a marked aesthetic character to experience. (Beardsley 1979: 728)

Greift der radikale Ästhetizismus auf diese Definition zurück, kann die ästhetische Bewertung durch die moralische tangiert werden? Die Aufmerksamkeit auf moralische Aspekte eines Werkes verhindert eine ästhetische Erfahrung, weil eine solche Erfahrung meist als desinteressiert bzw. als losgelöst von praktischen Belangen definiert wird.[25] Somit lässt sich aus dieser Form des radikalen Ästhetizismus ein moderater Autonomismus ableiten.

---

[18] Siehe Posner (1997).

[19] Tolstoi (1998) und Sidney (1966) scheinen eine Art von radikalem Moralismus zu vertreten.

[20] Siehe auch Carroll (1996: 226), (2000: 357); Gaut (2007: 67).

[21] Siehe auch Posner (1997: 5).

[22] Siehe auch Beardsmore (1971: 10 ff.); Gass (1987: 41).

[23] Siehe hierzu auch Gaut (2007: 26).

[24] Siehe auch Beardsley (1958: 531), (1979: 728).

[25] Siehe bsp. Beardsley (1958: 457); vgl. auch Dickie (1985: 4f.).

Doch was ist genau eine ästhetische Erfahrung, gibt es eine solche überhaupt und wie ist sie gegeben falls zu definieren?[26] In der Literatur finden sich unzählige Definitionsvorschläge der ästhetischen Erfahrung. Dabei offenbart sich eine grundlegende Schwierigkeit. Es ist nicht klar, welche Erfahrungen überhaupt als ästhetisch gelten sollen.[27] Eine ästhetische Erfahrung soll eine Erfahrung sein, die typischerweise, aber nicht ausschließlich von Kunstwerken hervorgerufen wird, und nicht jede Erfahrung eines Kunstwerkes soll eine ästhetische sein. Durch diese fehlende (vortheoretische) Klarheit darüber, welche Erfahrung eine ästhetische ist, scheint jeder Definitionsversuch mehr oder weniger *festzusetze*n, was eine ästhetische Erfahrung sein *soll*. Dies macht eine Kritik an diesen Definitionsversuchen so schwierig. Eine Definition des ästhetischen Werts, die nicht auf eine ästhetische Erfahrung referiert, ist daher zu bevorzugen.

Lassen wir diese Kritik einen Moment außen vor. Warum sollte man die monistische Forderung dieses radikalen Ästhetizismus akzeptieren? Das Problem ist, dass Kunstwerke nicht nur ästhetische Erfahrungen hervorrufen, sondern beispielsweise auch kognitive oder religiöse Erfahrungen.[28] Auch wird Kunst weder ausschließlich um der ästhetischen Erfahrung willen erschaffen, noch konsumiert. Hinzukommt, dass ästhetische Erfahrungen nicht ausschließlich durch Kunst hervorgerufen werden.[29] Warum sollte also allein die ästhetische Erfahrung relevant sein?

Im Folgenden wird eine Alternative zur instrumentalistischen Definition vorgeschlagen. Diese baut auf drei Annahmen auf. Zunächst scheint eine formalistische Einschränkung sinnvoll zu sein, wenn man den ästhetischen Wert im engeren Sinn definieren will: Allein die Form und Struktur bestimmt den ästhetischen Wert.[30] Unter „Form“ kann man sinnlich wahrnehmbare Eigenschaften wie Farben, Linien, Formen, Klänge, usw. verstehen, wohingegen „Struktur“ das Zusammenspiel dieser formalen Eigenschaften bezeichnet.

Die Form und Struktur sollte zweitens positiv bewertet werden, ohne dass diese positive Bewertung sich aus einem konkreten Zweck ableitet. Die Form und Struktur eines Flugzeugs kann beispielsweise dahingehend beurteilt werden, ob es besonders aerodynamisch geformt ist. In diesem Falle bewertet man das Flugzeug aber nicht ästhetisch. Hiermit greift man den Gedanken Kants auf, dass ein ästhetisches Urteil desinteressiert sein soll.[31] Es soll aber eine schwächere These als die Kants vertreten werden, da Kant von einem desinteressierten Urteil fordert, man dürfe kein Interesse an der Existenz des Objektes haben.[32] Es geht vielmehr darum, dass der Grund, warum die Form und Struktur positiv bewertet wird, nicht auf einer externen Zweckbetrachtung beruhen darf.[33]

Wie äußert sich nun eine solche positive Bewertung der Form und Struktur? Hier hilft Sibleys Theorie der ästhetischen Ausdrücke weiter.[34] Sibley selbst geht nicht auf die Frage ein, wie man den ästhetischen Wert definieren kann.[35] Er schließt aber auch nicht aus, dass man ihn über ästhetische Eigenschaften definieren könnte. Wenn wir etwas ästhetisch bewerten, sagen wir meist nicht „Das ist ästhetisch wertvoll“. Vielmehr sprechen wir davon, es sei

---

[26] Siehe bsp. Stolnitz (1960); Dickie (1964), (1965); Beardsley (1969), (1991); Carroll (2001); Matravers, (2003); Stecker (2005[2]: Kap. 5).

[27] Für einen ähnlichen Gedanken siehe Diffey (1986).

[28] Siehe hierzu Price (1979: 131).

[29] Siehe auch Kieran (2005: 98), Gaut (2007: 92 ff).

[30] Für einen ähnlichen Gedanken siehe Bell (1914); Zangwill (2001: 127).

[31] Siehe Kant (1963: AA V 211).

[32] Siehe Kant (1963: AA V 204).

[33] Siehe auch Stecker (2005[1]: 37).

[34] Siehe Sibley (1959), (1963).

[35] Siehe Sibley (1965: 136).

harmonisch, ausbalanciert, schön und so weiter. Dies sind ästhetische Ausdrücke im Sinne Sibleys. Mit diesen begründen wir unser ästhetisches Urteil.

Vereint man diese drei Gedanken, kann man definieren:

> Etwas ist genau dann ästhetisch wertvoll, wenn ihm in Bezug auf seine Form und Struktur ohne auf einen externen Zweck bezugnehmend positiv bewertbare bzw. vorwiegend positiv bewertbare ästhetische Eigenschaften (richtigerweise) zugesprochen werden.

Diese Definition kann als minimalistisch bezeichnet werden. Erstens bezieht sie sich nicht auf eine ästhetische Erfahrung. Zweitens liefert sie keine vollständige Liste all der ästhetischen Eigenschaften, die an eine ästhetische Bewertung ausdrücken können. Letzteres ist für den vorliegenden Kontext nicht notwendig, wie sich zeigen wird.

Kann der ästhetische Wert, minimalistisch definiert, durch den moralischen Wert tangiert werden? Pole bringt ein Argument vor, dass genau solch eine Interaktion zeigen will.[36] Er greift die ästhetische Eigenschaft der Inkohärenz heraus. Wenn ein Kunstwerk inkohärent ist, so Pole, wird sein ästhetischer Wert gemindert. Wenn nun ein Werk eine unmoralische Haltung ausdrückt, wird das Werk insoweit inkohärent. Somit vermindert eine unmoralische Haltung den ästhetischen Wert eines Werkes.

Warum wird ein Werk durch eine unmoralische Haltung inkohärent? Laut Pole ist es nicht die unmoralische Haltung selbst, die inkohärent ist. Vielmehr passt etwas in einem Werk nicht zusammen, drückt es eine unmoralische Haltung aus.[37] Diese Forderung bewahrheitet sich nicht, denkt man beispielsweise an den *Triumph des Willens*.[38] In diesem Film harmonieren formale Aspekte perfekt mit der moralischen Botschaft des Films.[39] Es handelt sich um ein besonders gelungenes Beispiel eines kohärenten Werkes.

Ein anderes Argument bringt Gaut vor, das Argument der moralischen Schönheit.[40] Wenn ein Werk schön ist, trägt dies positiv zu dem ästhetischen Wert des Werkes bei. Wenn ein Werk eine moralisch richtige Haltung ausdrückt, wird es zu einem (innerlich) schönen Werk. Somit trägt eine moralisch richtige Haltung positiv zu dem ästhetischen Wert eines Werkes bei.

Die zentrale Annahme dieses Arguments ist, dass „innere Schönheit“ kein metaphorischer Ausdruck ist, sondern es mindestens zwei Arten von Schönheit gibt, innere und äußere Schönheit. Gaut verweist hierfür zum einen auf die lange philosophische Tradition von innerer Schönheit zu sprechen und zum anderen auf unseren alltäglichen Sprachgebrauch.[41] Beides überzeugt nur bedingt, so lange keine allgemeine Theorie der Schönheit dargestellt wird, aus der sich ableiten lässt, dass innere Schönheit echte Schönheit ist. Abgesehen davon ist anzunehmen, dass ein moderater Autonomist die erste Prämisse des Arguments nur dann akzeptieren würde, wenn sie explizit auf äußere Schönheit eingeschränkt wird.

Muss man nun alle ästhetischen Eigenschaften, die sich auf die Form und Struktur eines Werkes beziehen, einzeln betrachten, um die Frage zu beantworten, ob der moralische Wert den ästhetischen Wert beeinflussen kann? Dies ist aber nicht der Fall, wenn man nochmals betont, dass der ästhetische Wert (in einem engen Sinn) sich auf die Form und Struktur eines Werkes bezieht. Wie oben ausgeführt, ist der moralische Wert eines Werkes an die inhaltliche Ebene des Werkes geknüpft. Wenn man sich allein auf die Form und Struktur eines Werkes konzentriert, muss man von dem Inhalt des Werkes und somit auch von dessen moralischen

---

[36] Siehe Pole (1962: 206).
[37] Siehe Pole (1962: 206).
[38] Siehe Gaut (1998: 190).
[39] Siehe Devereaux (1998).
[40] Siehe Gaut (2007: 115).
[41] Siehe Gaut (2007: 116 ff.).

Aspekten abstrahieren. Somit ist der ästhetische Wert, minimalistisch definiert, unabhängig von dem moralischen Wert eines Werkes.

Erneut drängt sich die Frage auf, ob und warum allein der ästhetische Wert eines Werkes relevant für die Kunstkritik sein sollte. Zunächst einmal wurde die formalistische Einschränkung einfach festgelegt. Dies kann man verteidigen, geht es darum, den ästhetischen Wert im engeren Sinne zu definieren. Problematisch wird es, wenn man den ästhetischen Wert im engeren Sinne mit dem im weiteren Sinne, also der Gesamtwert eines Kunstwerkes, gleichsetzt. Dies kann man auch nicht mit Verweis auf Sibleys Auflistung von ästhetischen Eigenschaften begründen. Vielmehr umfasst Sibleys Liste gerade auch ästhetische Eigenschaften, die sich auf den Inhalt oder die Wirkung eines Werkes beziehen, wie „rührend", „banal" oder „tiefgründig".[42] Gerade in unserer alltäglichen Kritik von Kunstwerken beschränken wir uns nicht nur auf die Kritik der Form und Struktur eines Werkes.

### *3.3 Kritik an monistischen Theorien*

Der radikale Ästhetizismus ist in abgeschwächter Form mit dem gleichen Problem konfrontiert wie der radikale Moralismus. Beide Werttheorien zeichnen eine Art, Kunstwerke zu bewerten, als einzig richtige aus und betrachten alle anderen als unangemessen für die Kunstkritik. Beide sind somit stark normativ aufgeladen. Es ist zu erwarten, dass diese starke Normativität alle monistischen Werttheorien eint. Damit kommt man zur grundlegenden Frage, ob man die Grundidee hinter einer monistischen Werttheorie verteidigen möchte.

Eine Möglichkeit, die monistische Strategie zu verteidigen, wäre auf ein Kriterium der allgemeinen Anwendbarkeit zu verweisen.[43] Ein solches fordert, dass man Kunstwerke nur anhand eines Maßstabes bewerten soll, der auf alle Kunstwerke anwendbar ist. Hieraus folgt nicht zwangsläufig eine monistische Werttheorie. Da es aber so schwierig ist ein Kriterium zu finden, dass auf alle Kunstwerke anwendbar ist, folgt meist eine monistische Theorie.

Jedoch wird das Kriterium der allgemeinen Anwendbarkeit erst dann plausibel, wenn man bereits eine monistische Werttheorie akzeptiert hat. Wenn es nur ein angemessenes Bewertungskriterium gibt, ist es wünschenswert, dass alle Kunstwerke anhand dieses bewertbar sind. Sobald man zulässt, dass es mehrere Bewertungskriterien gibt, mag ein Werk auf Basis eines Kriteriums nicht bewertbar sein und dennoch auf Basis der anderen Kriterien beurteilbar sein.

Eine andere Verteidigung der monistischen Strategie könnte auf der Forderung aufbauen, man solle Kunstwerke nur anhand der Eigenschaft bewerten, die sie zu Kunstwerken werden lässt. Viele Vertreter einer monistischen Werttheorie knüpfen diese Verbindung zwischen Werttheorie und Kunstdefinition.[44]

Mit diesem Vorschlag ist die Problematik verbunden, eine essentielle Eigenschaft für Kunstwerke zu finden. Aber selbst wenn es eine solche essentielle Eigenschaft gibt, folgt daraus nicht, dass Kunstwerke nicht auch anhand weiterer Kriterien bewertbar sind. Ein Roman mag Eigenschaften aufweisen, die er mit Gemälden nicht teilt und die somit nicht essentiell für Kunst sind, und dennoch mögen diese relevant für die angemessene Bewertung des Romans sein. Man denke beispielsweise an die Wortwahl und die Komposition der einzelnen Kapitel.

---

[42] Siehe Sibley (1959: 421 ff.).

[43] Siehe hierzu Carroll (1985: 330).

[44] Siehe bsp. Bell (1914); Tolstoi (1998: 75); Beardsley (1958: xix).

## 4. Skizze einer pluralistischen Werttheorie

Argumentiert man gegen monistische Werttheorien, gelangt man zum Gedanken einer pluralistischen Werttheorie. Gemäß dieser gibt es verschiedene angemessene Kunstbewertungskriterien. Der eigentliche Wert eines Kunstwerkes, sein Gesamtwert, ergibt sich aus den unterschiedlichen Bewertungen auf Basis der einzelnen Kriterien. Die entscheidende Frage ist nun, welche Kriterien in eine solche Theorie aufgenommen werden sollten?

Eine Möglichkeit, diese Frage zu beantworten, könnte darin bestehen, ein allgemeines Kriterium anzugeben, wann ein Bewertungsmaßstab angemessen ist. Wie wir im nächsten Abschnitt sehen werden, ist ein solches Kriterium problematisch. Daher soll ein anderer Weg gewählt werden. Den Ausgangspunkt auf der Suche nach einer angemessenen Werttheorie bildet unserer alltäglicher Umgang mit Kunstwerken und gut etablierte Beurteilungen von Kunstwerken. Die Frage ist, welche Kriterien hier mit hineinspielen. Zumindest vier Kriterien spielen eine Rolle:[45]

Damit sich die pluralistische Werttheorie nicht des Vorwurfs der Insensibilität gegenüber der Form und Struktur eines Kunstwerkes schuldig macht, sollte sie ein ästhetisches Bewertungskriterium beinhalten. Mit Hilfe dieses Kriteriums kann der (im engen Sinne) ästhetische Wert eines Kunstwerkes beurteilt werden.

Nun wurde bereits darauf hingewiesen, dass neben der Form und Struktur eines Werkes auch dessen Inhalt eine Rolle spielt. Denken wir beispielsweise an einen Roman wie *Anna Karenina*. Wenn wir sagen, dabei handele es sich um einen tiefgründigen Roman, machen wir eine positive Aussage über die Qualität des Werkes. Diese Intuition erfasst ein kognitives Bewertungskriterium. Es ermöglicht, den Inhalt eines Kunstwerkes dahingehend zu bewerten, ob man etwas von ihm lernen kann.[46] Typische Beurteilungen auf Basis dieses Bewertungsmaßstabes sind, „tiefgründig", „treffend" oder „banal".

Antikognitivisten wie Lamarque und Olson bestreiten, dass der kognitive Bewertungsmaßstab angemessen für die Kunstkritik ist. Sie verweisen darauf, dass Kunstkritiker nicht über die Wahrheit bzw. Falschheit von Aussagen diskutieren, die sie in Kunstwerken finden.[47] Aber auch wenn professionelle Kunstkritiker hierüber nicht reden, bedeutet es nicht, dass dies auch für „nicht-professionelle Kunstkritiker" gilt. Noch entscheidender ist ein anderer Punkt: Das kognitive Bewertungskriterium beurteilt nicht einfach, ob ein Werk wahr oder falsch ist. Bloße Wahrheit genügt nicht. Denken wir beispielsweise an Banalität. Beurteilen wir etwas als banal, machen wir nicht den Vorwurf, es sei nicht wahr. Beurteilen wir ein Werk aber als banal, beurteilen wir es in kognitiver Hinsicht negativ.

Als drittes Bewertungskriterium wird ein affektiv-antwortabhängiger Bewertungsmaßstab vorgeschlagen. Hiermit kann man einem Werk einen Wert attestieren, der sich daraus ableitet, ob es seinen eigenen Ansprüchen hinsichtlich affektiv-antwortabhängiger Eigenschaften gerecht wird.[48] Will ein Werk beispielsweise eine Komödie sein, ist es angemessen zu fragen, ob sie lustig ist. Charakteristische Beurteilungen auf Basis dieses Kriteriums sind „unterhaltsam", „komisch" oder „langweilig".

Um dieses Kriterium zu verteidigen, sollte man zwischen werk-internen und werk-externen Gründen für das Gelingen oder Scheitern, die intendierten Eigenschaften hervorzurufen, unterscheiden. Wenn eine schwer depressive Person eine Komödie ansieht, wird sie diese

[45] Diese vier Bewertungskriterien mögen nicht die einzig denkbaren Bewertungskriterien einer pluralistischen Werttheorie sein. Vielmehr sollte man diesen Vorschlag als Skizze einer pluralistischen Werttheorie begreifen, die u.U. erweitert werden kann.

[46] Siehe auch Gaut (2006: 122), (2007: 167).

[47] Siehe Lamarque und Olson (1994: 282 ff.); Lamarque (2006: 136 f.).

[48] Siehe auch Carroll (1998: 420).

wahrscheinlich nicht lustig finden. Hierfür kann man jedoch der Komödie keinen Vorwurf machen. Der Grund, warum sie nicht als lustig empfunden wurde, war ein werk-externer Grund. Wenn auf der anderen Seite ein Horrorfilm nicht furchterregend ist, weil die gezeigten Monster schlecht animiert wurden, handelt es sich um einen werk-internen Grund, wofür man das Werk kritisieren kann.

Viertens ist ein kunsthistorisches Bewertungskriterium angemessen. Stellen wir uns zwei scheinbar identische Gemälde vor, die wir folglich ästhetisch, kognitiv und antwortabhängig gleich bewerten sollten. Jedoch handelt es sich bei einem Werk um eine perfekte Kopie des anderen. Dies scheint ein für die Bewertung der Kunstwerke relevanter Unterschied zu sein. Man möchte das Original besser bewerten als die Kopie. Hierbei hilft ein kunsthistorischer Bewertungsmaßstab. Mit diesem kann man einen Wert zusprechen, der sich aus der kunsthistorischen Stellung des Werkes ableitet. Beurteilungen, an die man hier denken kann, sind „originell“, „kreativ“ oder „plagiiert“.

Gegen dieses Kriterium mag eingewandt werden, dass es kunsthistorische Aussagen keine Bewertungen beinhalten, sondern rein deskriptiv sind. Zu sagen „Werk x ist ein Original“, besagt einfach, dass es sich dabei um keine Kopie handelt. Teilweise mag es so sein, dass eine kunsthistorische Aussage rein deskriptiv ist, aber dies gilt nicht für alle Aussagen. Ein Werk als originell zu bezeichnen impliziert eine Bewertung.

Zusammenfassend ist festzuhalten: Eine pluralistische Werttheorie sollte zumindest vier Kriterien beinhalten, einen ästhetischen, einen kognitiven, einen antwortabhängigen und ein kunsthistorischen Bewertungsmaßstab. Wie deutlich wurde, baut die hier vorgestellte Werttheorie auf Sibleys Liste von ästhetischen Ausdrücken auf. Diese Ausdrücke sind typisch für die Kunstkritik und die vier Bewertungskriterien versuchen eine Klassifizierung dieser Ausdrücke.

Wichtig ist nochmals zu betonen, dass nicht gefordert wird, jedes Kunstwerk müsse anhand eines jeden Maßstabes beurteilbar sein. Außerdem kann es vorkommen, dass die Bewertungen auf Basis der einzelnen Bewertungen unterschiedlich positiv bzw. negativ ausfallen. Ein Werk kann in einer Hinsicht lobenswert sein, während es sich in anderer Hinsicht nicht auszeichnet. Will man den Gesamtwert eines Werkes bestimmen, muss man die unterschiedlichen Bewertungen abwägen und gewichten.

Der große Vorteil einer pluralistischen Werttheorie liegt darin, dass sie weniger normativ aufgeladen ist als eine monistische Theorie. Kunstwerke werden von unterschiedlichen Standpunkten aus betrachtet und bewertet. Diese Multidimensionalität der Kunstkritik kann eine pluralistische Werttheorie besser erfassen als eine monistische Werttheorie.

## 5. Argumente für einen indirekten Kontextualismus

Die für diesen Artikel entscheidende Frage ist, welche Rolle eine moralische Bewertung in der pluralistischen Werttheorie spielt. Zunächst sind zwei Möglichkeiten denkbar, wie der moralische Bewertungsmaßstab den Gesamtwert beeinflussen kann. Zum einen könnte man ein eigenständiges moralisches Bewertungskriterium in die pluralistische Theorie aufnehmen. In diesem Falle würde der moralische Wert eines Werkes dessen Gesamtwert *direkt* beeinflussen. Eine moralische Bewertung könnte den Gesamtwert auch *indirekt* tangieren. Hier würde dann eine Bewertung auf Basis der obigen vier Bewertungskriterien durch den moralischen Wert beeinflusst werden. Somit muss im Folgenden untersucht werden, ob der moralische Wert den Gesamtwert eines Kunstwerkes, direkt, indirekt oder vielleicht sogar direkt-indirekt beeinflusst.

### *5.1 Argumente für eine indirekte Einflussnahme*

Beginnen wir mit der Frage, ob der moralische Wert eines Werkes dessen Gesamtwert indirekt beeinflussen kann. Wie oben ausgeführt, ist der ästhetische Wert eines Werkes unabhängig von dessen moralischen Wert. Doch kann eine moralische Bewertung eine kognitive, affektiv-antwortabhängige oder kunsthistorische Beurteilung beeinflussen? Im Folgenden soll gezeigt werden, dass eine solche Beeinflussung möglich ist. Um diese zu beweisen, greifen wir jeweils eine typische Beurteilung auf Basis dieser Kriterien heraus, um dann Fälle zu skizzieren, bei welchen diese Beurteilung durch den moralischen Wert tangiert wird.

Ein Beispiel für eine positive kognitive Bewertung ist das Urteil, ein Werk sei tiefgründig. Charakteristischerweise gilt ein Werk als tiefgründig, wenn es unterschiedliche Aspekte und Perspektiven berücksichtigt, es nicht eindimensional und durchdacht erscheint. Ob ein Werk nun tiefgründig ist, kann u.a. von seinem moralischen Wert abhängen.

Dies kann man am Beispiel des Films *Das Meer in Mir* veranschaulichen. Dieser Film verdient es als tiefgründig bezeichnet zu werden. Er erzählt die Geschichte des Spaniers Ramon Sanpedro, der seit einem Badeunfall vor über 20 Jahren Tetraplegiker ist und vor Gericht für aktive Sterbehilfe kämpft. Den Film kann man als Plädoyer für die Zulässigkeit aktiver Sterbehilfe verstehen. Jedoch verkörpern die unterschiedlichen Charaktere des Films unterschiedliche Haltungen zum Thema Sterbehilfe, und in den gezeigten Diskussionen werden ihre Argumente dargestellt. Auch erfährt der Zuschauer viel über ihre emotionale Verfasstheit. All dies lässt den Film tiefgründig werden. Was hier beschrieben wird, ist aber genau das, was den moralischen Wert des Films beschreibt. Somit beeinflusst der moralische Wert die Tiefgründigkeit des Werkes.

Wichtig ist, darauf hinzuweisen, dass Tiefgründigkeit nicht Wahrheit impliziert.[49] Etwas kann tiefgründig sein, ohne wahr zu sein. Anselms Gedanken über Gott mögen nicht wahr sein, können aber dennoch tiefgründig sein. Übertragen auf das Beispiel des Films *Das Meer in Mir* bedeutet dies, dass man den Film auch dann als tiefgründig bezeichnen kann, wenn man Sterbehilfe für moralisch unzulässig ansieht. Somit kann der moralische Wert eines Werkes, unabhängig davon ob er positiv oder negativ ist, zur Tiefgründigkeit des Werkes beitragen.

Kommen wir zu einem Beispiel einer Beurteilung auf Basis des affektiv-antwortabhängigen Maßstabes: Unterhaltsamkeit. Bezeichnen wir ein Werk als unterhaltsam, attestieren wir ihm einen positiven antwortabhängigen Wert. Charakteristischerweise ist ein Werk unterhaltsam, wenn es die Aufmerksamkeit seines Publikums gewinnt, deren Interesse erweckt und es amüsiert, überrascht oder fasziniert. Ob ein Werk unterhaltsam ist, kann nun teilweise von seinem moralischen Wert abhängen, welcher sowohl positiv als auch negativ sein kann.

Denken wir zunächst zur Veranschaulichung an Jane Austens Roman *Emma*. Hierin wird das Versteckspiel von Frank Churchill und seiner heimlichen Verlobten verurteilt, weil er dadurch Emmas Gefühle zu verletzen. Dies kann man als moralisch richtige Haltung ansehen. Das Buch wird nun u.a. genau wegen der Art und Weise, wie dieses Versteckspiel dargestellt wird, unterhaltsam.

Oder erinnern wir uns an das bereits eingangs zitierte Beispiel *Lolita*. *Lolita* ist ein unterhaltsamer Roman. Seine Unterhaltsamkeit scheint untrennbar mit der problematischen bzw. unklaren Haltung des Werkes zur Pädophilie von Humbert verknüpft zu sein. Würde diese klar verurteilt, würde das Werk also eine moralisch richtige Haltung ausdrücken, wäre es viel weniger unterhaltsam.

Greifen wir drittens ein Beispiel einer kunsthistorischen Bewertung auf, nämlich Originalität. Unterscheidet sich ein Werk ausreichend von seinen Vorgängern, kann es dadurch das

[49] Siehe Lamarque (2006: 130).

positive Werturteil der Originalität verdienen.[50] Auch die Originalität eines Werkes kann nun von seinem moralischen Wert abhängen.[51]

De Sades *Die 120 Tage von Sodom und Gomorra* ist ein originelles Werk. Seine Originalität hängt zumindest teilweise davon ab, dass in diesem Buch Gewalt auf eine bisher nicht dagewesene Art und Weise dargestellt und verherrlicht wird. Dies ist aber genau das, was wir dem Werk in moralischer Hinsicht vorwerfen.

Oder denken wir an Sands *Leila*. Dieser Roman verdankt seine Originalität teilweise der Art und Weise wie die weibliche Protagonistin dargestellt wird, nämlich als eine Frau mit einem differenzierten Innenleben.[52] Solch eine Darstellung einer Frau war für die damalige Zeit neuartig. Eine Frau so darzustellen stellt aber genau einen positiven moralischen Wert dar.

Weitere Beispiele wären wünschenswert, um diese Einflussnahme ausführlicher zu illustrieren: Kann ein Werk beispielsweise durch seine moralisch richtige Haltung banal und sentimental werden? Man denke beispielsweise an Rosamunde-Pilcher-Romane. Oder kann eine unmoralische Haltung die Ironie eines Werkes untergraben? Man denke beispielsweise an Elis *American Psycho*. Die bisherigen Beispiele genügen jedoch die Möglichkeit zu zeigen, dass der moralische Wert den Gesamtwert von Kunstwerken indirekt kontextualistisch beeinflussen kann.

Zwei Fragen sind im Hinblick auf diese indirekte kontextualistische Einflussnahme noch offen. Erstens beeinflusst der moralische Wert eines Werkes dessen Gesamtwert *immer* indirekt oder gibt es Fälle, bei welchen ein Werk zwar moralisch bewertbar ist, diese Bewertung aber keine andere tangiert? Diese Möglichkeit sollte eingeräumt werden. Stellen wir uns ein Landschaftsgemälde vor, dessen Titel „Landschaftsgemälde (Töten ist falsch)“ lautet. In dem Titel wird eine moralische Haltung explizit ausgedrückt und somit scheint das Werk einen moralischen Wert zu haben. Dieser interagiert jedoch mit keinem Aspekt in dem Bild und kann somit keine Bewertung des eigentlichen Gemäldes beeinflussen.

Zweitens kommt die Frage auf, ob man allgemeine Regeln angeben kann, *wann* ein moralischer Wert eines Werkes *wie welche* Bewertungen beeinflusst. Es ist nicht möglich, solche Regeln anzugeben. Dies kann man unter Rückgriff auf Sibleys Theorie der ästhetischen Ausdrücke begründen. Das Besondere an ästhetischen Ausdrücken ist, dass sie nicht positiv durch Bedingungen bestimmt sind.[53] Vielmehr braucht es Geschmack („taste“), um sie richtig zuzuschreiben.[54] Nur im direkten Kontakt mit einem Kunstwerk kann entschieden werden, welche ästhetischen Eigenschaften es hat. Nimmt man diesen Gedanken ernst, kann es auch im Hinblick auf die Interaktion zwischen moralischem Wert und ästhetischer Eigenschaft eines Werkes keinen allgemein bestimmbaren Bedingungen geben. Die indirekte Einflussnahme des moralischen Werts auf den Gesamtwert kann also am besten als unsystematisch beschrieben werden.

Der moralische Wert eines Werkes kann also den Gesamtwert eines Werkes durch eine kognitive, affektiv-antwortabhängige oder kunsthistorische Bewertung indirekt beeinflussen. Diese Beeinflussung fällt kontextualistisch aus, wobei die Möglichkeit eingeräumt wird, dass es zu keiner Interaktion kommt. Auch kann sie am besten als unsystematisch beschrieben werden.

---

[50] Für Argumente, das die Originalität eines Werkes dessen Wert beeinflusst, siehe Hoaglund (1976); Osborne (1979); Sibley (1985); Kieran (2005).

[51] Dieses Argument ist inspiriert durch eine kurze Anmerkung Gauts, auch wenn Gaut sich gegen den Kontextualismus und für den Ethizismus ausspricht, siehe Gaut (2007: 60).

[52] Siehe Schlientz (2008).

[53] Siehe Sibley (1959: 424).

[54] Siehe Sibley (1959: 421).

### *5.2 Argumente gegen eine direkte Einflussnahme*

Kann der moralische Wert eines Werkes dessen Gesamtwert auch direkt beeinflussen, d.h. sollte die pluralistische Werttheorie auch ein moralisches Bewertungskriterium als eigenständiges Beurteilungskriterium beinhalten? Eine Möglichkeit, für eine direkte Einflussnahme zu argumentieren, kann darin bestehen ein allgemeines Angemessenheitskriterium für Bewertungsmaßstäbe anzugeben, um dann zu zeigen, dass der moralische Maßstab dieses Kriterium erfüllt. Booths Freundschaftsargument und Dickies Argument des essentiellen Teils scheinen solch ein Vorhaben zu verfolgen.

Laut Dickie ist jede Kritik an einem essentiellen Teil eines Kunstwerkes eine angemessene Kritik an dem Kunstwerk. Zweitens geht er davon aus, dass die moralische Perspektive (Haltung) eines Kunstwerkes ein essentieller Teil dieses Werkes sei. Somit ist es angemessen die moralische Haltung eines Kunstwerkes zu kritisieren.[55]

Die grundlegende Schwierigkeit besteht darin zu verstehen, was genau ein essentieller Teil eines Kunstwerkes ist. Dickie zu Folge ist etwas ein essentieller Teil eines Kunstwerkes, wenn das Werk in hohem Maße sehr verändert würde, würde der Teil entfernt.[56] Doch ab wann ist eine Veränderung bedeutend genug? Man könnte sehr strikt argumentieren, dass jede Veränderung an einem Werk bedeutend ist. So gesehen, wäre alles an einem Kunstwerk essentiell. Somit wäre die genaue Anzahl von „i“s in einem Gedicht ein essentieller Teil des Gedichts und es wäre angemessen das Gedicht dahingehend zu bewerten.[57] Ob dies wirklich eine relevante Beurteilung des Gedichts ist, darf bezweifelt werden. Sieht man nicht jede Veränderung eines Kunstwerkes als relevant an, entsteht das notorische Problem zu erklären, wie viel Veränderung bedeutend genug ist. Sucht man nach einem „objektiven“ Kriterium der Angemessenheit eines Bewertungsmaßstabes, ist dies nicht wünschenswert.

Booths Argument baut auf der Annahme auf, wir würden den implizierten Autor eines Buches wie einen Freund behandeln. Wenn wir nun den impliziten Autor wie einen Freund beurteilen, handelt es sich dabei um eine angemessene Kritik. Da wir Freunde moralisch bewerten, ist eine moralische Kritik angemessen.[58]

Das Angemessenheitskriterium dieses Arguments ist problematisch. Vieles, was wir von (realen) Freunden erwarten, erwarten wir nicht von implizierten Autoren, beispielsweise dass sie uns zu unserem Geburtstag anrufen. Einem impliziten Autor dies vorzuwerfen, wäre sich keine angemessene Kritik. Außerdem gibt es Eigenschaften, die wir an impliziten Autoren schätzen, jedoch nicht unbedingt bei realen Freunden. Wir suchen nicht unbedingt einen Freund, der viel witziger, sprachgewandter, und intelligenter ist als wir selbst, auch wenn wir solche Eigenschaften bei einem implizierten Autor durchaus schätzen.[59] Außerdem würde der Verweis auf die Freundschaft wahrscheinlich zu einer sehr subjektiven Werttheorie führen, außer man verweist auf das Konzept der vollkommenen Freundschaft. Hier entsteht aber das Problem, dass wir zwar teilweise implizite Autoren wie Freunde behandeln, selten aber wie vollkommene Freunde.

Die grundlegende Schwierigkeit bei dieser Art von Argumenten ist es ein gutes Angemessenheitskriterium für Kunstbewertungskriterien zu finden, welches weder zu strikt noch zu weit gefasst ist. Auch darf bezweifelt werden, ob ein moderater Autonomist je ein Angemessenheitskriterium akzeptieren würde, welches das moralische Kriterium erfüllt.

---

[55] Siehe Dickie (1964: 64).

[56] Siehe Dickie (1964: 64).

[57] Siehe Gaut (1998: 190).

[58] Siehe Booth (1988: 169 ff.). Die Rekonstruktion des Arguments basiert auf Gauts Analyse desselben, siehe Gaut (2007: 109). Booth konzentriert sich literarische Werke. Gegeben falls könnte man das Argument auf Kunstwerke im Allgemeinen erweitern.

[59] Siehe Gaut (2007: 113).

Stellen wir diese Argumentationsstrategie zurück. Was würde es bedeuten, wäre der moralische ein eigenständiger Bewertungsmaßstab? Zunächst wäre der moralische Wert eines Kunstwerkes immer relevant für dessen Gesamtbewertung, wenn das betreffende Kunstwerk moralisch bewertbar ist. Ein moralisches Defizit würde sich immer negativ und ein positiver moralischer Wert würde sich immer positiv auf den Gesamtwert auswirken. In dieser Hinsicht würden wir dem Ethizismus Recht geben müssen.[60]

Das Problem, welches auch der Ethizismus sieht, ist, dass man nicht möchte, dass der moralische Wert eines Kunstwerkes *immer* relevant ist. Erinnern wir uns an das obige Beispiel des Landschaftsgemäldes. Oder denken wir an ein Buch mit einem Appendix von moralischen Wahrheiten, die vollkommen losgelöst vom Rest und Inhalt des eigentlichen Buches sind.[61] Der Appendix drückt explizit moralische Haltungen aus. Somit ist das Buch insoweit moralisch bewertbar und der positive moralische Wert müsste den Gesamtwert positiv beeinflussen. Man hätte eine relativ einfache Möglichkeit gefunden den Gesamtwert eines Werkes zu optimieren.

Um solche Probleme zu umgehen, schlägt Gaut vor, ein zusätzliches Kriterium der ästhetischen Relevanz zu berücksichtigen, welches Auskunft darüber gibt, wann ein moralischer Wert zu berücksichtigen ist.[62] Zunächst entsteht hier die Schwierigkeit ein Kriterium anzugeben, welches ein echtes Kriterium ist und nicht nur als Alibi für eine bereits vorab getroffene intuitive Einschätzung über die Relevanz des moralischen Werts fungiert. Abgesehen davon braucht man ein solches Kriterium wirklich oder wäre es nicht methodisch ratsamer, ohne ein solches auszukommen?

Man kann auf ein solches Kriterium verzichten, wenn man nicht davon ausgeht, dass der moralische Wert den Gesamtwert direkt beeinflusst. Wie bereits ausgeführt, beeinflusst der moralische Wert eines Werkes den Gesamtwert nicht immer indirekt. Wenn es zu keiner indirekten Einflussnahme kommt, wie im Falle des Appendix mit moralischen Botschaften, hat man dadurch bereits eine Art indirektes Kriterium der ästhetischen Relevanz.

## 6. Konklusion

Ziel dieses Artikels war es die Frage zu beantworten, ob Kunstwerke moralisch bewertbar sind und ob und wie der moralische Wert eines Werkes dessen Gesamtwert beeinflusst. Eine Frage, die zu der grundlegenderen Frage nach einer Werttheorie für Kunstwerke führte. Zwei Möglichkeiten, wie eine moralische Bewertung von Kunstwerken möglich ist, wurden aufgezeigt und somit wurde gegen den radikalen Autonomismus argumentiert. Auch der moderate Autonomismus überzeugt nicht, da er eng mit einer monistischen Werttheorie verbunden ist, welche zu normativ aufgeladen ist. Eine pluralistische Werttheorie erfasst die Vieldimensionalität der Kunstkritik besser und eröffnet die Möglichkeit, dass der moralische Wert relevant für die Gesamtbewertung ist. Indem berücksichtigt wird, dass der moralische Wert auf Basis pluralistischen Werttheorie den Gesamtwert nicht nur direkt, sondern auch indirekt beeinflussen kann, wird die Möglichkeit einer kontextualistischen Einflussnahme des moralischen Werts auf den Gesamtwert denkbar. Anhand von beispielhaften Bewertungen wurde für solch einen indirekten Kontextualismus argumentiert. Weitere Bespiele könnten ausgearbeitet und dargestellt werden. Entscheidend ist jedoch, dass durch den Gedanken der indirekten Einflussnahme den kontextualistischen Zug erklären kann, was man allein auf einer direkten Einflussnahme nicht vermag. Folglich stellt der indirekte Kontextualismus unsere Antwort auf die Frage nach dem Einfluss des moralischen Werts auf den Gesamtwert eines Kunstwerkes dar.

---

[60] Siehe Gaut (2007: 55).

[61] SieheKieran (2001: 29).

[62] Siehe Gaut (2007: 85).

**Lisa Katharin Schmalzried**
Universität Luzern
lisa.schmalzried@unilu.ch

## Literatur

Anderson, J. & Dean, J. 1998: „Moderate Autonomism", *British Journal of Aesthetics* 38, 150–166.

Beardsley, M. 1958: *Aesthetics: Problems in the Philospohy of Criticism*. New York: Harcourt Brace Jovanovich.

– 1969: „Aesthetic Experience Regained", *Journal of Aesthetics and Art Criticism* 28, 3–11.

– 1979: „In Defense of Aesthetic Value", *Proceedings of the Aristotelian Society* 52, 723–749.

– 1991: „Aesthetic Experience", in R. Smith und A. Simpson (Hrg.): *Aesthetics and Arts Education*, Chicago: University of Illinois Press, 72–84.

Beardsmore, R.W. 1971: *Art and Morality*. London: Mac Millan, 1971.

– 1973: „Learning from a Novel", *Royal Institute of Philosophy Lectures*, 23–46.

Bell, C. 1914: *Art*. London: Chatto & Windus.

Booth, W. 1988: *The Company We Keep*. London: University California Press.

– 1998: „Why Banning Ethical Criticism is a Serious Mistake", *Philosophy and Literatrue* 22, 366–393.

Carroll, N. 1985: „Formalism and Critical Evaluation", in P. McCormick (Hrg.): *The Reasons of Art*, Montreal: University of Ottawa Press, 327–335.

– 1986: „Art and Interaction", *The Journal of Aesthetics and Art Criticism* 45, 57–68.

– 1996: „Moderate Moralism", *British Journal of Aesthetics* 36, 223–238.

– 1998: „Moderate Moralism versus Moderate Autonomism", *British Journal of Aesthetics* 40, 419–424.

– 2000: „Art and Ethical Criticism. An Overview of Recent Directions of Research", *Ethics* 110, 350–387.

– 2001: „Four Concepts of Aesthetic Experience", in *Beyond Aesthetics*, Cambridge: Cambridge University Press, 41–62.

– 2002: „The Wheel of Virtue", *The Journal of Aesthetics and Art Criticism 60*, 3–26.

– 2008: „Narrative and Ethical Life", in G. Hagberg (Hrg.): *Art and Ethical Criticism*, Oxford: Blackwell, 35–62.

Currie, G. 1998: „Realism of Character and the Value of Fiction", in J. Levinson (Hrg.): *Aesthetics and Ethics*, Cambridge : Cambridge University Press, 161–181.

Devereaux, M. 1998: „Beatuy and Evil: The Case of Leni Riefenstahl's Triumph of the Will", in J. Levinson (Hrg.): *Aesthetics and Ethics–Essays at the Intersection*, Cambridge: Cambridge University Press, 227–256.

– 2004: „Moral Judgement and Works of Art: The Case of Narrative Literature", *The Journal of Aesthetics and Art Criticism* 62, 3–11.

Dickie, G. 1964: „The Myth of the Aesthetic Attitude", *American Philosophical Quarterly* 1, 56–65.

– 1965: „Beardsley's Phantom Aesthetic Experience", *The Journal of Philosophy* 62, 129–136.

– 1985: „Evaluating Art", in *British Journal of Aesthetics* 25, 3–16.

Diffey, T. 1986: „The Idea of Aesthetic Experience", in M. Mitias (Hrg.): *Possibility of the Aesthetic Experience*, Dordrecht: Martinus Nijhoff Publishers, 3–12.

Dziemidok, B. 1993: „Artistic Formalism: Its Achievements and Weaknesses", *The Journal of Aesthetics and Art Criticism* 51, 185–193.

Gass, W. 1987: „Goodness Knows Nothing of Beauty", *Harper's* 274, 37–44.

Gaut, B. 1998: „The Ethical Criticism of Art", in J. Levinson (Hrg.): *Aesthetics and Ethics: Essays at the Intersection*, Cambridge: Cambridge University Press, 182–203.

– 2003: „Art and Knowledge", in J. Levinson (Hrg.): *The Oxford Handbook of Aesthetics*, Oxford: Oxford University Press, 436–450.

– 2006: „Art and Cognition", in M. Kieran (Hrg.): *Contemporary Debates in Aesthetics and the Philosophy of Art*, Oxford: Blackwell, 115–142.

– 2007: *Art, Emotion and Ethics*. Oxford: Oxford University Press.

Giovanelli, A. 2007: „The Ethical Criticism of Art: A New Mapping of the Territory", *Philosophia* 35, 117–127.

Harold, J. 2008: „Immoralism and the Valence Constraint", *British Journal of Aesthetics* 48, 45–64.

Hoaglund, J. 1976: „Originality and Aesthetic Value", *British Journal of Aesthetics* 16, 46–55.

Jacobson, D. 1997: „In Praise of Immoral Art", *Philosophical Topics* 25, 155–199.

– 2006: „Ethical Criticism and The Vice of Moderation", in M. Kieran (Hrg.): Contemporary *Debates in Aesthetics and the Philosophy of Art*, Oxford : Blackwell, 342–355.

Kant, I. 1963: *Kritik der Urteilskraft*. Stuttgart: Reclam.

– 1977: „Über den Gemeinspruch", in W. Weischedel (Hrg.), *Immanuel Kant: Werke in zwölf Bänden. Band 11*, Frankfurt a.M.: Suhrkamp.

Kieran, M. 1996: „Art, Imagination, and the Cultivation of Morals", *Journal of Aesthetics and Art Criticism* 54, 337–351.

– 2001: „In Defense of the Ethical Evaluation of Narrative Art", *British Journal of Aesthetics* 41, 26–38.

– 2002: „On Obscenity: The Thrill and Repulsion of the Morally Prohibeted", *Philosophy and Phenomenological Research* 64, 31–55.

– 2003: „Forbidden Knowledge: the Challenge of Immoralism", in J. Bermudez und S. Gardner (Hrg.): *Art and Morality*, London; New York: Routledge, 56-73.

– 2005: *Revealing Art*. Oxon, New York: Routledge.

Lamarque, P. 2006: CCognitive Values in the Arts: Marking the Boundaries" in M. Kieran (Hrg.): *Contemporary Debates in Aesthetics and the Philosophy of Art*, Oxford: Blackwell, 127–139.

Lamarque, P. und Olsen, S.H. 1994: *Truth, Fiction, and Literature–A Philosophical Perspective*. Oxford : Clarendon Press.

Nabokov, V. 1999: *Lolita*. Reinek: rororo.

Novitz, D. 1980: „Fiction, Imagination and Emotion", *The Journal of Aesthetics and Art Criticism* 38, 279–288.

Nussbaum, M. 1990: „Finely Aware and Richly Responsible", in *Love's Knowledge*, Oxford: Oxford University Press, 148–167.

Matravers, D. 2003: „The Aesthetic Experience", *British Journal of Aesthteics* 43, 158–174.

Osborne, H. 1979: „The Concept of Creativity in Art", *British Journal of Aesthetics* 19, 224–231.

Pole, D. 1962: „Morality and the Assessment of Literature", *Philosophy* 37, 193-207.

Posner, R. 1997: „Against Ethical Criticism“, *Philosophy and Literature* 21, 1–27.

Price, K. 1979: „What Makes an Experience Aesthetic?“, *British Journal of Aethetics* 19, 131–143.

Putnam, H. 1975–76: „Literature, Science and Reflection“, *New Literary History* 7, 483–491.

Schlientz, G. 2008: „Nachwort zu Lelia“, in G. Sand: *Lelia*, München: DTV.

Sibley, F. 1959: „Aesthetic Concepts“, *The Philosophical Review* 68, 421–450.

– 1965: „Aesthetic and Nonaesthetic“, *The Philosophical Review* 72, 135–159.

– 1985: „Originality and Value“, *British Journal of Aesthetics* 25, 169–184.

Sidney, P. 1966: *The Defense of Poesy*. Oxford: Oxford University Press.

Stecker, R. 2005[1]: „The Interaction of Ethical and Aesthetic Value“, *British Journal of Aesthetics* 45, 138–150.

– 2005[2]: *Aesthetics and the Philosophy of Art*. Lanham, MD: Rowman & Littlefield.

Stolnitz, J. 1960: *Aesthetics and Philosophy of Art Criticism*. Cambridge, Mass.: The Riverside Press.

– 1992: „On the Cognitive Triviality of Art”, *British Journal of Aesthetics* 32, 191–200;

Tolstoi, L. 1998: *Was ist Kunst?*. Schutterwald: Wissenschaftlicher Verlag, 1998.

Walton, K. 1994: „Morals in Fiction and Fictional Morality“, *Proceedings of the Aristotelian Society* 68, 27–50.

Wollheim, R. 1980: „Seeing-as, seeing-in, and pictoral representation“, in *Art and its objects*, Cambridge: Cambridge University Press, 205–226.

Zangwill, N. 2001: „Aesthetic/Sensory Dependence“, in *The Metaphysics of Beauty*, London: Cornell University Press, 127–146.

# Praemotio physica und leibnizianischer Molinismus

Ruben Schneider

In der analytischen Religionsphilosophie ist die systematische Untersuchung der im Gnadenstreit mit dem Molinismus konkurrierenden thomistischen Theorie der *praemotio physica* ein veritables Forschungsdesiderat. Hier soll der Ansatz einer systematischen Rekonstruktion der thomistischen Lehre versucht werden – und im Zuge dessen soll aufgewiesen werden, dass der Molinismus gegenüber kausaltheoretischen Einwänden von thomistischer Seite zur sog. *Essence-solution* der *Grounding objection* verpflichtet ist. Hierbei soll sowohl aus historischen als auch aus systematischen Gründen eine von Leibniz inspirierte Version der *Essence-solution* vorgeschlagen werden.

Während sich der Molinismus in der gegenwärtigen analytischen Religionsphilosophie eines großen Interesses erfreut,[1] findet die in seiner Entstehungszeit mit ihm rivalisierende thomistische Theorie der *praemotio physica* bzw. des *concursus praevius* kaum noch Beachtung.[2] Einen Grund dafür stellen mit Sicherheit die erbitterten exegetischen Kämpfe dar, die bis Mitte des 20. Jahrhunderts um die Frage geführt wurden, ob Thomas von Aquin selbst in seinen Werken eine „praemotio physica" vertrete oder nicht. Dieser Frage kann hier nicht nachgegangen werden.[3] Im Folgenden wird vielmehr der (notwendigerweise höchst

[1] Cf. Flint (1998) und für ausführliche Literatur Perszyk (2011), 3–19 and 303 – 317.

[2] Neuere Ausnahmen sind Osborne, Thomas M. Jr. (2006), „Thomist Premotion and Contemporary Philosophy of Religion", in: *Nova et Vetera*, English Edition, Vol. 4, No. 3; Kondoleon, Theodore J. (1983), „The Free Will Defense", in: *The Thomist* 47, 1–42; Long, Steven A. (2002), "Providence, liberté et loi naturelle", in: *Revue Thomiste* 102, 355–406; und Flint, Thomas P. (1988), „Two Accounts of Providence", in: Morris, T. (ed.), *Divine and Human Action. Essays in the Metaphysics of Theism*, Ithaca/New York, 147-181.

[3] Cf. als prominentesten Opponenten der Prämotionslehre: Stufler, Johannes (1923), *Divi Thomae Aquinatis doctrina de Deo operante in omni operatione naturae creatae praesertim liberi arbitrii*, Innsbruck; ders. (1936), *Gott, der erste Beweger aller Dinge. Ein neuer Beitrag zum Verständnis der Konkurslehre des hl. Thomas von Aquin*, Innsbruck; u.a. Zur intensiven Kontroverse um Stuflers Thesen vgl. Berger, David (2005), *In der Schule des hl. Thomas von Aquin*, Bonn, 210-219. Cf. auch Jorissen, H. (1988), „Schöpfung und Heil. Theologiegeschichtliche Perspektiven zum Vorsehungsglauben nach Thomas von Aquin", in: Schneider, T., Ullrich, L. (Hrsg.), *Vorsehung und Handeln Gottes*, 94–108; Pesch, O. H. (1963), „Freiheitsbegriff und Freiheitslehre bei Thomas von Aquin und Luther", in: *Catholica* 17, 197–244; Pesch, O. H. (1962), „Philosophie und Theologie der Freiheit bei Thomas von Aquin in quest. disp. 6 De malo. Ein Diskussionsbeitrag", in: *Münchner Theologische Zeitschrift* 13, 1–25; Congar, Y. M.-J. (1934), „‚Praedeterminare' et ‚Praedeterminatio' chez Saint Thomas", in: *Revue des sciences philosophiques et théologiques* 23, 363–371; Pesch, O. H. (1967), *Theologie der Rechtfertigung bei Martin Luther und Thomas von Aquin*, 862, n. 32; Sertillanges, A. D. (1954), *Der hl. Thomas von Aquin*, 2. Aufl., 345–348; Siewerth, G. (1954), „Einführung", S. 107–111, in: *Thomas von Aquin. Die menschliche Willensfreiheit*, 7–134; cf. Jonkenheere, Anna-Maria (1995), *God als vrij-geleide voor goed leven en handelen. Thomas van Aquino's systematische teksten over de predestinatie*. Theologische Faculteit Tilburg Studies 24. Tilburg, 13, Fn. 46. – Klassische Proponenten der *praemotio physica* sind vorallem Garrigou-Lagrange, R. (1936), *God: His Existence and His Nature*, tr. by B. Rose, 2 v., St. Louis; Dummermuth, A.M. (1886), *S. Thomas et Doctrina Praemotionis Physicae*, Paris; ders. (1895), *Defensio doctrinae S. Thomae Aquinatis de praemotione physica*, Paris; Del Prado, N. (1907), *De gratia et libero arbitrio, pars secunda: Concordia liberi arbitrii cum divina motione juxta S. Augustinum et D. Thomam*, Freiburg.

unvollkommene) Versuch unternommen, die Lehre der physischen Vorherbewegung, wie sie von Domingo Bañez (1528–1604) und seinen Schülern verfochten wurde, in systematischer Weise zu rekonstruieren und damit der analytischen Debatte zugänglich zu machen. Auf dieser Basis soll sodann ein thomistischer Einwand gegen die molinistische Lehre vom reinen *concursus simultaneus et indifferens* dargestellt werden – und aufgezeigt werden, wie eine molinistische Antwort auf diesen spezifischen kausaltheoretischen Einwand zugleich eine Form der „Essence solution" der berüchtigten „Grounding objection" gegen den Molinismus[4] mit sich führt.

## 1. Concursus simultaneus und concursus praevius

### *1.1 Definitionen und Vorklärungen*

Als Basis einer strukturalen Rekonstruktion der Lehre von der *praemotio physica* seien zunächst einige wesentliche Definitionen und Bemerkungen gegeben.

**(1.1) Definition.** Der Term *concursus simultaneus* bzw. *concomitans* denotiert das wirkursächliche Verhältnis der Erstursache (*causa prima*) zu Tätigkeit (*actio*) und Wirkung (*effectus*) der Zweitursachen (*causae secundae*), durch welches die Erstursache der Zweitursache das *Sein* ihrer Tätigkeit und das *Sein* der Wirkung ihrer Tätigkeit mitteilt. Dieser *concursus* wird „simultan" bzw. „konkomitant" genannt, weil er mit der Ausübung der Tätigkeit der Zweitursache und der Hervorbringung ihrer Wirkung parallel einhergeht bzw. sie „begleitet".[5]

**(1.2) Bemerkung.** Der *concursus simultaneus* muss unter zwei Rücksichten betrachtet werden:

**1°** Aus der Perspektive Gottes stellt der simultane Concursus nichts anderes dar als das göttliche Wirken, durch das die Erstursache wirkursächlich das Sein der zweitursächlichen Tätigkeit und ihrer Wirkung hervorbringt.

**2°** Aus der Perspektive des Geschöpfs ist der simultane Concursus die Tätigkeit und ihre Wirkung selbst, *sub ratione entis*.[6]

**(1.3) Definition.** Der Term *concursus praevius* bzw. *praemotio physica* (Vorherbewegung) denotiert das wirkursächliche Verhältnis der Erstursache zur Zweitursache selbst, durch welches die Erstursache die Zweitursache zum Wirken, d.h. zur ursächlichen Hervorbringung ihrer Tätigkeit bringt. Die praemotio wird „prae-" genannt im Sinne einer a-temporalen Ordnungsrelation und „physica", um anzuzeigen, dass es sich nicht um einen finalursächlichen Einfluss handelt.[7]

**(1.4) Bemerkung.** Auch bezüglich der *praemotio physica* muss zwischen zwei Rücksichten unterschieden werden:

**1°** Aus der Perspektive Gottes ist die physische Vorherbewegung nichts anderes als das göttliche Wirken selbst, durch das die Erstursache wirkursächlich die Zweitursache (als Instrumentalursache) zur Tätigkeit bringt.

**2°** Aus der Perspektive des Geschöpfs ist die physische Vorherbewegung eine transeunte Partizipation an der göttlichen Allmacht, kraft derer die Zweitursache sich selbst zum Wirken

[4] Cf. Flint (1998), 123-126; Hasker (1989); Adams (1977), 30.
[5] Cf. Gredt (1953), 247; Dummermuth (1886), 17-19.
[6] Cf. Gredt (1953), 247.
[7] Cf. S.th. I, q. 82, a. 2, 4; I-IIae, q. 9, a. 1; q. 10, a. 2; cf. O'Brien (1981), 671.

bringt (sich vom *actus primus* [Akt des Daseins] in den *actus secundus* [hier: Akt der *operatio*] überführt).[8]

Die Lehre von *concursus simultaneus* und *praemotio physica* setzt als argumentativen Hintergrund die thomistische Lehre von Akt und Potenz voraus. Das Akt-Potenz-Schema steht im Zentrum der thomistischen Seinstheorie und Metaphysik und besitzt eine ungeheure Explikationskraft. Sie kann im vorliegenden Rahmen nicht eingehend dargestellt werden – daher sollen nur einige grundlegende architektonische Merkmale dieser Theorie wiedergegeben werden, soweit sie für den Fortgang vonnöten sind:

**(1.5) Akt und Potenz**:[9]

**1°** Akt und Potenz stellen *modi* des Seins dar.

**2°** Im geschaffenen Bereich sind Akt und Potenz keine monadischen Qualifikationen, sondern bezeichnen relationale bzw. korrelative Verhältnisse: $A_n$ ist ein Akt (oder „im Akt“) relativ zu einem $P_n$, wobei $P_n$ sodann eine Potenz (oder „in Potenz“) relativ zu $A_n$ ist. $P_n$ selbst jedoch kann wiederum ein Akt $A_{n-1}$ sein, relativ zu einem Dritten, $P_{n-1}$, welches relativ zu $A_{n-1}$ Potenz (oder „in Potenz“) ist, usw. (Cf. ScG, II, c. 77.) Analoges gilt vom Begriffspaar „Form“ und „Materie“.

**3°** Man erhält damit eine *lineare*, *irreflexive* und *asymmetrische Ordnung* aus Akt-Potenz-Korrelationen: Ein $P_1$ steht in Potenz zu einem korrelierten Akt $A_1$, das Paar $(P_1,A_1)$ wiederum kann eine Potenz $P_2$ in Bezug auf einen höheren Akt $A_2$ darstellen, was selbst in Potenz $P_3$ relativ zu einem höheren Akt $A_3$ stehen kann, usw. (Cf. ScG, II, c. 54.) Jeder Akt komplettiert dabei nur die in dieser Ordnung direkt unter ihm stehende Potenz (cf. ScG, II, c. 73). Der untere Abschluß dieser Hierarchie ist die *materia prima* und der obere Abschluß der *actus purus*.

**4°** Ein Akt oder „im Akt“ zu sein bedeutet, von höherem ontologischem Rang zu sein als die korrelative Potenz: Ein Akt ist ontologisch reichhaltiger als die zugehörige Potenz, er hat eine gewisse ontologische Saturiertheit erreicht. In der in 3° genannten Hierarchie sind damit jeweils höhere Aktstufen reichhaltiger/saturierter als niedrigere. Der *actus purus* besitzt die höchste Saturiertheit (Vollkommenheit aller Vollkommenheiten).

**5°** Für jede Potenz $P_i$ gilt: Ihr korrelativer Akt $A_i$ ist extrinsisch zu ihr (Realdistinktion zwischen Akt und Potenz).

**6°** Die Potenz ist das „Prinzip der Limitation“: Im Allgemeinen kann eine Potenz nicht den vollen zugehörigen Akt aufnehmen (dies ist lediglich bei den „materiellen Formen“ möglich).

### *1.2 Der concursus simultaneus*

**(2.1) Satz.** Gott als Erstursache bringt im *concursus simultaneus* wirkursächlich und unmittelbar das Sein der zweitursächlichen Tätigkeit und ihrer Wirkungen hervor.

**Begründung.**[10] Es gilt die vollständige Disjunktion: Entweder wird das Sein der zweitursächlichen Tätigkeit (a) durch die Zweitursache selbst, oder (b) durch eine durch Zweitursachen vermittelte Tätigkeit Gottes oder (c) durch eine unmittelbare Tätigkeit Gottes hervorgebracht.

Ad (a): Das Sein der zweitursächlichen Tätigkeit wird nicht durch die Zweitursache selbst hervorgebracht. Sei x eine geschaffene Zweitursache (Zx).[11] Dann gilt:

---

[8] Cf. Gredt (1953), 251.

[9] Cf. zum Folgenden Schneider (2007), 222–235.

[10] Cf. Gredt (1953), 247f.; Feldner (1890), 62–84.

[11] Hierbei sei x auf die Menge der materiellen Substanzen beschränkt. Die in der thomistischen Theorie auch auftretenden *substantiae creatae spirituales*, die substanziell unveränderlich, aber akzidentell

(1) Keine geschaffene Zweitursache ist reiner Akt, *actus purus*. [$\forall x(Zx \rightarrow \neg APx)$] (Def. des Geschaffenen.)

(2) Alles, was nicht reiner Akt ist, ist nicht essentiell im Akt. [$\forall x(\neg APx \rightarrow \neg EAx)$] (mit 1.6, 4°)[12]

(3) Also ist keine geschaffene Zweitursache essentiell im Akt. [$\forall x(Zx \rightarrow \neg EAx)$]

(4) Wenn x nicht essentiell im Akt ist, ist es (für einen beliebigen Akt $A_i$) zu einem ersten Moment *t* nicht realisiert, dass x in Relation zu $A_i$ im Akt ist [$\forall x(\neg EAx \rightarrow \neg R_{t,i}Ax)$] (lies: $R_{t,i}Ax$ = „es ist (für einen beliebigen Akt $A_i$) zu einem ersten Moment *t* realisiert, dass x in Relation zu $A_i$ im Akt ist". – Eventuelle vorherige Akte $A_{i-1}$ sind bereits an ihre korrespondierende Potenz $P_{i-1}$ gebunden und stehen im Paar $(P_{i-1},A_{i-1})$ in Potenz zu $A_i$, vgl. 1.6, 3°).

(5) Also: Für alle Zweitursachen x ist es (für einen beliebigen Akt $A_i$) zu einem ersten Moment *t* nicht realisiert, dass x in Relation zu $A_i$ im Akt ist. [$\forall x(Zx \rightarrow \neg R_{t,i}Ax)$]

(6) Etwas kann (für einen beliebigen Akt $A_i$) zu einem ersten Moment *t* nur dann (wirk)ursächlich tätig sein, wenn es zu *t* in Relation zu $A_i$ im Akt ist. [$\forall x(R_{t,i}Ux \rightarrow R_{t,i}Ax)$][13]

(7) Also: Keine Zweitursache ist (für einen beliebigen Akt $A_i$) zu einem ersten Moment *t* in Relation zu $A_i$ (wirk)ursächlich tätig. [$\forall x(Zx \rightarrow \neg R_{t,i}Ux) \leftrightarrow \neg\exists x(Zx \wedge R_{t,i}Ux)$]

(8) Sei $A_i$ nun eine Tätigkeit $\tau$. Wenn x eine neue Tätigkeit $\tau$ zu *t* hervorbringt, dann bringt x $\tau$ vom Nichtsein ins Sein. [$\forall\tau\forall x(H_t x\tau \leftrightarrow NSx\tau)$] („radical novelty")

(9) Bei diesem Übergang vom Nichtsein zum Sein muss x bezüglich $\tau$ (wirk)ursächlich tätig sein. [$\forall\tau\forall x(NSx\tau \rightarrow R_{t,\tau}Ux)$] (Mit 1.6, 4°, 5° und dem metaphysischen Kausalprinzip: Neues Sein entsteht nicht durch spontane Emanation aus Nichts.)

(10) Also kann keine Zweitursache anfänglich ihre eigene Tätigkeit hervorbringen (und damit auch nicht ihr Sein und das ihrer Wirkung). [$\forall\tau\forall x(Zx \rightarrow \neg H_t x\tau)$]

Ad (b): Es folgt direkt aus (a), dass Gott nicht vermittelt durch Zweitursachen das Sein ihrer Tätigkeiten hervorbringen kann, wenn die Zweitursachen selbst noch gar nicht tätig sind.

Ergo gilt (c).

**(2.2) Bemerkung**.[14]

**1°** Jede Zweitursache ist ein *ens per participationem*, d.h. es hat in beschränkter Weise teil am subsistierenden Sein selbst (*esse ipsum subsistens*, d.i. der *actus purus*). Das subsistierende Sein selbst enthält alles endliche Sein in eminenter Weise in sich. Die Partizipation des endlichen Seins am subsistierenden Sein selbst kann jedoch nicht formalursächlich sein, da sonst das subsistierende Sein selbst das Sein der endlichen Dinge wäre. Die Mitteilung des Seins durch das subsistierende Sein muss also *wirkursächlich* sein:

---

veränderlich sind, müssen in diesem Rahmen außer Acht gelassen werden. Für sie gilt aber eine analoge Argumentation.

[12] Die Tätigkeiten der Zweitursachen sind ein Akzidenz. Jedoch kein *accidens proprium* (eine Substanz kann nicht ohne ihre *accidentia propria* existieren, sehr wohl aber ohne ihre jeweilig verschiedenen Tätigkeiten, die ihr ja nicht notwendig zukommen), sondern ein *accidens per accidens*, cf. Feldner (1890), 145f. und 214.

[13] Cf. ScG, II, c. 59: „Id quo aliquid operatur, oportet esse formam eius: nihil enim agit nisi secundum quod est in actu; actu autem non est aliquid nisi per id quod est in forma eius [...]."

[14] Cf. hierzu Gredt (1953), 248.

Die endlichen Seienden bestehen in ihrem limitierten Sein in fortwährender Abhängigkeit vom *esse ipsum subsistens*.[15]

**2°** Die Zweitursachen als Potenzialitäten können nicht *rein aus sich heraus* einen Akt hervorbringen, weil dieser ontologisch reichhaltiger ist als die anfängliche Potenzialität. Aus Weniger wird nicht aus sich heraus Mehr – eine Potenz als Potenz ist nicht suffizient für das „bringing about" der ontologischen Saturiertheit des korrespondierenden Aktes.

**(2.3) Korollar.**

**1°** Erst- und Zweitursache verhalten sich im *concursus* als Totalursachen und nicht als Teilursachen.[16] Beide bringen die Aktsetzung und die Wirkung ganz hervor – dies bedeutet jedoch *keine Überdetermination* des Effekts: Beide bringen ihn unter verschiedener Rücksicht hervor bzw. unter verschiedener Ordnung – die Erstursache unter der unbeschränkten Hinsicht des Seins (*sub illimitato modo entis*), die Zweitursache hingegen unter der Hinsicht des beschränkten Soseins der Tätigkeit (*sub limitato modo actionis*).[17]

**2°** Die thomistische Theorie der Kausalität der Erstursache und der Realisierung ihrer Wirkung in der Zeit schließt die Lehre sogenannter „gemischter Relationen" (*relationes mixtae*) ein, deren erstes Relatum im Modus der Ewigkeit existiert und deren zweites Relatum in der Zeit vorkommt. Darauf soll hier jedoch nicht näher eingegangen werden.[18]

### *1.3 Der concursus praevius (praemotio physica)*

Eine Zweitursache ist, wie wir sahen, nicht essentiell im Akt, und damit bezüglich ihrer Aktsetzungen in einem Status reiner Potenzialität. Sie steht zu ihren Aktsetzungen bzw. Tätigkeiten jedoch in zweifacher Weise in Potenz:[19] (a) Sie ist in Potenz zum *actus secundus formalis*, und (b) sie ist in Potenz zum *actus secundus causalis*. In formaler Hinsicht wird die Zweitursache durch den *concursus simultaneus* zur Aktivität überführt, in kausaler Hinsicht ist jedoch mehr vonnöten:

Das durch den *concursus simultaneus* partizipierte Sein der Tätigkeit kann nicht nur wie ein gewöhnliches Akzidenz rein passiv aufgenommen werden, denn sonst würde es sich nicht um eine Tätigkeit *der Zweitursache* handeln, d.h. dann geht sie nicht *wirkursächlich aus der Zweitursache hervor* bzw. wird nicht *wirkursächlich* von ihr hervorgebracht. Die Tätigkeit muss also vielmehr bereits aktiv mitvollzogen werden, es muss eine Aufnahme im Selbstvollzug stattfinden (es muss also eine Rezeptivität *in* Aktivität bestehen). Dies setzt aber voraus, dass die geschaffene Potenz bereits „vor" (nicht temporal, sondern der Natur nach) der Aufnahme des Seins der Tätigkeit im *concursus simultaneus* bereits „aktiviert" wurde, und damit den *concursus simultaneus* aktiv mitvollziehen kann und damit wirkursächliche Macht über das Entstehen der Tätigkeit besitzt.[20]

Dies führt zur These des *concursus praevius* bzw. der *praemotio physica*:[21]

**(3.1) Satz.** Gott als Erstursache bewegt im *concursus praevius* bzw. in der *praemotio physica* die Zweitursachen intrinsisch zu ihren Tätigkeiten vorher.

**Begründung.**

---

[15] Cf. Gredt (1936), 239.

[16] Cf. Osborne (2006), 626–628.

[17] Cf. Gredt (1953), 248f.

[18] Zur Theorie der gemischten Relationen (mixed relations) und einer „timeless causation" cf. Kretzmann/Stump (1981), 448; Alston (1985), 12f.; Davies (1985), 170; Leftow (1991), 290-312.

[19] Cf. Gredt (1953), 254.

[20] Cf. Gredt (1936), 241. Cf. auch Weissmahr (2005), 154ff.

[21] Cf. zum Folgenden Gredt (1953), 250–255.

(1) In der Begründung von Satz 2.1 wurde bereits klar, dass die Zweitursachen als *potentiae in actu primo* rein passive Potenzen darstellen, die sich damit in einem Status in völliger Inaktivität befinden. [$\forall x(Zx \rightarrow Px)$]

(2) Was jedoch in rein passiver Potenz ist, bedarf, um selbst tätig zu werden, eines Bewegtwerdens von einem Anderen. [$\forall x(Px \rightarrow Bx)$]

(3) Also müssen die Zweitursachen von einem anderen dazu bewegt werden, selbst tätig zu sein. [$\forall x(Zx \rightarrow Bx)$][22]

(4) Was von einem Anderen bewegt wird, wird entweder extrinsisch oder intrinsisch bewegt. [$\forall x(Bx \rightarrow (Ex \vee Ix)$]

(5) Zur Selbsttätigkeit kann aber nicht rein extrinsisch (etwa eine andere Zweitursache) bewegt werden, denn rein extrinsische Ursachen verleihen der Zweitursache nicht die intrinsische Macht, aus sich selbst heraus eine Tätigkeit hervorzubringen,[23] sie wirken nur als externer Anstoß bzw. prägen der Potenz von außen ein Akzidenz ein, das rein passiv aufgenommen wird. [$\neg Ex$]

(6) Eine Verursachung, die der Zweitursache die intrinsische Macht verleiht, aus sich selbst heraus eine Tätigkeit hervorzubringen, muss also der Zweitursache intrinsisch sein. [Ix]

(7) Zudem muss eine Verursachung, die der Zweitursache die intrinsische Macht verleiht, aus sich selbst heraus eine Tätigkeit hervorzubringen, der hervorzubringenden Tätigkeit vorangehen und darf nicht nur konkomitant sein (der rein simultane *concursus* reicht nicht aus).

Die einzige Kausalität, die der Zweitursache nicht schlechthin extrinsisch ist, aber auch nicht schlechthin mit ihr identisch ist, ist die Kausalität der Erstursache. Somit bewegt Gott als Erstursache die Zweitursachen zu ihren Tätigkeiten vorher.[24]

**(3.2) Korollar.**

**1°** Der concursus simultaneus setzt den concursus praevius bzw. die praemotio physica voraus.[25]

**2°** Die Thomisten sprechen bei der *praemotio physica* auch von einer „vorübergehenden Einwirkung" Gottes oder einer von Gott mitgeteilten *entitas vialis* bzw. *instrumentalis*.[26] Gott gibt bei jedem Akt der Zweitursachen zwar eine spezielle Vorherbewegung, aber die Unterscheidung zwischen verschiedenen Vorherbewegungen Gottes als verschiedener göttlicher Tätigkeiten ist nur ein Unterschied *modo nostro intelligendi*, bzw. stellt eine *extrinsische Denomination* dar. Diese extrinsische, aus der Zeit heraus getätigte Prädikation

[22] Cf. Feldner (1890), 151. Cf. S.th. I-IIae, q. 9, a. 1.

[23] Cf. Gredt (1953), 171.

[24] Cf. beispielsweise S.Th. I, q.8, a.1, co.: „[...] Deus [est] in omnibus rebus, et intime"; S.Th. I, q.8, a.1, ad 1: „[...] Deus est supra omnia per excellentiam suae naturae, et tamen est in omnibus rebus [...]"; S.Th. I, q. 105, a.4: „Deus in omnibus intime operatur"; S.Th. I-IIae, q. 109, a.1, co.: „Et ideo quantumcumque natura aliqua corporalis vel spiritualis ponatur perfecta, non potest suum actum procedere nisi moveatur a Deo." Und bezüglich des Willens explizit ScG III, c. 88: „[...] quod motus voluntarius eius [sc. hominis] sit ab aliquo principio extrinseco quod non est causa voluntatis, est impossibile. Voluntatis autem causa nihil aliud esse potest quam Deus"; De Ver, q. 22, a. 8: „operatur intra voluntatem". Cf. ebenso: S.th. I, q. 106, a. 2; q. 111, a. 2; I-IIae, q. 80, a. 1; De Pot, q. 3, a. 7, ad 5 („applicat actioni"!); De Malo, q. 3, a. 3; De Ver, q. 22, a. 9; ScG III, c. 89. "In fact, the will is more plainly in need of divine assistance than is any other power precisely because no other mover can directly act on the will", O'Brien (1981), 671.

[25] Cf. Gredt (1953), 254.

[26] Cf. Gredt (1953), 251.

hat jedoch ein reales Fundament in Gott – auch wenn damit nur auf eine einzige Wirklichkeit in Gott verwiesen wird.

**3°** Die göttliche Kausalität verursacht in der *praemotio physica* nicht den *Akt der Zweitursache*, sonst wären beide in derselben Ordnung. Der *terminus ad quem* der Vorherbewegung ist nicht der *actus* der Zweitursache, sondern die *actuatio potentiae*.[27] „[D]ie motio [divina, R.S.] gibt keine neue Form, verändert die Kraft nicht und ist insofern keine Kraft (virtus quae); sie *aktuiert* aber eine bereits bestehende Kraft (virtus quâ).“[28] Diese *actuatio potentiae* ist also nicht identisch mit dem Akt der Zweitursache selbst. Wir haben es mit einer *Substanzkausalität* von Seiten der Zweitursache zu tun, die von der Erstursache zu dieser Eigenkausalität ermächtigt wird. Die göttliche Vorherbewegung bewirkt also eine *Selbstdeterminierung* und eine *Selbsttranszendenz* der Zweitursachen hin zur Saturiertheit ihrer Akte (cf. 2.2, 2°).

**4°** Die *praemotio physica* ist kein äußerlicher „Fußtritt“ für die Zweitursachen. Wie Leo Elders betont, gilt: Gott ist es, der die „Dinge in ihrer tiefsten Tiefe [...] berührt und aus seiner Kausalität hervorfließen lässt“.[29]

## 2. Prämotion, Freiheit und Prädetermination

### *2.1 Prämotion und Freiheit*

Für den Willen gilt noch viel mehr als für andere Zweitursachen, dass er in passiver Potenz indifferent ist und dass er nicht von einer rein externen Ursache bewegt werden darf, da sonst eine Außendeterminierung vorläge. Doch wie verhalten sich Vorherbewegung und Freiheit des Willens zueinander? Was unter Freiheit zu verstehen ist, sei mit *Peter van Inwagen* definiert:

**(4.1) Definition.** (Freiheit *simpliciter*) Ein Willensakt ist frei, wenn gilt: „We are sometimes in the following position with respect to a contemplated future act: we simultaneously have both the following abilities: the ability to perform that act and the ability to refrain from performing that act.“[30]

**(4.2) Korollar** (aus 3.1 und 3.2)**.**

**1°** Auch hier gilt im Besonderen, was in 3.2, 3° im Allgemeinen dargelegt wurde: Der *terminus* der Vorherbewegung Gottes ist nicht der *actus* des Willens, sondern die *actuatio potentiae* (bei Thomas wird diese *actuatio potentiae* auch „intentio“ genannt, cf. De Pot, q. 3, a. 7)[31]. Diese *actuatio potentiae* ist also nicht identisch mit dem Wollen (*velle*) des Willens selbst.[32] Der freie Wille exerziert eine **Akteurskausalität**. Es „geht nach Thomas der Wille selbst ‚in Bewegung über‘, d.h. setzt aus eigener Kraft den *Willensakt*, aber die Kraft und Macht dazu gibt ihm Gott, durch die motio in der natürlichen Ordnung, durch das auxilium gratiae in der übernatürlichen Ordnung [...] (I-II q. 109 a. 2 ad 1).“[33]

**2°** Gott verletzt den freien Willen nicht: Ein *actus violentus* gegen den Willen würde nur dann vorliegen, wenn „das bewegende Subjekt den Akt, zu dem es bewegt wird, nicht selbst

---

[27] Cf. Schultes (1925), 288.

[28] Schultes (1925), 288.

[29] Elders (1987), 314.

[30] Van Inwagen (2008), 337.

[31] Cf. Schultes (1925), 287.

[32] Schultes (1925), 473, Fn. 1.

[33] Schultes (1925), 473f. Cf. O’Brien (1981), 671: „The premotion does not anticipate the will’s act; it makes possible the act’s exercise. Nor does it deprive the will of its own causality; rather, in bringing about the transition to act, it makes this causality effective.“

setzt“[34]. Der freie Wille kann als freier von Gott zu nichts anderem bewegt werden als *zu seinem eigenen Akt*[35]: „Wenn nämlich Gott den Willen nötigte, so würde das besagen, daß der Wille nicht selbst wollte (non esset cum actu voluntatis), ja daß der Wille nicht bewegt würde [...].“ „Ein motus violentus, ergänzt Thomas noch, wäre dann gegeben, wenn der motus ‚esset contrarius motui voluntatis‘, d.h. wenn das, was Gott im Willen bewirkt, der Willensbewegung entgegengesetzt wäre. Das sei aber im gegebenen Falle unmöglich: ‚quia sic idem vellet et non vellet‘, d.h. der Wille würde zugleich wollen und nicht wollen.“[36]

**(4.3) Explikationen und Bemerkungen.**

**1°** Der Wille ist als aktives Prinzip nocheinmal unterteilt in ein passives und ein aktives Moment:[37] Er ist zunächst **passive Potenz** (*potentia in actu primo*[38] bzw. *potentia non operativa* bzw. *pincipium patiendi ab alio*[39]) im Zustand der Untätigkeit. Der Wille als passive Potenz ist **passiv indifferent**, d.h. in völliger Indifferenz gegenüber Tätigkeit und Untätigkeit und gegenüber jedwedem möglichen Objekt des Wollens. Er ist **aktive Potenz** (*actus secundus* bzw. *potentia operativa*[40]), wenn er einen Akt exerziert. Jedoch auch die *aktive Potenz* ist nicht reiner Akt, sondern noch einmal in passiver Potenz zu ihrer Wirkung, die ihr gegenüber Akt ist – denn die Tätigkeit des Willens ist real verschieden von seiner aktiven Potenz.[41] Der Wille als aktive Potenz ist **aktiv indifferent**, d.h. er bleibt auch im aktuellen Vollzug Herr seiner Tätigkeit und kann wieder von ihr ablassen.[42] Diese Indifferenz steht in Gegensatz zur absoluten Notwendigkeit.[43] Die aktive Indifferenz gehört wesentlich zur Willensfreiheit,[44] denn wo eine bloße *Möglichkeit* zu einem Akt besteht, kann noch keine Herrschaft über den Akt vorliegen.[45] Die aktive Indifferenz besagt, dass ein Wesen die Macht hat, einen Akt zu exerzieren oder ihn zu unterlassen. Es ist die aktive Potenz, durch welche

---

[34] Schultes (1924), 188.

[35] Cf. S.Th. I-IIae, q. 9, a. 3.

[36] Schultes (1924), 192 und 186, unter Berufung v.a. auf S.th. I-IIae, q. 9, a. 4, ad 2.

[37] Cf. S.th. I-IIae, q. 10, a. 1, ad 2; De Potentia, q. 1, a. 1. Cf. Feldner (1890), 63f. Cf. Super Sent., 1, dist. 42, q. 1, a. 1, ad 1: „potentia primo imposita est ad significandum principium actionis; sed secundo translatum est ad hoc ut illud etiam quod recipit actionem agentis, potentiam habere dicatur; et haec est potentia passiva; ut sicut potentiae activae respondet operatio vel actio, in qua completur potentia activa; ita etiam illud quod respondet potentiae passivae, quasi perfectio et complementum, actus dicatur. Et propter hoc omnis forma actus dicitur, etiam ipsae formae separatae; et illud quod est principium perfectionis totius, quod est Deus, vocatur actus primus et purus, cui maxime illa potentia convenit.“

[38] Cf. Feldner (1890), 82.

[39] Cf. Gredt (1953), 38f.

[40] Cf. Gredt (1953), ebd.

[41] Cf. Feldner (1890), 76.

[42] Feldner (1890), 66. Gott muss besitzt bezüglich der Geschöpfe eine aktive Indifferenz: Er kann die Geschöpfe so und so erschaffen oder auch nicht erschaffen. Es muss aber jedwede passive Indifferenz von ihm ausgeschlossen werden, da er nicht vervollkommnet werden kann (jeder Akt ist die Vollkommenheit seiner Potenz) und da seine Tätigkeit kein Akzidenz ist; dies ist nur da möglich, wo Wesen und Existenz verschieden sind, cf. Feldner (1890), 67.

[43] Notwendigkeit bedeutet in diesem Kontext, dass der Wille nur *ein* eindeutig bestimmtes Objekt anstreben kann und dieses auch anstreben *muss*. In Gott gibt es eine Notwendigkeit des Wollens bezüglich seiner selbst, die jedoch keinen äußeren Zwang bedeutet: Als *actus purus* muss er seine Verstandes- und Willenstätigkeit immer ausüben. Er will notwendig sein eigenes Wesen, das die Vollkommenheit aller Vollkommenheiten und das absolute Gute ist; liebt und will er noch etwas anderes außer seiner selbst, dann um seiner selbst willen: Jeder Wille hat ein doppeltes Objekt: Sein *Hauptobjekt* und sein *sekundäres Objekt*. Gott will die Geschöpfe nur als sekundäres Objekt um des Hauptobjekts seines Willens wegen, das er selbst ist (er will die Geschöpfe aufgrund ihrer Ähnlichkeit zu seinem Wesen). Hinsichtlich des sekundären Objekts seines Willens hat Gott volle Wahlfreiheit, denn Gott ist allseitig vollkommen selbst dann, wenn gar keine Kreatur existiert. Cf. Feldner (1890), 69 – 71; cf. S.th. I, q. 19, a. 3; De Ver, 1. 23, a. 4.

[44] Cf. Feldner (1890), 69.

[45] Cf. Feldner (1890), 82, unter Berufung v.a. auf Super Sent., 2, dist. 24, q.1, a.1, ad 2.

ein Akt hervorgebracht wird, die passive Potenz hingegen nimmt den Akt in sich auf. Sie kann also eben gerade nicht tun, was sie will. Die Vollkommenheit, Macht zu haben über den eigenen Akt und damit frei zu sein, kommt folglich der aktiven Potenz zu.[46] In der Exerzierung des Aktes muss die Potenz, den Akt auch wieder zu unterlassen, bestehen bleiben. Denn andernfalls wäre der Wille im Akt gleich dem *actus purus*, dem gar keine Potenzialität mehr beigemischt ist.[47]

**2°** Die *praemotio physica* verletzt also den Thomisten zufolge die Freiheit nicht, sie ist vielmehr sogar notwendig, um die Freiheit formal und als solche zu wahren.[48] Ein Individuum ist frei zu nennen, wenn es der *dominus suorum actuum*, der Herr seiner Akte ist.[49] Zu diesem wird es durch die *praemotio physica*:[50] „Die Potenz *in actu* ist frei, indem sie ihre Thätigkeit in der Weise vollzieht, dass sie die Möglichkeit zum Gegenteil beibehält."[51] Die Freiheit liegt also nicht in der Inaktivität und in der passiven Indifferenz.[52] Gott selbst besitzt keinerlei Inaktivität und passive Indifferenz, er ist reiner Akt und besitzt nur *aktive Indifferenz* und besitzt dennoch Freiheit, sogar in ihrem Höchstmaß. Auch beim Geschöpf ist die Freiheit formal und eigentlich dann vorhanden, wenn es *in actu* ist.[53] Würde die Freiheit formal in der passiven Indifferenz liegen, dann würde der Wille beim Übergang zum Akt seine Freiheit selbst zerstören (das Problem besteht also auch ganz unabhängig von der *praemotio physica*).[54]

**3°** Wenn der Wille *in actu* ist, den Akt exerziert, dann kann er, *solange* er den Akt exerziert, diesen nicht *nicht* exerzieren. Es ist notwendig *ex suppositione*, dass er diesen exerziert (auch: ***necessitas suppositionis*** bzw. *compositionis* bzw. s*ecundum quid*,[55] bzw. *necessitas in sensu composito* = den faktischen Vollzug des Aktes hinzugenommen). Aber das bedeutet *keine* absolute Notwendigkeit, denn die Potenz, den Akt wieder unterlassen zu

---

[46] Cf. Feldner (1890), 74, 75 und 82f.

[47] Cf. Feldner (1890), 76 und 85. Cf. ebenso ScG, I, c. 85. Gott besitzt Freiheit im höchstmöglichen Grad. Er besitzt, wie oben angeführt, keine passive Indifferenz. Er ist nicht „manchmal *agens in potentia*, manchmal *agens in actu*; noch auch verhält sich seine active Potenz *passiv* mit Bezug auf seine Thätigkeit. Denn obgleich sein Willensact immanent ist, so bildet er doch nicht ein Accidens, welche der Potenz inhäriert, sondern ist in der Wirklichkeit, real die Potenz selber", Feldner (1890), 74. Auch in Gott wird von Thomas eine Aktivpotenz angenommen (*potentia operativa increata*), die jedoch nicht real von ihrem Akt, dem *actus purus*, verschieden ist – im Gegensatz zur Aktivpotenz der Geschöpfe, cf. Gredt (1953), 163-167. Durch das sekundäre Objekt seines Willens (die Geschöpfe) gewinnt Gott nicht an Vollkommenheit, und dieses Objekt wirkt auch nicht bewegend auf ihn, er ist also nicht passiv indifferent bezüglich des Objekts seines Willens. Die passive Indifferenz kann also nicht *an sich* zur Freiheit gehören (cf. Feldner (1890), 71. Cf. ScG, IV, c. 19; S.th. I-IIae, q. 10, a. 1, ad 2). Ansonsten wäre *jedes* geschaffene Seiende, auch ein Stein, frei, da alle passive Potenz in sich besitzen, cf. Feldner (1890), 201.

[48] Cf. Feldner (1890), 199f.

[49] Cf. Feldner (1890), 201: „Solange ein Geschöpf *unthätig*, in der Möglichkeit oder Potenz zur Thätigkeit, ein Agens *in potentia* ist, kann man offenbar nicht behaupten, es besitze formell die Herrschaft über seine Thätigkeit. Über das, was Jemand *nicht* besitzt, kann er unmöglich Herr sein, er wäre König ohne Land und Unterthanen", unter Berufung auf Super Sent, 2, dist. 24, q. 1, a. 1, ad 2; 4, dist. 24, q. 1, a. 1; q. 2, qu. 2, ad 3.

[50] Cf. Feldner (1890), 200.

[51] Feldner (1890), 202. Cf. Joannes a S. Thoma, *Philosophia naturalis*, IV, q. 12, a. 3: "Sufficit dicere, quod positis omnibus requisitis, etiam motione Dei, prout praebet determinationem ad actum tollendo quidem indifferentiam potentialem et suspensivam, praebendo autem et conservando modum indifferentiae actualis et potestatis dominativae, potest oppositum facere voluntas", cf. Osborne (2006), 629, fn. 64.

[52] Cf. Feldner (1890), 201.

[53] Cf. Feldner (1890), 203.

[54] Cf. Feldner (1890), 202, 220f. und 203; Cf. Joannes a S. Thoma, *Philosophia naturalis*, IV, q. 12, a. 3; *Cursus theologicus*, d. 25, a. 4, n. 1; d. 25, a. 4, n. 9 – 11. Cf. Osborne (2006), 629, fn. 64.

[55] Cf. Bannez, *Scholastica commentaria in primam partem angelici doctoris D. Thomae [In ST I]*, q. 19, a. 3 und a. 8. Cf. Osborne (2006), 613 und 616–624.

können, bleibt – wie oben dargelegt – ungeschmälert bestehen (*in sensu diviso*). Es ist nur die *passive* Indifferenz aufgehoben, nicht aber die *aktive* Indifferenz. Und da letztere wesentlich zur Freiheit gehört, ist die Freiheit nicht verletzt.[56] An die *praemotio* ist also nur eine *necessitas suppositionis* geknüpft.[57] Die *praemotio* bewirkt die Selbstdeterminierung des Individuums (cf. 3.2, 3°), d.h. „[d]eterminieren sich die Creaturen *ebenfalls*, so sind deren Thätigkeiten *freie*. Werden sie *nur* von Gott determiniert, so nennen wir ihre Thätigkeiten *unfreie* und nothwendige."[58] „Wird nun angenommen, die *praemotio physica* determiniere den Willen der freien Geschöpfe auf solche Weise, dass die Thätigkeit *mit absoluter Nothwendigkeit* folgt, so behauptet man, die *praemotio* wirke *gegen* die Neigung, *gegen die Natur und das Wesen* der Freiheit. Alles aber, was gegen die Natur und Neigung eines Dinges geht, das ist *Zwang*, *Gewalt*. Den Willen vermag in Bezug auf seine Thätigkeit, hinsichtlich des *actus elicitus* keine Macht zu zwingen. Dies vermag selbst Gott nicht."[59] Gott ändert nicht den *modus* der Natur, die er zu ihrem Wirken appliziert[60].

### *2.2 Prädetermination und Sünde*

**1° Prädetermination:** Da Gott den Thomisten zufolge zu *einem* spezifischen Ziel bestimmt (die Potenz *unfehlbar* zu *einem* bestimmten Akt vorherbewegt),[61] sprechen die Thomisten auch von einer „*Praedeterminatio physica*". Praemotio und Praedeterminatio sind real nicht verschieden, aber während die praemotio sich begrifflich eher auf die Wirkursächlichkeit bezieht,[62] bezieht sich die *praedeterminatio* auf die Hinordnung zum Ziel. Beides, Wirkursache und Ziel sind jedoch dasselbe: Gott.[63] Gott intendiert sich in der *praemotio physica* selbst als Endziel *in ordine morali*.[64] Die Freiheit des Menschen ist für die Thomisten insbesondere die Indifferenz gegenüber endlichen Bestimmungen: Kein endliches Gut kann den Menschen befriedigen.[65] Die *praemotio physica* ist "gerade die Befreiung aus [den endlichen] Fesseln auf ein unendliches Gut hin, an dem der Mensch seine Erfüllung findet."[66]

**2° Prämotion und Sünde:** Bezüglich des Verhältnisses von Vorherbewegung und sündhaften Akten können in diesem Rahmen nur einige grundsätzliche Bemerkungen gemacht werden: Gemäß dem thomistischen Grundaxiom „ens et bonum convertuntur" wird das Böse als Privation *in ordine morali* aufgefasst. Jede Privation aber hat ein Seiendes als zugrundeliegendes Subjekt – und jedes Seiende ist ein „bonum per participationem". Durch die *praemotio* ist Gott die Erstursache des positiven Seins der sündhaften Akte (des zugrundeliegenden Subjekts), aber nicht die Ursache des Mangels an Gutheit (der Privation). Die Privation als solche ist nicht positiv verursacht, sie hat keine *causa efficiens*, sondern nur eine *causa deficiens* – welche in der Zweitursache liegt, und nicht in der Erstursache. Keine moralisch gute Handlung kann ohne göttliche Vorherbewegung (d.h. hier ohne *gratia ex se*

---

[56] Cf. Feldner (1890), 84. Cf. S.th. I, q. 14, a. 3, ad 3; q. 19, a. 8, ad 1; S.th I-IIae, q. 10, a. 4, ad 3.

[57] Cf. Feldner (1890), 214.

[58] Feldner (1890), 215f.

[59] Feldner (1890), 216. Cf. S.th. I-IIae, q. 10, a. 4.

[60] Feldner (1890), 222. Cf. S.th. I, q. 19, a. 8; I-IIae, q. 10, a. 4. Cf. auch S.Th. I, q. 23, a. 5, co: „Non est autem distinctum quod est ex libero arbitrio, et ex praedestinatione; sicut nec est distinctum quod est ex causa secunda, et causa prima, divina enim providentia producit effectus per operationes causarum secundarum, ut supra dictum est. Unde et id quod est per liberum arbitrium, est ex praedestinatione."

[61] Cf. ScG, III, c. 89 und c. 149; De pot, q. 3, a. 7, ad 5.

[62] Jede rein moralische oder finale Verursachung wäre für Thomas eine unzureichende Erklärung, cf. S.th. I, q. 82, a. 2, 4; I-IIae, q. 9, a. 1; q. 10, a. 2; cf. O'Brien (1981), 671. "Because the will is in potency to actual operation, efficient (physical) causality must be exercised on the faculty itself", ibid, 671.

[63] Cf. Feldner (1890), 212.

[64] Feldner (1890), 235.

[65] Cf. Ramelow (1997), 49.

[66] Ramelow (1997), 49.

*efficax*) vollzogen werden – die sündhaften Akte aber folgen aus der moralischen Defizienz der Zweitursachen (cf. 1.6, 6°).[67]

### *2.3 Prädetermination, Skotismus und „counterfactual power"*

**1° Parallelen zum Skotismus:** Dass der Wille im Akt *in sensu diviso* die Potenz für das Gegenteil behält, ist eine Lehre, die wohl zum ersten mal von Duns Scotus in seinem Opus Oxoniense, I. Sent, dist. 39, explizit ausformuliert wurde: Der Satz „Voluntas volens A potest non velle A" sei *in sensu composito* falsch, *in sensu diviso* aber wahr.[68] Die „potentia ad oppositum" geht dabei nach Scotus dem Willensakt nicht zeitlich, aber *in ordine naturae* voraus. Es besteht nicht die Möglichkeit, im gleichen Zeitpunkt dasselbe zu wollen und nicht zu wollen, aber die *Fähigkeit*, zugleich auch nicht zu wollen, kann zugleich mit dem Akt bestehen: „Simul habeo potentiam ad opposita, sed non ad opposita simul."[69]

**2° „Counterfactual power":** Die Potenz zur Unterlassung des Aktes bzw. zu einem anderen Akt bleibt, wie wir gesehen haben, auch für die Thomisten unter der Prämotion und Prädetermination Gottes erhalten. Aber ist dies eine tatsächliche Macht, unter den gleichen, die Prämotion einschließenden Umständen anders handeln zu können? Oder handelt es sich nur um eine „counterfactual power", wie die Ausführungen des bannezianischen Thomisten Joseph Gredt nahelegen:

> Indem die praemotio physica den Willen „zu einer bestimmten Willensentscheidung führt, bringt sie alle im Willen sich befindenden Möglichkeiten, auch die Möglichkeit des Nichtwollens, des Widerstrebens und des Anderswollens [...] zum Wirken. Sie bringt diese Möglichkeiten zum Wirken, *nicht* in dem Sinne, als wenn der Wille *tatsächlich widerstände* und anders wollte, sondern in dem Sinne, dass der Wille bei Setzung der Willensentscheidung, zu der er von Gott bewegt wird, das Widerstehenkönnen und Anderswollenkönnen, das er vor der Vorherbewegung nur [...] als leidentliche Möglichkeit hatte, jetzt tatsächlich ausübt: Er hat die tatsächliche Kraft, zu widerstehen und anders zu wollen. Nicht als wenn er diese Tätigkeit des Widerstehens ausübte, aber in der Tätigkeit, die er unter der Vorherbewegung setzt, übt er das Widerstehenkönnen tatsächlich aus, weil er diese Tätigkeit frei setzt."[70]

Der Wille übt seine Macht, anders handeln zu können, also Gredt zufolge nicht *tatsächlich* aus (d.h. er übt sie nur kontrafaktisch aus: wenn er sie faktisch ausüben würde, wäre nicht die entsprechende Prämotion vorhergegangen). Hier zeigt sich wieder eine Parallele zum Skotismus: Die Scotus-Schüler *Francis of Marchia* (ca. 1290–1344) und *William of Rubione* (*1290) sprechen bezüglich der im Willensakt verbleibenden Potenz zum Gegenteil von einer „indeterminatio de possibili", durch die in anderer Terminologie das ausgedrückt wird, was die Thomisten die „aktive Indifferenz" nennen.[71] *William of Rubione* sagt hierbei ausdrücklich, dass dies eine bloße Möglichkeit bleibe, die niemals Wirklichkeit wird: „potest, dico, de possibili, non tamen illud facit de facto."[72] Die Freiheit des geschaffenen Willens besteht also letztlich in der *reinen Möglichkeit*, anders zu handeln, die trotz der unfehlbaren Determinierung durch Gott nicht aufgehoben ist.[73] Wir hätten es demnach letztlich mit einer reinen *Vollzugsfreiheit*, nicht aber mit *Wahlfreiheit* zu tun. Reicht dies aber für Freiheit wirklich aus?

---

[67] Cf. Feldner (1890), 223–240. Cf. auch Grant (2009).

[68] Cf. Duns Scotus, Opus Oxoniense, I. Sent., dist. 39, 134–135. Cf. Schwamm (1934), 21.

[69] Cf. Scotus, Op. Ox., ibid., 135. Cf. Schwamm (1934), 22.

[70] Gredt (1935), 235f. Herv. R.S.

[71] Im Gegensatz zur „determinatio de inesse" durch Gott. Cf. Schwamm (1934), 248f.

[72] Wilhelm de Rubione, Sentenzenkommentar II, dist. 38 q. 1. Cf. Schwamm (1934), 266f. and 333.

[73] Cf. Schwamm (1934), 267.

**3° Reine Unterlassungen:** Eine weitere kritische Anfrage an die Theorie der Vorherbewegung wäre, dass die Aktivierung der Potenz zu einem gegenteiligen Akt den thomistischen Prinzipien zufolge einer weiteren Vorherbewegung durch Gott bedürfte. Doch dann würde sich die gesamte Problematik schlichtweg wiederholen. Ein möglicher Ausweg wäre jedoch die Annahme „reiner Unterlassungen", die nicht eine Aktualisierung einer Potenz zu einem spezifischen positiven Akt darstellen, sondern den Willen einfach „offline" gehen lassen. Solche reine Unterlassungen würden eben keiner Vorherbewegung bedürfen. Dann aber – so ein schwerwiegender Einwand von *Sebastian Izquierdo* – entziehen sie sich auch jedweden prädeterminierenden Dekreten Gottes und unterminieren die thomistische Theorie von den göttlichen Dekreten als *medium quo* des göttlichen Wissens um zukünftig kontingente Handlungen.[74]

## 3. Molinistischer *concursus simultaneus* und die „Essence-solution"

Die Molinisten lehnen die Lehre vom *concursus praevius* bzw. von der *praemotio physica* ab, da eine Vorherbewegung den freien Willen prädeterminiert und damit schlichtweg die Freiheit zunichte mache. Mit Molina nehmen die meisten Molinisten einen rein allgemeinen, simultanen *concursus* an:

> Concursus Dei generalis non est influxus Dei in causam secundam, quasi illa prius eo mota agat, et producat suum effectum, sed est influxus immediate cum causa in illius actionem et effectum. (Concordia, q. XIV, a.13, dist. 26.)

Erstursache und Zweitursache wirken daher für Molina auch nicht als Totalursachen, sondern als Partialursachen:

> Cum dicimus neque Deum per concursum universalem, neque causas secundas esse integras, sed partiales causas effectuum, intelligendum id est de partialitate causae, ut vocant, non vero de partialitate effectus: totus quippe effectus et a Deo est, et a causis secundis; sed neque a Deo, neque a causis secundis, ut a tota causa, sed ut a parte causae, quae simul exigit concursum et influxum alterius: non secus ac cum duo trahunt navim, totus motus proficiscitur ab unoquoque trahentium, sed non tanquam a tota causa motus, siquidem quivis eorum simul efficit cum altero omnes ac singulas partes ejusdem motus. (Concordia, q. XIV, a.13., dist, 26.)[75]

Der Molinismus lehrt bezüglich des freien Willens überdies, dass Gott dem Willen zunächst seine entitative Mitwirkung „anbietet" (*concursus oblatus*), welche der freie Wille durch seine eigenständige Entscheidung dann zur *tatsächlich geleisteten* Mitwirkung (*concursus collatus*) werden lässt[76] (auch der Molinismus hat damit eine Prämotionslehre: Nur geht hier die Prämotion vom Geschöpf aus und wirkt umgekehrt auf Gott![77]). Die Thomisten lehnen diese Lehre wiederum als in sich widersprüchlich ab: Der *concursus collatus* **gibt** der Willensentscheidung die **Existenz**, aber um den *concursus collatus* zu „aktivieren", muss die Willensentscheidung **bereits existieren**. Die Willensentscheidung kann aber nicht zugleich existieren und nicht existieren.[78]

Der Molinist muss hierauf mit einer Unterscheidung antworten: Gott gewährt seine Mitwirkung nur dort, wo er weiß, dass die Kreatur auch „mitspielt". Dies weiß er kraft der

---

[74] Cf. Ramelow (1997), 51, Fn. 108, 174f, 175 Fn. 284.
[75] Cf. Dummermuth (1886), 17f.
[76] Cf. Gredt (1953), 260.
[77] Cf. Feldner (1890), 191.
[78] Cf. Gredt (1953), 262.

*scientia media futuribilium*, welche Gott auch die nötige Souveränität über die Geschöpfe verleiht, um seine Providenz exerzieren zu können.[79] In der *scientia media* hat die Entscheidung noch kein „physisches Sein" (*esse physicum*), welches ihr nur durch den *concursus collatus* vermittelt wird, sondern sie besitzt lediglich ein „intentionales Sein" (*esse intentionale*).[80] Tatsächlich wurde von Molinisten wie *Antonio Perez* (1599–1716) und *Martin de Esparza Artieda* (*1621) die These einer *intentionalen Präexistenz* möglicher freier Entscheidungen im göttlichen Verstand (*electio in mente divina*) vertreten. Die menschliche Wahl hat demzufolge eine zweifache Existenz: Einmal die tatsächliche Existenz in einer aktualen Schöpfung, und einmal als sog. von Ewigkeit her existierende „*futuritio formalis*" eine *intentionale Realität* logisch vorgängig zum Schöpfungsakt im Geiste Gottes. Das aktual existierende Ereignis ist hierbei nicht vorausgesetzt, sondern dieses wird durch die *futuritio formalis* erst konstituiert. Diese *futuritio formalis* ist für den göttlichen Verstand nur *intentional*, nicht aber wirkursächlich bestimmend: Nicht eine außergöttliche Realität wirkt auf den göttlichen Verstand in der *scientia media*, sondern der in Gottes Wesen enthaltene[81] rein formale konditionale Zusammenhang zwischen möglichen Umständen und geschöpflicher Wahl. Er besitzt gegenüber Gottes Wissen keine *prioritas ut quod* (= *prioritas existentiae*), sondern nur eine *prioritas ut quo* (= *prioritas eligentis*). Das göttliche Wissen ist damit einerseits nicht die Ursache der Willensentscheidungen, und andererseits ist dieses Wissen nicht von außen in Gott verursacht.[82]

Im Zuge dieser molinistischen Theoriebildung kam konsequenterweise die Idee auf, dass das Mittlere Wissen auf eine absolut vollständige Repräsentation der möglichen Individuen im göttlichen Geist gegründet werden müsse. Einschlägig hierfür ist vorallem der Molinist *Hieronymus Fasolus* (1568–1639), welcher Molinas Gedanken der göttlichen „*supercomprehensio*", d.h. der unendlichen Repräsentationskraft des göttlichen Wesens in Richtung einer vollständigen Kenntnis von *individuellen Essenzen* weiterdenkt.[83] Nach *Suárez* ist eine solche Repräsentation nicht nötig, da es ausreiche, dass kontrafaktische Konditionale entweder wahr oder falsch seien (d.h. dass das Gesetz des Kontrafaktisch ausgeschlossenen Dritten gelte) – ein allwissendes Wesen weiß dann per definitionem, welches der Disjunkte wahr ist. *Leibniz*' Kritik an dieser sog. „Veritas-determinata"-Lösung bestand nun in der Rückfrage, *woher* die kontrafaktischen Konditionale denn ihre Wahrheitswerte erhielten: Sie können nicht extensional verifiziert werden, da sie vorgängig zur Schöpfung noch keine Extension besitzen. Sie können allenfalls intensional verifiziert werden: Durch Einsicht in die individuellen Essenzen der möglichen Akteure. Auch wenn *Leibniz* selbst den Begriff des Mittleren Wissen ablehnt, lässt sich dennoch zeigen, dass seine Idee einer „*notio completa*" eines Individuums im Kontext der molinistisch-thomistischen Debatte entstanden ist und eine starke Affinität zur molinistischen Position besitzt.[84] Seine Intuition bezüglich des göttlichen Wissens um zukünftig kontingente Handlungen steht in einer Linie mit den Entwicklungen in der molinistischen Debatte:[85]

---

[79] The Molinist theory of "[...] of a motion that is merely moral and simultaneous concurrence are what make the *scientia media* indispensable to Molinism", O'Brien (1981), 669.

[80] Cf. Ramelow (1997), 228.

[81] Cf. Ramelow (1997), 226.

[82] Cf. Ramelow (1997), 219. Cf. zu dieser gesamten Theorie der *futuritio formalis* ebd., 222–230.

[83] „Mente Molinae [...] causa libera [...] non potest perfectissimo modo obiective cognosci, nisi simul cognoscantur et omnia, quae sunt in causa, et praeterea omnia quae ex causa vel esse possunt, vel erunt, vel sunt, vel fuerunt, *vel essent*; nam effectus etiam, atque adeo omnes isti effectus, sunt aliquid causae; ergo qui cognoscit perfectissimo modo causam, eius etiam effectus, quavis ratione ab ea pendentes, cognoscat necesse est [...]. Quod autem haec perfectissima cognitio respectu effectuum futurorum esse debeat infinita, patet." (In primum partem Summae D. Thomae Commentariorum, T.2, Lyon 1629, 269a. Cf. Knebel (1991), 3; cf. Brüntrup/Schneider (2011); cf. Brüntrup/Schneider (2013).)

[84] Cf. Hübener (1988), 114; Ramelow (1997), 401–419. Cf. Brüntrup/Schneider (2013).

[85] Cf. Brüntrup/Schneider (2013).

> [Gott] bewahrt unser Sein und bringt es kontinuierlich hervor, und zwar so, dass uns die Gedanken spontan oder frei in derjenigen Ordnung begegnen, die der *Begriff unserer individuellen Substanz* trägt, in welchem man sie von Ewigkeit her voraussehen konnte (Discours de métaphysique, § 30).

## 4. Ausblick: Molinistische vollständige Begriffe

In seinen Briefen an *de Volder* führt Leibniz weiter aus, dass die vollständigen Begriffe eines Individuums als Funktionen aufgefasst werden müssen.[86] Ein entsprechendes mathematisches Modell für eine analytische Rekonstruktion von „molinistischen vollständigen Begriffen" wurde in Brüntrup/Schneider (2011) geliefert. Einige Angelpunkte einer solchen Rekonstruktion seien abschließend angeführt:[87]

(a) Die vollständigen Begriffe sind keine volldeterminierten, rein aktualistischen Essenzen (Superessentialismus), sondern reine *Begriffe* bzw. Ideen im Geiste Gottes, die zudem das gesamte Möglichkeitsspektrum eines Individuums enthalten. Sie sind *keine* Eigenschaftsbündel, sondern haben eine „*transworld-identity*". (b) Diese Begriffe sind zudem die von der „Grounding objection" geforderten metaphysischen „*truth-maker*" der kontrafaktischen Konditionale, aber sie existieren als Begriffe nicht unabhängig vom Geist Gottes. Und (c) sie sind nicht algorithmisch bzw. gesetzesartig. Sie enthalten eine eindeutige „thin red line" des Individuums für jede mögliche Welt, die jedoch nicht deterministisch ist: Für eine Determiniertheit ist mehr erforderlich als bloße Eindeutigkeit eines Weltverlaufs – dieser Weltverlauf muss zudem noch einer gesetzesartigen Fortschreibung gehorchen. Gesetzesartigkeit impliziert Eindeutigkeit, jedoch Eindeutigkeit impliziert vice versa nicht Gesetzesartigkeit.[88]

**Ruben Schneider**

Hochschule für Philosophie, Philosophische Fakultät SJ, München
ruben.schneider@hfph.de

## Literatur

Adams, R. 1977: „Middle Knowledge and the Problem of Evil", *American Philosophical Quaterly* 14, 109-117.

Alston, W. P. 1985: „Divine-Human Dialogue and The Nature of God, *Faith and Philosophy* 2, 5-20.

Benz, W. 1936: „Das göttliche Vorherwissen der freien Willensakte der Geschöpfe bei Thomas von Aquin", *Divus Thomas* 14, 255-273.

Brüntrup, G. und Schneider, R. 2011: „How Molinists Can Have Their Cake and Eat It Too", in Ch. Kanzian, W. Löffler, und J. Quitterer (Hrsg.): *The Ways Things Are*, Frankfurt: ontos, 221-241.

- 2013, "Complete Concept Molinism", in: European Journal for Philosophy of Religion, forthcoming.

Davies, B. 1985: *Thinking About God*. Introducing Catholic Theology, Vol. 5. London.

[86] Cf. den Brief an *de Volder* vom 24. März/3. April 1699, Hauptschriften, 1996, Bd. II, 475, und an *de Volder* vom 21. Januar 1704, Hauptschriften, 1996, Bd. II, 513–518.

[87] Cf. Brüntrup/Schneider (2013).

[88] Cf. Schneider (2009), 130-134.

Dummermuth, A. M. 1886: *S. Thomas et Doctrina Praemotionis Physicae*. Paris.

Elders, L. 1987: *Die Metaphysik des Thomas von Aquin in historischer Perspektive*. 2 Bde. Salzburg/München.

Feldner, G. 1890: *Die Lehre des heil. Thomas von Aquin über die Willensfreiheit der vernünftigen Wesen*. Graz.

Flint, Th. P. 1998: *Divine Providence. The Molinist Account*. Ithaca: Cornell University Press.

Grant, M. W. 2009: „Aquinas on How God Causes The Act of Sin Without Causing Sin Itself", in *The Thomist* 73, 455–496.

Gredt, J. 1935: *Die aristotelisch-thomistische Philosophie*, Bd. 2, Freiburg i. Br.

- 1936: „Die göttliche Mitwirkung im Lichte der thomistischen Lehre von Wirklichkeit und Möglichkeit", *Divus Thomas* 14, 237-242.

- 1953: *Elementa Philosophiae Aristotelico-Thomisticae*, Vol. II, Editio Decima Recognita. Freiburg i.Br.

Hasker, W. 1994: "A Philosophical Perspective", in Ch. Pinnock, R. Rice, J. Sanders, W. Hasker, und D. Basinger (Hrsg.): *The Openness of God: A Biblical Challenge to the Traditional Understanding of God*, Downer's Grove IL: InterVarsity Press, 126-154.

Hübener, W. 1988: „‚Notio completa.' Die theologischen Voraussetzungen von Leibniz' Postulat der Unbeweisbarkeit der Existentialsätze und die Idee des logischen Formalismus", in *Studia Leibnitiana*, Sonderheft 15, 107-116.

Knebel, S. 1991: „Necessitas moralis ad optimum. Zum historischen Hintergrund der Wahl der besten aller möglichen Welten", *Studia Leibnitiana* 23, 3-24.

Kretzmann, N. und Stump, E. 1981: „Eternity", *Journal of Philosophy* 78, 429-458.

Leftow, B. 1991: *Time and Eternity*. Ithaca NY/London: Cornell Studies in the Philosophy of Religion.

O'Brien, C. 1981: „Premotion, Physical", *New Catholic Encyclopedia* XI. Washington, 741-743.

Osborne, Th. M. Jr. 2006: „Thomist Premotion and Contemporary Philosophy of Religion", *Nova et Vetera*, English Edition, Vol. 4, No. 3, 607-632.

Perszyk, K. (Hrsg.) 2011: *Molinism. The Contemprary Debate*. Oxford: Oxford University Press, 96-117.

Ramelow, T. 1997: *Gott, Freiheit, Weltenwahl. Die Metaphysik der Willensfreiheit zwischen A. Perez S.J. (1599 – 1649) und G.W. Leibniz (1646 – 1716)*. Leiden: Brill.

Schneider, Chr. 2007: *Spontaneität und Freiheit. Ein ontologischer Theorieansatz*. Habilitationsschrift eingereicht bei der Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, Ludwig-Maximilians-Universität München.

- 2009: *Metaphysische Freiheit – Kohärenz und Theorie*, München: Philosophia.

Schultes, R. 1924: „Die Lehre des hl. Thomas über die Einwirkung Gottes auf die Geschöpfe", *Divus Thomas* 2, 176-195, 277-307.

- 1925: „Die Entwicklung der Stufler-Kontroverse", *Divus Thomas* 3, 360-369, 464-482.

Schwamm, H. 1934: *Das göttliche Vorherwissen bei Duns Scotus und seinen ersten Anhängern*. Innsbruck.

Van Inwagen, Peter 2008: „How to Think about the Problem of Free Will", *Journal of Ethics* 12, 327–341.

Weissmahr, B. 2005: „Selbstüberbietung und die Evolution des Kosmos auf Christus hin", in: Schöndorf, H. (Hrsg): *Die philosophischen Quellen der Theologie Karl Rahners*. Quaestiones Disputatae 213. Freiburg / Basel / Wien: Herder, 143-177.

# 7. Angewandte Ethik, politische Philosophie, Rechts- und Sozialphilosophie

# Problems of Advance Directives in Psychiatry

Simone Aicher

Even though psychiatric advance directives (PADs) are a powerful tool for people who suffer from severe mental illness, professionals like psychiatrists, psychologists and social workers sometimes raise critical voices, pointing to implementation problems on the one hand and questioning its moral authority on the other hand. Although patients' and professionals' attitudes are appropriately studied with the help of empirical research methods, ethical theory poses equally important questions about (surrogate) decision-making as well as (precedent) autonomy, paternalism, coercion, trust and personal identity - concepts relevant to the bioethical debate surrounding PADs. In this very paper, I want to address some philosophically relevant questions concerning PADs: First of all, I am interested in how psychiatric advance directives differ from other medical advance directives (e.g. in palliative care) and why it is important to acknowledge this difference. Second, I will turn to the bioethical principle of autonomy and the concept of voluntary informed consent insofar as it relates to the debate surrounding PADs. Finally, I will focus on the problem of (surrogate) decision-making in case of mentally ill patients and the moral authority of PADs.

## 1. Introduction

Advance directives are a highly debated issue in bioethics. In Germany alone, several committees like the "Kutzer-Kommission", the "Bioethik-Kommission Rheinland-Pfalz" and the "Enquete-Kommission" as well as the German National Ethics Council[1] and the German Medical Association (Bundesärztekammer) entered into discussions about legal and ethical aspects concerning advance directives, which resulted in draft bills[2] and finally in the law reform of 2009.[3] Although advance directives are generally considered legally binding and even more so after the law reform of 2009, (involuntary) hospital commitment which is most relevant in the field of psychiatric care remains a matter for the courts.[4]

While advance directives in the field of terminally ill people are more or less satisfactorily covered by the new legislation in Germany, advance directives in psychiatry have not been discussed extensively within German research literature. Publications on so-called PADs, i.e. psychiatric advanced directives, mainly have their origin in the U.S., the UK and the Netherlands. Still, there are some German publications, including the recent comprehensive brochure edited by *Aktion psychisch Kranke e.V.* in 2010. Furthermore, the Swiss National Advisory Commission on Biomedical Ethics issued a report on advance directives in 2011, which particularly takes into account dementia.[5]

In the present paper I will argue that advance directives are a powerful tool for patients to manage treatment prior to a state of incompetence not only for end of life situations but also amongst people who suffer from severe mental illness. I will address several philosophically relevant questions concerning PADs: First of all, I am interested in how psychiatric advance directives differ from other medical advance directives (e.g. in palliative care) and why it is

[1] See: Nationaler Ethikrat 2005.

[2] See: Landwehr 2007, pp. 67-87.

[3] See: Choi 2010, pp. 30-33. See also: Bundesministerium der Justiz 2012.

[4] See: Olzen & Schneider 2010.

[5] See: NEK-CNE 2011.

important to acknowledge this difference. Second, I will turn to the bioethical principle of autonomy and the concept of voluntary informed consent insofar as it relates to the debate surrounding PADs. Finally, I will focus on the problem of (surrogate) decision-making in case of mentally ill patients and the moral authority of PADs.

## 2. PADs and End-of-Life Advance Directives

Although advance directives for the terminally ill are in some respects similar to advance directives in mental health, there are also several differences that have to be acknowledged. While both instruments are designed to foster patient autonomy, psychiatric patients often have a different appearance from other patients. Most of them are conscious, physically able to express a will, able to give consent or to refuse treatment and not terminally ill. They are often faced with coercion, prevented from harming themselves (i.e. protected) as well as restrained from acting in certain ways. Their freedom of action and will is restricted by external and internal (medication, delusions) forces. They are temporarily (or in random cases permanently) in a difficult mental state. The question is not how they want to die or if they desire artificial life extension. It is not about euthanasia, neither active nor passive. Instead, as Backlar claims, "advance directive[s] for end-of-life decisions and an advance directive for psychiatric treatment are similar kettles which contain quite different fish."[6]

One of the main differences between PADs and medical advance directives which has been neglected in the research literature, stems from the knowledge and experience that the psychiatric patients has about her disorder. Psychiatric advance directives are usually written after a mental crisis has already occurred. They do not emerge "out of the blue". Instead the patient who suffers from mental illness learns from what he experienced during past crisis and she decides how she wants to be treated in the future when another crisis occurs that results in her being no longer competent to make medical decisions for herself. Furthermore, mental diseases are often chronic. As Lidz et al. found out by participant observation, people who suffer from serious chronic diseases are very interested in active participation when it comes to treatment decisions, since they

> cannot so easily give up responsibility for their treatment, deny the reality or seriousness of their illness, and wait to be cured. To do so mean giving up responsibility for larger parts of their life while waiting for something that will probably never come.[7]

The Scottish psychologist Jacqueline Atkinson who wrote a book on *Advance directives in Mental Health* in 2007 summarizes the concept of (P)ADs as follows: "At its simplest, the concept of an advance directive is that, when well (capable/ competent), a person indicates what they want to happen to them when they are ill and, crucially, not capable of making that decision for themselves".[8] Thereby she argues that the most important distinction regarding the structure of PADs is between "opt-in" and "opt-out". To be more precisely: It is possible for mental health patients to exercise precedent autonomy either by giving voluntary informed consent[9] or by refusing treatment in advance. Thereby it is of utmost importance that advance directives are formulated by competent agents who are able to give their consent[10] since otherwise the authority of advance directives is questionable from the outset. In addition to that, the directive only comes into effect when the patient is no longer capable

[6] Backlar 1997, p. 261.
[7] Lidz et al. 2012, p. 305.
[8] Atkinson 2007, p. 39.
[9] See: Helmchen 2010, pp. 209-226.
[10] See: Vollmann 2008.

to make decisions or to give her consent to medical treatment. For if she is capable of making decisions, her current informed consent outweighs the advance directive. In what follows, I will analyze the notions of voluntary informed consent, autonomy and competence insofar as they relate to psychiatric advance directives.

## 3. PADs and the Principle of Autonomy

The principle of autonomy is the most important principle when it comes to advance directives, since "[p]resumably, advance directives are a manifestation of patient autonomy and should be respected for this reason".[11] Therefore, I will focus my analysis on this very principle. Nevertheless, we may not forget that all of Beauchamp and Childress'[12] principles are intertwined: While the right to refuse treatment is closely connected with the principle of non-maleficence, since a physician's behaviour is considered an assault in case he is treating a patient without his consent, the best interest standard is justified by the principle of beneficence. But the act of writing an advance directive and thereby determining how one wants to be treated when no longer competent is mainly based on the principle of autonomy.

The very act of writing down one's wishes and preferences for future treatment already shows that obtaining informed consent is not all there is to respect the autonomy of a particular patient. Instead, the active involvement autonomy implies and requires is very complex. Several attempts have been made to explain what the notion of autonomy means, what it requires and why it is important. Joel Feinberg is one of the first philosophers who has given a "faithfully vague account" of the concept of autonomy in his "twelve-part sketch" which "refers to a congeries of virtues all of which derive from a conception of self-determination, though sometimes by considerable extension of that idea".[13] Amongst the virtues that make up autonomy are self-possession, distinct self-identity, authenticity, self-creation, self-legislation, moral authenticity, moral independence, integrity, self-control and last but not least responsibility for self.

According to a more recent publication by Nomy Arpaly there are eight possible meanings of the term "autonomy" amongst them self-control (agent-autonomy), reason-responsiveness, authenticity and identification with one's actions.[14] Additionally, Gerald Dworkin proposes several "criteria for a satisfactory theory of autonomy" which include logical consistency, empirical possibility, value conditions, ideological neutrality, normative relevance and judgmental relevance.[15] So the first lesson we learn is that the ideal of autonomy has been construed in multiple ways and that it seems impossible for a person whether with or without competence to fulfil all of the components of the ideal of autonomy as described for example by Feinberg, Arpaly or Dworkin.

According to some philosophers, autonomous action is the result of a rational being's own choices and decisions. Rational capacities are a precondition for autonomy – these philosophers argue – and respecting autonomy means respecting a person's rational capacities, which enable her to make her own decisions according to her own reasons and ends.[16] I doubt that rationality plays a central role in acting autonomously. Instead, it seems to be one criterion amongst many others that might be helpful for successfully making decisions. Like Bortoletti et al., I think it is important to first distinguish between "(a) whether one has the capacity to govern oneself and (b) whether one is successful at governing

[11] Brauer 2008, p. 233.
[12] See: Beauchamp & Childress 2001.
[13] See: Feinberg 1986, p. 31.
[14] See: Arpaly 2003, pp. 117ff.
[15] Dworkin 1989, pp. 55f.
[16] See: Schermer 2002, pp. 2-4.

oneself".[17] Furthermore, it is important to keep the concept of autonomy separate from the concept of decision-making competence, even though both notions are central in the present debate.

Additionally, as Radoilska puts it in her recent publication: "we should be able to reliably distinguish between irrational treatment refusals that are protected by an absolute right and incompetent treatment refusals that can be overridden on paternalist grounds, i.e. in the patient's best interests".[18] So-called "incompetent" patients are considered to lack the capabilities that are necessary to make medical decisions. But they may declare in advance, before they enter the state of incompetence, how they want to be treated when an emergency occurs and when they are no longer able to consent or to refuse treatment. This is called "precedent autonomy". The concept of "precedent autonomy" differs from that of autonomy in that it applies only to cases, where a patient that used to be competent in the past and at the time of competency expressed wishes regarding her future treatment, is now incompetent. According to Beauchamp and Childress precedent autonomy means that "whether or not a formal advance directive exists, caretakers should accept prior autonomous judgments".[19] So the existence of an advance directive is not a necessary condition but a helpful hint for respecting someone's precedent autonomy.

While Nancy Rhoden holds the opinion that the autonomy involved in psychiatric advance directives is not any different from ordinary concepts of autonomy[20], most philosophers seem to argue to the opposite. Willigenburg and Delaere claim that it is "not autonomy of sovereignty, but autonomy as authenticity" that is at stake when using advance directives in psychiatry. In fact there is some truth in what they say when arguing that "[p]atients do not experience signing a precommitment directive as a triumph of self-determination".[21] Instead, while advance directives cannot restore a patient's sovereignty, the "deep concerns that constitute his or her identity are safeguarded"[22] and therefore autonomy as authenticity is protected. Davis also provides an argument to solve what he calls the "Delayed Self-Determination Problem" that shows why we have to respect precedent autonomy as much as ordinary autonomy in cases where former preferences are highest order preferences.[23] This is in accord with theories that perceive autonomy as authenticity.

Susanne Brauer also holds the opinion that "the concept of decisional autonomy used for informed consent is not sufficient for capturing the self-determination made possible through advance directives"[24] and that additional components such as "relationship to others" and "character" are needed to define the special kind of autonomy that is involved when talking about advance directives. By appealing to the role other people play in decision-making, Brauer gives a hint to the danger formulated so aptly by Jennings at the end of his article on autonomy in *The Oxford Handbook of Bioethics*, namely that there is "an excessive emphasis on autonomy and too little appreciation of human interdependence and mutual responsibility"[25] inherent in bioethics today. Therefore, relational concepts of autonomy have been discussed during the last few years.[26]

---

[17] Bortoletti et al. 2012, p. 100.

[18] Radoilska 2012, p. xv.

[19] Beauchamp & Childress 2001, p. 137.

[20] See: Rhoden 1990, p. 856.

[21] Willigenburg & Delaere 2005, p. 403.

[22] Ibd., p. 407.

[23] See: Davis 2002.

[24] Brauer 2008, p. 232.

[25] Jennings 2007, p. 88.

[26] See: Mackenzie & Stoljar 1999.

## 4. Voluntary Informed Consent and Thick Respect for Autonomy

Philosophers often distinguish two kinds of respect for persons as autonomous. The first is *simple* respect for autonomy, which is often represented by libertarians. It says that each person has the right to make her own decisions regardless of how rational they are and regardless of potential harmful consequences. Paternalistic intervention is never allowed according to thin respect for autonomy. By contrast, according to thick respect for autonomy, each person has a right to be treated as autonomous in the sense that it is to be ensured that the standards of making rational choices are met. "By insisting on the importance of informed consent we *make* it *possible* for individuals to choose autonomously"[27], O'Neill argues. Paternalistic intervention seems to be allowed to a certain degree – e.g. by seeking voluntary informed consent – if a person obviously is not aware of the consequences of her actions or finds herself in a situation where she cannot make competent decisions.

The criterion of voluntary informed consent is therefore closely connected with thick respect for autonomy. Beauchamp and Childress present a "five-element definition" of informed consent that is widely acknowledged: "One gives an informed consent to an intervention if (and perhaps only if) one is competent to act, receives a thorough disclosure, comprehends the disclosure, acts voluntarily, and consents to the intervention".[28] There has been lots of debate whether voluntary informed consent is necessary and/ or sufficient for autonomy. Many philosophers agree that the notion of autonomy is not sufficiently defined by equating it with voluntary informed consent. As we have seen from the discussion above, there are multiple ways to define autonomy and most of them are not restricted to voluntary informed consent. In medical ethics most philosophers argue that voluntary informed consent is necessary for autonomy. If a person who is able to give voluntary informed consent does not give voluntary informed consent, his right to autonomy is not respected. Still, the right to self-determination does not include the capacity to make use of this right.

Therefore, even though the patient's decision and autonomy should be respected in as many cases as possible, there are cases in which persons lack the capability of making competent decisions. This seems to justify paternalistic intervention:

> Our obligations to respect autonomy do not extend to persons who cannot act in a sufficiently autonomous manner (and who cannot be rendered autonomous) because they are immature, incapacitated, ignorant, coerced, or exploited. Infants, irrationally suicidal individuals, and drug-dependent patients are examples.[29]

There seem to be cases in which it is justified to disregard a patient's autonomy and to act paternalistically on him. There are two forms of paternalism. Soft paternalism says that it is permissible to prevent someone from acting to make sure whether he satisfies conditions of voluntary informed consent. Hard paternalism means that it is permissible to prevent a person from acting in order to prevent harm from coming to that person. I share Gerald Dworkin's opinion that "the difficulty is in specifying in advance, even vaguely, the class of cases in which [paternalistic] intervention will be legitimate".[30] In order to lay down the boundaries for autonomy, it is important to determine under which circumstances it should be allowed to override a person's will, because these cases are an *exception* to the general rule that a person's autonomy always trumps.

---

[27] O'Neill 2002, p. 37.

[28] Beauchamp & Childress 2001, p. 120.

[29] Beauchamp & Childress 2001, p. 105.

[30] Dworkin 1983, p. 33.

## 5. (Surrogate) Decision-Making with the Mentally Ill

According to Radden, most cases in psychiatry are borderline cases, where it is not clear whether patients are competent or not.[31] Even if there are clear results on a formal test, the division into two classes – the "competent" and the "incompetent" – falls short of taking into account that decision-making competence is by its nature a threshold concept. Therefore it is important not to overestimate the results of formal tests that claim to assess competency. Although there are different standards of competence, only a "process standard of decision-making competence"[32] is appropriate. Instead of focusing on the content of a decision, the process of decision-making is evaluated.

It is often asked when to question a patient's competence. Suspicions whether a patient is capable of making a decision are "usually aroused when healthcare workers or proxies do not feel the patient's choice is in his or her best interest."[33] This is often the case when patients do not follow doctors' recommendations or when they refuse treatment. Nevertheless, treatment refusals are not sufficient for declaring a patient incompetent, as Brock argues: "It bears emphasis that mere non-compliance with treatment, or refusal of treatment recommendation, in themselves are no evidence of the patient`s incompetence, but at most should trigger a competence evaluation."[34] Non-compliance and refusal of treatment happen frequently with patients who suffer from a psychiatric disorder. Both Brock and Bærøe point out that there is much room for arbitrariness when it comes to assessing competence in mentally ill patients. They emphasize that people with psychiatric diseases often are not taken serious due to their disease and are therefore particularly vulnerable.

The idea of introducing an instrument to strengthen the autonomy of mentally ill people has its roots in the anti-psychiatric movement. Szasz proposed that mentally ill people use the so-called "psychiatric will"[35], a form of advance directive, to preclude future treatment and to ensure their voices will be heard and their wishes respected before they enter mental hospital or psychiatric treatment in general. It is important to bear in mind that even in psychiatry every medical intervention without consent is considered an assault and therefore unlawful, so it is necessary that the physician provides information to the patient, who in turn makes his or her decision to agree to or to refuse treatment. Although physicians need patients' voluntary informed consent in order to treat them, this is often difficult in the case of mentally ill persons, since "it is probably the case that psychiatric disease affects adversely the decision-making capacities the patient needs in order to be competent to give consent more frequently than does physical illness".[36] Nevertheless mental illness is not a sufficient criterion for assessing incompetence. The assessment of competence and incompetence is a complex procedure. Therefore even in the current literature it is claimed that "the competence of experts for assessing competence is not sufficiently defined yet. This applies to the medical as well as to the psychological and legal experts".[37]

Considering the complexity of this issue, it is not surprising that the first chapter in Buchanan and Brock's book on surrogate decision-making is dedicated to competence and incompetence. They argue that competence in this context means nothing but decision-making capacity and claim that this capacity is always task-relative.[38] A person might be

[31] See: Radden 1994, pp. 797f.
[32] Buchanan & Brock 1990, p. 50.
[33] Bærøe 2010, p. 91.
[34] Brock 1993, p. 255.
[35] See: Szasz 1982.
[36] Brock 1993, p. 247; also Buchanan & Brock 1990, pp. 311, 318f.; Helmchen 2010, p. 218.
[37] Helmchen 2010, p. 224.
[38] See: Buchanan & Brock 1990, p. 18.

capable to decide one thing but might fail to be competent to decide another thing. There are decisions that we consider more fundamental than others. "According to the decision-relative concept of competence, the greater the potential harm to the individual of accepting his or her choice, the higher the standard of competence."[39] Because of the importance of medical decisions to their well-being, most patients choose a physician whom they trust, when they suffer from illness. It is important to them that someone takes care of them who is competent with regards to his profession, but who also honors their own attitude, perspective and values. This is also in accord with the ideal of informed consent which

> recognizes that while the physician commonly brings to the physician-patient encounter medical knowledge, training and experience that the patient lacks, the patient brings knowledge that the physician lacks: knowledge of his or her particular subjective aims and values that are likely to be affected by whatever decision is made.[40]

Therefore, self-determination as a value in itself balances the paternalistic best interest standard which is commonly applied by physicians. But even if deciding autonomously is important to many patients, not all of them are capable to exercise their rights. So the main question seems to be: What abilities does competence require? According to Buchanan and Brock there are three major prerequisites for making competent decisions. The first is *understanding and communication*, the second is *capacities for reasoning and deliberation* and the last is *a set of values or conception of what is good*.[41] Helmchen names several structured tools for assessing competence.[42] The most important tool is the MacCAT-T.[43]

In case incompetence is assessed, there is a fixed order of "guidance principles"[44] that holds during the decision-making process. These principles do partly contradict each other insofar as they are aiming either at the patient's right to self-determination and autonomy or at what is in his best interest. The first principle that comes into effect when incompetency is assessed is the advance directive principle. When a patient is considered incompetent, the first thing that is taken into account in order to figure out how the patient wants to be treated is the advance directive. If the patient did not make an advance directive or if it is invalid or of no use for some reason the substituted judgment principle comes into play. A surrogate decision-maker is appointed whose task it is to decide how the patient would have decided, if he were competent. Value histories are an important tool to figure out what the patient would have wanted.[45] Finally, if for some good reason the substituted judgment principle also fails, there is a last principle to guide the decision-making process. This is the principle of best interest. Instead of fostering a patient's autonomous will, it works on paternalistic grounds. Advance directives take precedence over the other guidance principles since respect for autonomy trumps paternalism. It is important to inquire the moral authority of advance directives also in order to find out why so "many physicians tend to be unduly paternalistic."[46]

## 6. Moral Authority of PADs and Overriding PADs

Concerns about the moral authority of advance directives often have to do with the possibility of overriding (P)ADs. Atkinson lists more than twenty reasons, why PADs are overridden.[47]

[39] Ibd., p. 62.
[40] Ibd., pp. 29f.
[41] See: Ibd., pp. 23-25.
[42] See: Helmchen, pp. 221ff.
[43] See: Grisso et al. 1997.
[44] Buchanan & Brock 1990, pp. 93ff.
[45] See: Buchanan & Brock 1990, p. 121.
[46] Buchanan & Brock 1990, p. 111.
[47] See: Atkinson 2007, pp. 68f.

One reason for overriding advance directives is the principle of beneficence. Physicians may for paternalistic reasons disregard an advance directive and act in line with their own professional opinion, since their superior status gives them the power to question the authority of such documents. Not only do they have more knowledge about the patient's disease, they are also the ones who have to carry out the decision. Compliance with advance directives therefore is not self-evident. Referring to the Pennsylvania Act 194, which they consider a PAD statute that is "quite typical", Swanson et al. list "three specific sections devoted to ensuring that physicians can override these directives with few (if any) consequences".[48]

First of all, civil commitment law is always more powerful than any PAD statute. This concerns legislation inside as well as outside the U.S. So it is possible that involuntary committed patients are "trapped" even if they are competent, since according to the law, they have no right to refuse treatment. Nevertheless, this was called into question by the federal court's decision in the case *Hargrave v. Vermont* "which challenged the state's power to treat an involuntary patient whose PAD indicated her preference to avoid all antipsychotic medications"[49]. Second, another section in Pennsylvania's PAD statute made it clear that physicians cannot be forced to comply with an advance directive if their conscience does not allow it. They may refuse complying with a directive, if it infringes accepted clinical standards. Last but not least, "a physician who acts in good faith may not be subject to criminal or civil liability or disciplined for unprofessional conduct as a result of refusing to comply with a PAD".[50] Many of the legislative regulations weaken PADs as an instrument for self-determination.

But reasons for overriding PADs are not only grounded in the noncompliance of physicians. As Brock states, there are three more situations that trigger overriding PADs.[51] Doubt whether the advance directive really states what the patient would have wanted suffices for superseding PADs. Another reason for overriding directives is given if there are interests of other persons involved that justify ignoring the patients previous wishes. Finally, when there are conflicts between what the directive states and the interests that are currently attributed to the patient or when the directive is in conflict with the (altered) identity of the individual, this also gives rise to overriding an advance directive. Kennett argues that "[a] person suffering from a mental illness or disorder may differ dramatically from his or her previous well self".[52]

## 7. The "Non-Identity Thesis" alias the "Someone Else Problem"

A frequently discussed reason for overriding advance directives and questioning its moral authority is the appeal to the so-called "nonidentity thesis" or the "someone else problem"[53], which claims that the author of the advance directive is not identical to the person to whom it would apply. This is what Buchanan calls "the slavery argument, since it portrays advance directives not as vehicles for self-determination, but as sinister devices to subjugate other persons".[54] Although Buchanan and DeGrazia point out that the reasoning involved in this kind of argument is not sound[55], there is a huge amount of literature that takes the problem of

[48] Swanson et al. 2006b, p. 386.
[49] Appelbaum 2006, p. 397.
[50] Ibd.
[51] See: Brock 1991.
[52] Kennett 2007, p. 91.
[53] DeGrazia 2005, p. 165.
[54] Buchanan 1988, p. 282.
[55] See also: Quante 1999.

personal identity regarding the moral authority of advance directives serious. Even as early as in 1979 Jon Elster explained this phenomenon involved in self-binding by saying that "[i]n somewhat fanciful terms we might speak here of an alliance between the early and the late self against the intermediate and more docile self".[56] Especially when talking about people with mental illness, the question of personal identity arises. As Jennifer Radden explains: "Episodes of mental disorder expunge and distort memories and change cognitive functions, beliefs, and values; they alter capabilities, personality, mood, emotional style, and response. They disrupt normal psychological functioning of all kinds and interrupt in lives in ways that are often devastatingly far reaching".[57] Therefore it is important to turn to issues of personal identity, when discussing the moral authority of advance directives in psychiatric care.

The main issue regarding the so called "someone else problem" is, whether there are in fact different "selves" involved and if so, whether they represent different "persons". Personhood is one of the most important concepts in bioethics, especially in the debates on abortion and euthanasia. It seems difficult to grasp the notion of "personhood", although several philosophers tried to formulate necessary and sufficient criteria for what it means to be a person. When it comes to advance directives and especially to PADs, it is common to appeal to Derek Parfit's psychological view[58] about personal identity. Parfit claims that psychological continuity is the main criterion for our persistence over time and for the definition of personhood and identity. Although there are many more philosophers, who dealt with issues of personhood and personal identity, it is common to ask along Parfitian lines whether personal identity is to be understood in terms of psychological continuity and if so, how much continuity we need in order to speak of one and the same person. Buchanan, like many others, discusses Parfit's example of the "Russian nobleman" and comes to the conclusion that it does not show, along with his other examples, that "neurological damage is severe enough to undercut the moral authority of an advance directive by destroying its author while leaving in his place another *person*".[59] Nevertheless he adheres to the Parfitian thesis "that psychological continuity is at least a necessary condition for personal identity" and takes it as a premise to argue that "neurological damage [...] destroys at least some of the necessary conditions for personhood".[60]

DeGrazia challenges the psychological continuity view in a powerful way. He starts with the basic assumption that we are essentially animals and not persons. Therefore, "we must reject the psychological-continuity view of our identity. The psychological-continuity theory is a plausible account only for persons."[61] With regards to numerical identity DeGrazia's opinion seems to be clear: He argues in favor of "the biological view, which holds that we are essentially human animals and that human identity consists in sameness of biological life"[62]. But numerical identity is not all there is when it comes to (personal) identity. As French phenomenologist Paul Ricoeur has argued at the end of the third volume of *Time and Narrative*, there is also narrative identity that may help explaining many of the "puzzling-cases" introduced by Derek Parfit and that "offers a solution to the aporias concerning personal identity"[63]. DeGrazia agrees with Ricoeur that

[56] Elster 1979, p. 41.
[57] Radden 2004, p. 133.
[58] Parfit 1984, pp. 204ff.
[59] Buchanan 1988, p. 292.
[60] Ibd., p. 299.
[61] DeGrazia 1999, p. 387.
[62] DeGrazia 2005, p. 73.
[63] Ricoeur 1991, p. 76.

> the narrative constructs the durable character of an individual, which one can call his or her narrative identity, in constructing the sort of dynamic identity proper to the plot [...] which creates the identity of the protagonist in the story.[64]

Although DeGrazia argues that numerical identity is a necessary condition for narrative identity[65], it is the latter "sense of human identity that most concerns people in everyday life"[66].

## 8. How to Deal with Revocations

There is one final aspect that I want to discuss regarding the moral authority of advance directives. PADs not only lose their power when they are overridden by professionals, as has been shown earlier, they can also be revoked by their authors. There is much debate about whether revocation of PADs (and especially of so-called "Ulysses contracts"[67]) should be possible at any time, i.e. not only when the patient is competent to refuse treatment[68], but even if he is considered incompetent.[69] One reason for this debate is that "states that have adopted specific statutes about mental health care directives, all note the irrevocability of the directive after loss of capacity. In states without such statutes, the necessity of competency for revocation of mental health advance directives is less clear"[70].

Radden gave a detailed analysis of second thoughts by authors of advance directives. She borrows the idea from Feinberg that there are "hard cases" and "easy cases". While we are clear about honoring an earlier commitment and ignore possible future revocations when we are confronted with "easy cases", we do not know whether we should give more weight to precommitments or rather to revocations when we are confronted with "hard cases". The case of Ulysses is easy since there we are dealing with a situation "where the later self's contrary 'choice' results from coercion or fraud"[71]. But when it comes to "hard cases", there is no "qualitative difference"[72] between the earlier commitment and the later revocation and therefore it is not clear whether the advance directive is supposed to be honored. There are certain circumstances that indicate when the earlier commitment must be honored, because the latter decision is the result of undue influences. These circumstances are presented by Radden in "an extended version of Feinberg's list: (1) incompetence (2) moral weakness (3) coercion or fraud (4) distracting states (of pain, emotion, etc.) and 'promises'"[73]. If due to these circumstances a patient is revoking his advance directive, we may ignore the revocation.

For Radden the conflict with "hard cases" lies in the fact that the revocations are to be considered autonomous and therefore cannot be trumped by the earlier commitment with reference to an argument of precedent autonomy that outweighs later revocations. "Our ability to entertain second thoughts," she argues "to reconsider, adapt and change direction in the light of a new piece of information, or a telling experience, is deeply bound up with what makes us autonomous human beings".[74] She holds the opinion that "hard cases" are common in the psychiatric sphere. Not being satisfied with the rough distinction between "hard" and

---

[64] Ricoeur 1991, p. 77.
[65] DeGrazia 2005, p. 114.
[66] Ibd., p. 113.
[67] See: Hallich 2011.
[68] See: Davis 2008.
[69] See: Srebnik 2005, p. 43.
[70] Srebnik & La Fond 1999, p. 922.
[71] Radden 1994, p. 789.
[72] Ibd., p. 791.
[73] Ibd.
[74] Ibd., p. 793.

"easy" cases, Radden even goes deeper into the matter by analyzing the phenomenological differences that occur when we change our mind. A change of mind that is characterized by a change of values, beliefs and attitudes generally is to be respected as an expression of autonomy. By contrast, the revocation of an earlier commitment often is due to an inner struggle, a state of ambivalence, a divided mind which may not be mixed up with what Radden considers a competent change of mind. Other than Feinberg, she argues that the revocation trumps the earlier commitment when it comes to hard cases, since changes of mind are central to the concept of autonomy. Last but not least, a patient's "present embodiment constitutes a kind of privileged relationship. It is she, after all, who must endure most immediately the suffering, mental and physical, which her decision imposes"[75].

Other than Radden, whose analysis of revocation is grounded in philosophical theory, Srebnik conducted empirical studies in order to find out whether authors of advance directives wish to be able to revoke the directives they have written even in case of incapacity. She showed that of 106 persons, 40% wanted that their directives are revocable at any time, i.e. even in times of incompetency.[76] Srebnik also came to the conclusion that revocation of psychiatric advance directives is a "rare event"[77] and that a good solution to this problem would be to give authors of advance directives the opportunity to choose whether they want their directive to be revocable or irrevocable.

## 9. Conclusion

In sum, the differences between PADs and advance directives at the end of life are important even though there are of course similarities. The most important difference that has been widely neglected in the research literature stems from knowledge and experience with prior crises and treatment. While problems with anticipating future health care situations occur frequently with advance directives for end of life situations, problems in anticipating future situations are hardly a hindrance for psychiatric advance directives. Nevertheless, there are more hindrances with regards to decision-making competence in people with mental diseases and the question occurs whether PADs really have the power to foster patient autonomy in mental health patients.

It turns out that the notions of "autonomy" and "voluntary informed consent" that have been examined by many authors in the bioethical debate are too idealistic. And even though rationality is not necessary for autonomy, it is still helpful for making wise decisions. Additionally, the special character of what has been called "precedent autonomy" needs to be taken into account. Recently, relational autonomy and social embeddedness of decisions has become more and more important. Furthermore, we need to bear in mind that decision-making competence is neither the same as autonomy nor as voluntary informed consent. The notion of competence is the most important notion – since it is a precondition for voluntary informed consent and therefore relevant for thick respect for autonomy – even though it is particularly difficult to assess decision-making competence in mental health patients.

The prerequisites for moral and legal authority of PADs are strict. The fact that it is legally possible that PADs are overridden by professionals questions the authority of PADs. Revocation of PADs by the patient is also a familiar phenomenon that often has to do either with second thoughts of the authors of PADs or with a change in the author's personality. This leads to the so-called "Non-Identity Problem". The psychological continuity view is not the only answer to solve this problem. Narrative identity in addition to biological continuity might be a better account for dealing with revocations.

---

[75] Radden 1994, p. 800.

[76] See: Srebnik 2005, p. 43.

[77] Ibd., p. 44.

Discussion of philosophical problems of advance directives in psychiatry therefore have to take into account 1) the difficulty of assessing decision-making competence in mentally ill people, 2) the difficulty of defining and respecting (precedent) autonomy in the special case of advance directives, 3) the difficulty of deciding over the moral authority of PADs and making decisions on behalf of others and finally 4) explaining revocation of PADs by the authors including issues of altered identity.

**Simone Aicher**

Universität Regensburg
simone.aicher@gmx.de

## References

Aktion Psychisch Kranke e. V. (eds.) 2010: *Dokumentation des Workshops „Patientenverfügung und Behandlungsvereinbarung bei psychischen Erkrankungen" am 07. Juli 2010 in Berlin, Rathaus Schöneberg*. Bonn: Psychiatrie-Verlag.

Appelbaum, P. S. 2006: 'Commentary: Psychiatric Advance Directives at a Crossroads –When Can PADs be Overridden?', *Journal of the American Academic Psychiatry Law* 34(3), 395-397.

Arpaly, N. 2003: *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Atkinson, J. M. 2007: *Advance Directives in Mental Health: Theory, Practice and Ethics*. London: Jessica Kingsley .

Backlar, P. 1997: 'Ethics in community mental health care. Anticipatory Planning for Psychiatric Treatment Is Not Quite the Same as Planning for End-of-Life Care', *Community Mental Health Journal* 33(4), 261-268.

Bærøe, K. 2010: 'Patient autonomy, assessment of competence and surrogate decision-making: a call for reasonableness in deciding for others', *Bioethics* 24(2), 87-95.

Beauchamp, T. L. and Childress, J. F. 2001: *Principles of Biomedical Ethics*. 5th edition. New York: Oxford University Press.

Bortoletti, L., R. Cox, M. Broome, and M. Mameli 2012: 'Rationality and self-knowledge in delusion and confabulation: implications for autonomy as self-governance', in L. Radoilska (ed.) 2012, 100-122.

Brauer, S. 2008: ‚Die Autonomiekonzeption in Patientenverfügungen ‑ Die Rolle von Persönlichkeit und sozialen Beziehungen', *Ethik in der Medizin* 20(3), 230-239.

Brock, D. W. 1991: 'Trumping advance directives', *Hastings Center Report* 21(5), 5-6.

Brock, D. W. 1993: 'A proposal for the use of advance directives in the treatment of incompetent mentally ill persons', *Bioethics* 7(2/3), 247-256.

Buchanan, A. 1988: 'Advance Directives and the Personal Identity Problem', *Philosophy & Public Affairs* 17(4), 277-302.

Buchanan, A. E., and D.W. Brock 1990: *Deciding for Others: The Ethics of Surrogate Decision Making*. Cambridge: Cambridge University Press.

Bundesministerium der Justiz: *Pressemitteilung: Betreuungsrecht – Neuregelung hilft psychisch Kranken*. Erscheinungsdatum: 07.11.12.

Capron, A. M. 2009: 'Advance Directives', in H. Kuhse and P. Singer (eds.) 2009, 299-311.

Choi, J. S. 2010: *Patientenverfügung und Patientenautonomie zwischen Rechtsdogmatik und Rechtswirklichkeit*. Frankfurt am Main: Lang.

Davis, J. K. 2002: 'The concept of precedent autonomy', *Bioethics* 16(2), 114-133.

Davis, J. K. 2007: 'Precedent Autonomy, Advance directives, and End-of-Life Care', in: B. Steinbock (ed.) 2007, 349-374.

Davis, J. K. 2008: 'How to justify enforcing a Ulysses contract when Ulysses is competent to refuse', *Kennedy Institute of Ethics Journal* 18(1), 87-106.

DeGrazia, D. 1999: 'Advance Directives, Dementia, and "the Someone Else Problem"', *Bioethics* 13(5), 373-391.

DeGrazia, D. 2005: *Human Identity and Bioethics*. Cambridge: Cambridge University Press.

Dresser, R.: 'Advance directives, self-determination, and personal identity', in C. Hackler, R. Moseles, and D.E. Vawter (eds.) 1989, 155-170.

Dworkin, G. 1983: 'Paternalism', in: R. Sartorius (ed.) 1983, 19-34.

Dworkin, G. 1988: *The theory and practice of autonomy*. Cambridge: Cambridge University Press.

Elster, J. 1979: *Ulysses and the sirens. Studies in rationality and irrationality*. Cambridge: Cambridge University Press.

Feinberg, J. 1986: *The Moral Limits of the Criminal Law. Vol. III. Harm to self*. Oxford: Oxford University Press.

Grisso, T., P.S. Appelbaum, and C. Hill-Fotouhi 1997: 'The MacCAT-T: A Clinicl Tool to Assess Patients' Capacities to Make Treatment Decisions', *Psychiatric Services* 48(11), 1415-1419.

Hackler, C., R. Moseley, and D.E. Vawter 1989: *Advance directives in medicine*. New York: Praeger.

Hallich, O. 2011: 'Selbstbindungen und medizinischer Paternalismus. Zum normativen Status von "Odysseus-Anweisungen"', *Zeitschrift für philosophische Forschung* 65, 151-172.

Helmchen, H. 2010: *Ethics in psychiatry: European contributions. International library of ethics, law, and the new medicine: Vol. 45*. Dordrecht: Springer.

Holland, S. 2012: *Arguing about bioethics*. New York: Routledge.

Jennings, B. 2007: 'Autonomy', in B. Steinbock *(ed.), 72-90*.

Kennett, J. 2007: 'Mental Disorder, Moral Agency, and the Self', in B. Steinbock *(ed.) 2007, 90-114*.

Kuhse, H., and P. Singer 2009: *A Companion to Bioethics*. 2nd edition. Malden, MA: Wiley-Blackwell.

Lidz, C. W., A. Meisel, M. Osterweis, J.L. Holden, J.H. Marx, and M.R. Munetz 2012: 'Barriers to Informed Consent', in S. Holland (ed.) 2012, 299-307.

Landwehr, C. 2007: *Selbstbestimmung durch Patientenverfügung*. Regensburg: Roderer.

Mackenzie, C. and N. Stoljar 1999: *Relational autonomy. Feminist perspectives on automony, agency, and the social self*. New York: Oxford University Press.

Nationale Ethikkommission im Bereich Humanmedizin 2011: *Patientenverfügung. Stellungnahme Nr. 17/2011*. Retrieved from www.nek-cne.ch. (08/21/2012).

Nationaler Ethikrat 2005: *Patientenverfügung – Ein Instrument der Selbstbestimmung. Stellungnahme*. Retrieved from www.ethikrat.org. (08/21/2012).

Olzen, D., and F. Schneider 2010: 'Das Patientenverfügungsgesetz (PatVG) vom 1. 9. 2009 – Eine erste Bilanz. Unter besonderer Berücksichtigung der Auswirkungen auf die Unterbringung psychisch Kranker', *Medizinrecht* 28(11), 745-751.

O'Neill, O. 2002: *Autonomy and trust in bioethics*. Cambridge: Cambridge University Press.

Parfit, D. 1984: *Reasons and persons*. Oxford: Clarendon Press.

Quante, M. 1999: 'Precedent Autonomy and Personal Identity', *Kennedy Institute of Ethics Journal* 9(4), 365-81.

Radden, J. 1994: 'Second Thoughts: Revoking Decisions Over One's Own Future', *Philosophy and Phenomenological Research* 54(4), 787-801.

Radden, J. 2004: 'Identity: personal identity, characterization, identity and mental disorder', in J. Radden (ed.) 2004, 133-146

Radden, J. (ed.) 2004: *The Philosophy of psychiatry: a companion*. New York: Oxford University Press.

Radoilska, L. 2012: *Autonomy and mental disorder*. Oxford: Oxford University Press.

Radoilska, L.: 'Introduction: personal autonomy, decisional capacity, and mental disorder', in L. Radoilska (ed.) 2012, ix-xli.

Rhoden, N. K. 1990: 'The Limits of Legal Objectivity', *North Carolina Law Review* 68, 845-865.

Ricoeur, P. 1991: 'Narrative Identity', *Philosophy Today* 35(1), 73-81.

Sartorius, R. 1983: *Paternalism*. Minneapolis: University of Minnesota Press.

Schermer, M. 2001: *The Different Faces of Autonomy. Patient Autonomy in Ethical Theory and Hospital Practice.* Dordrecht: Kluwer Academic Publishers.

Srebnik, D. S., and La Fond, J. Q. 1999: 'Advance Directives for Mental Health Treatment', *Psychiatric Services* 50(7), 919-925.

Srebnik, D. 2005: 'Issues in applying advance directives to psychiatric care in the United States', *Australasian Journal on Ageing* 24, 42-45.

*Steinbock, B. 2007:* Oxford Handbook of Bioethics. Oxford: Oxford University Press.

Swanson, J.W., Van McCrary, S., Swartz, M.S., Elbogen, E.B., and Van Dorn, R. A. 2006: 'Superseding Psychiatric Advance Directives: Ethical and Legal Considerations', *Journal of the American Academy of Psychiatry and the Law* 34(3), 385-94.

Szasz, T. 1982: 'The psychiatric will', *American Psychologist* 37(7), 762-770.

Van Willigenburg, T., and P.J.J. Delaere 2005: 'Protecting Autonomy as Authenticity Using Ulysses Contracts', *Journal of Medicine and Philosophy* 30(4), 395-409.

Vollmann, J. 2008: *Patientenselbstbestimmung und Selbstbestimmungsfähigkeit. Beiträge zur klinischen Ethik*. Stuttgart: Kohlhammer.

# Bildung als Gegenstand der fairen Chancengleichheit bei Rawls

Claudia Blöser

Gegenstand dieses Beitrags ist eine der wichtigsten Gerechtigkeitsforderungen im Bildungsbereich, die Forderung nach Chancengleichheit. Zentrale Frage ist, welche Aussagen Rawls' Gerechtigkeitstheorie in Bezug auf die *Realisierung* von Chancengleichheit machen kann. Zunächst wird dargestellt, was faire Chancengleichheit in Bezug auf Ämter und Positionen in Rawls' zweitem Gerechtigkeitsgrundsatz bedeutet, um in einem zweiten Schritt für die These zu argumentieren, dass die Realisierung von Chancengleichheit in Bezug auf Ämter und Positionen notwendig voraussetzt, dass Chancengleichheit in der Bildung verwirklicht wird, d.h. dass Bildung der primäre Gegenstand der Chancengleichheit sein muss. Im dritten Teil diskutiere ich als grundlegende Schwierigkeit bei der Realisierung von Chancengleichheit den Einfluss der Familie. Angesichts dieses Faktors scheint Rawls in seinem Werk *Gerechtigkeit als Fairness* zu resignieren, was die Realisierbarkeit des Chancenprinzips und damit dessen Vorrang vor dem Differenzprinzip angeht. Dieses Problem werde ich im Rahmen der Debatte um ideale und nicht-ideale Theorie bei Rawls diskutieren. Es ist idealen Theoretikern vorgeworfen worden, im besten Fall irrelevante, im schlimmsten Fall ideologisch verzerrende Beiträge zur Lösung von Gerechtigkeitsproblemen unter nicht-idealen Bedingungen zu liefern. In der Tat bietet das Problem der Chancengleichheit Anlass zu derartiger Kritik an Rawls' idealer Theorie, die deshalb der Ergänzung durch Regeln einer nicht-idealen Theorie bedarf.

In diesem Beitrag werde ich eine der wichtigsten Gerechtigkeitsforderungen im Bildungsbereich, die Forderung nach Chancengleichheit, auf der Grundlage von Rawls' Gerechtigkeitstheorie beleuchten. Meine zentrale Fragestellung ist, welche Aussagen die Rawls' Theorie in Bezug auf die *Realisierung* von Chancengleichheit machen kann.

Im ersten Teil stelle ich dar, was „faire Chancengleichheit in Bezug auf Ämter und Positionen" in Rawls' zweitem Gerechtigkeitsgrundsatz bedeutet. In einem zweiten Schritt werde ich für die These argumentieren, dass die Realisierung von Chancengleichheit in Bezug auf Ämter und Positionen notwendig voraussetzt, dass Chancengleichheit in der *Bildung* verwirklicht wird. Das bedeutet, dass Bildung als primärer Gegenstand der Chancengleichheit gelten muss. Im dritten Teil diskutiere ich grundlegende Schwierigkeiten bei der Realisierung von Chancengleichheit und setze diese in Bezug zur Debatte um ideale und nicht-ideale Theorie bei Rawls.[1] Es ist idealen Theoretikern vorgeworfen worden, im besten Fall irrelevante, im schlimmsten Fall (ideologisch) verzerrende Beiträge zur Lösung von Gerechtigkeitsproblemen unter nicht-idealen Bedingungen zu liefern.[2] Das Problem der Chancengleichheit bietet in der Tat Anlass zu derartiger Kritik an der idealen Theorie. Auf der Grundlage dieser Kritik wird deutlich, dass Rawls' ideale Theorie der Ergänzung durch Regeln einer nicht-idealen Theorie bedarf.

[1] Eine Rekonstruktion der methodologischen Unterscheidung zwischen idealer und nicht-idealer Theorie bei Rawls liefert z.B. Simmons 2010. Vgl. für einen Überblick über die Debatte auch Schaub 2010.

[2] Vgl. z.B. Boettcher (Boettcher 2009, 238).

## 1. Was heißt „faire Chancengleichheit“ bei Rawls?

Das Prinzip der fairen Chancengleichheit (im folgenden auch kurz „Chancenprinzip“ genannt) spielt in Rawls' zweitem Gerechtigkeitsgrundsatz eine zentrale Rolle. Dieser Grundsatz formuliert die Bedingungen für gerechtfertigte soziale und ökonomische Ungleichheiten in zwei Teilen: Das erste Teilprinzip ist das *Chancenprinzip*, das fordert, dass Ungleichheiten nur die Folge einer Verteilung von „Ämtern und Positionen“ sein dürfen, „die allen unter Bedingungen fairer Chancengleichheit offenstehen“ (vgl. z.B. PL, 70f.). Das zweite Teilprinzip, das *Differenzprinzip*, besagt, dass sich die Ungleichheiten „zum größtmöglichen Vorteil für die am wenigsten begünstigten Gesellschaftsmitglieder auswirken“ (ebd.) müssen.

Obwohl Rawls dem Chancenprinzip einen prominenten Stellenwert einräumt, äußert er sich nur an wenigen Stellen zu der Frage, worin die geforderte faire Chancengleichheit genauer besteht. Da es im Chancenprinzip um die Verteilung von Ämtern und Positionen geht, liegt es nahe, die Forderung nach Chancengleichheit als Forderung nach *diskriminierungsfreien gesetzlichen Rechten* auf berufliche Positionen zu verstehen. Tatsächlich ist es ein Bestandteil der fairen Chancengleichheit, den Zugang zu Ämtern und Positionen gesetzlich so zu regeln, dass keine Gruppen aus religiösen, ethnischen oder anderen Gründen prinzipiell ausgeschlossen oder benachteiligt werden. Doch diese Art der Chancengleichheit nennt Rawls bloß „formal“ (TdG, 92). Was der formalen Chancengleichheit nach Rawls fehlt, ist die Berücksichtigung von „natürlichen und gesellschaftlichen Zufälligkeiten“ (TdG, 92). Rawls bezieht sich damit auf die Tatsache, dass Menschen natürliche und gesellschaftliche Bedingungen als gegeben vorfinden, die ihre Chancen auf Ämter und Positionen beeinflussen, ohne dass sie selbst Einfluss auf diese Bedingungen nehmen können – in diesem Sinn sind die Faktoren „zufällig“. Die beiden Klassen von Zufälligkeiten sind (i) natürliche Fähigkeiten und Leistungsbereitschaft und (ii) die ökonomische Stellung der Familie.

Über die formale Chancengleichheit hinaus hängen die realen Möglichkeiten einer Person stark davon ab, welche natürlichen Fähigkeiten und Talente sie hat und welche ökonomischen Möglichkeiten ihre Familie besitzt. Diese Abhängigkeit kritisiert Rawls als „die krasseste Ungerechtigkeit“ (TdG, 93), da die genannten Faktoren „unter moralischen Gesichtspunkten so willkürlich sind“ (ebd.). Unter Berücksichtigung dieser Überlegung gelangt Rawls zu folgender Bestimmung der fairen Chancengleichheit:

> In allen Teilen der Gesellschaft sollte es für ähnlich Begabte und Motivierte auch einigermaßen ähnliche kulturelle Möglichkeiten und Aufstiegschancen geben. Die Aussichten von Menschen mit gleichen Fähigkeiten und Motiven dürfen nicht von ihrer sozialen Schicht abhängen (TdG, 93). [3]

Diese Formulierung der Bedeutung fairer Chancengleichheit ist in einer Hinsicht überraschend. Nach Rawls gibt es, wie erwähnt, zwei Arten von Bedingungen – nämlich natürliche Fähigkeiten und ökonomische Stellung der Familie –, die gleichermaßen zufällig sind. So sagt Rawls:

> [W]enn man einmal mit dem Einfluß entweder gesellschaftlicher oder natürlicher Zufälle auf die Verteilung unzufrieden ist, dann wird man durch Nachdenken dazu geführt, mit *beidem* unzufrieden zu sein. Vom moralischen Gesichtspunkt aus erscheint beides als *gleich willkürlich* (TdG, 95, H.v.m.).

Eine Person kann weder etwas dafür, mit welchen natürlichen Fähigkeiten sie auf die Welt kommt, noch hat sie Einfluss darauf, in welche soziale Schicht sie hineingeboren wird. Und dennoch besteht faire Chancengleichheit nach Rawls nicht darin, die Verteilung von Ämtern und Positionen unabhängig von *beiden* zufälligen Faktoren zu machen, sondern nur von sozial-ökonomischen Zufälligkeiten. Er stellt nicht die Maximalforderung, dass der Zugang zu

[3] Diese Bestimmung behält Rawls in *Gerechtigkeit als Fairness* bei (vgl. GaF, 79).

Ämtern und Positionen für *alle* Menschen gleichwahrscheinlich sein soll, sondern nur dass er für alle Menschen *mit denselben natürlichen Fähigkeiten (*ungefähr) gleichwahrscheinlich sein soll. Der zufällige Faktor der natürlichen Fähigkeiten darf also Rawls zufolge weiterhin eine diskriminierende Rolle bei der Vergabe von Ämtern spielen.

Es lassen sich meines Erachtens zwei plausible Gründe für Rawls' Position angeben. *Erstens* müssen nach dem ersten Gerechtigkeitsgrundsatz die Grundrechte und Grundfreiheiten einer Person geschützt werden. Zu diesen zählt auch die freie Berufswahl. Man kann nun eine Person, die für eine bestimmte Position die entsprechenden Begabungen mitbringt, z.B. pädagogisches Geschick vorweist, nach dem ersten Gerechtigkeitsgrundsatz nicht davon abhalten, den Lehrerberuf anzustreben. Bleibt zu begründen, warum die begabtere bzw. geeignetere Person bei der Bewerbung auf dieselbe Stelle der unbegabteren vorzuziehen ist. Das Effizienzargument, dass eine bessere Ausübung der Ämter und Positionen einer schlechteren vorzuziehen ist, und dadurch größerer gesellschaftlicher Wohlstand erreicht werden kann, steht Rawls nicht zur Verfügung, da Chancengleichheit gegenüber gesellschaftlicher Wohlfahrt Vorrang hat. Als Rawlssches Argument kann gelten, dass es die *Achtung vor der Person* erfordert, bei der Bewerbung einzig ihr Eignungspotential zu berücksichtigen. Andere Verfahren könnten dazu führen, die sozialen Grundlagen der Selbstachtung zu untergraben.[4]

Ein zweiter Grund, der für Rawls' Vorgehen spricht, von der Maximalforderung der Chancengleichheit Abstand zu nehmen, ist, dass die Realisierung der Chancengleichheit in diesem maximalen Sinn schlicht *unmöglich* zu sein scheint. Es scheint aussichtslos, Menschen mit völlig unterschiedlichen Fähigkeiten in die Lage zu versetzen, dass jede und jeder jeden beliebigen Beruf ausüben kann. Ob ein solches Szenario überhaupt wünschenswert ist, sei dahingestellt, aber es ist sicherlich unrealistisch. So erscheint es beispielsweise sinnlos, einer Person mit „zwei linken Händen" den Beruf eines Chirurgen zuzumuten.

Das Chancenprinzip fordert also nur gleiche Chancen auf Ämter und Positionen unabhängig von gesellschaftlichen Zufälligkeiten und *nicht* unabhängig von natürlichen Zufälligkeiten. Rawls' These, dass die Abhängigkeit der Einkommensverteilung von natürlichen Fähigkeiten unter moralischen Gesichtspunkten ebenso ungerecht ist wie die Abhängigkeit von gesellschaftlichen Zufälligkeiten, findet schließlich doch noch in seinem Differenzprinzip Berücksichtigung, das die „willkürlichen Wirkungen der natürlichen Lotterie" mildern soll (TdG, 94). Dieses Prinzip besagt, dass „die besseren Aussichten der Begünstigten genau dann gerecht [sind], wenn sie zur Verbesserung der Aussichten der am wenigsten begünstigten Mitglieder der Gesellschaft beitragen" (TdG, 96). An einem Beispiel kann die Grundidee verdeutlicht werden: Das Prinzip der fairen Chancengleichheit besagt, dass zwei gleichermaßen begabte Personen die gleiche Chance haben sollen, Arzt zu werden, obgleich eine der beiden aus einer Familie stammt, in der die Eltern nicht studiert haben und nicht die finanziellen Mittel haben, ihrem Kind das Studium zu ermöglichen. Wenn jedoch aufgrund verschiedener natürlicher Fähigkeiten nur eine der beiden Personen dazu in der Lage ist, den Arztberuf auszuüben, während die zweite nicht die entsprechende Begabung besitzt, soll die zweite laut Differenzprinzip *Vorteile* aus der guten Stellung der ersten haben.[5]

---

[4] Vgl. zu diesem Punkt auch Wallimann 2008, 25, außerdem Sher 1988. Eine andere Möglichkeit, die in diesem Zusammenhang diskutiert wird, ist die Einführung von Quoten, die ja gerade die Annahme in Frage stellen, dass nur Eignung bzw. Begabung der ausschlaggebende Faktor bei der Vergabe von Ämtern sein soll. Für eine Diskussion siehe Rössler 1993.

[5] Es ist eine interessante Frage, ob aus Rawls' Differenzprinzip impliziert, dass die weniger Begabten besonders gefördert werden sollten. Auf den ersten Blick scheint dies nahezuliegen, denn das Differenzprinzip richtet den Blick auf die am schlechtesten Gestellten in der Gesellschaft und fordert, deren Vorteile zu maximieren, d.h. auch im „Bildungswesen die Anstrengungen auf die Verbesserung der langfristigen Aussichten der am wenigsten Bevorzugten lenken" (TdG, 122). Allerdings könnte dies

Alles in allem scheint Rawls' Konzeption der fairen Chancengleichheit eine vielversprechende Bestimmung des Prinzips zu sein. Das Chancenprinzip formuliert eine normative Forderung, die in unserer Gesellschaft noch lange nicht erfüllt ist: Dass der Zugang zu Ämtern und Positionen unabhängig von der sozialen Herkunft der Person sein soll. Dies sagt allerdings für sich genommen noch nichts darüber aus, wie Chancengleichheit in Bezug auf Ämter und Positionen verwirklicht werden kann. Anders gefragt: Wie kann die Gleichheit der beruflichen Chancen trotz unterschiedlicher sozialer Herkunft hergestellt werden?

## 2. Bildung als primärer Gegenstand von Chancengleichheit

Im Folgenden werde ich für die These argumentieren, dass die Realisierung fairer Chancengleichheit in Bezug auf Ämter und Positionen voraussetzt, dass Chancengleichheit in der Bildung verwirklicht wird. Anders gesagt: Bildung muss der primäre Gegenstand von Chancengleichheit sein. Rawls selbst erwähnt durchaus, dass man den Grundsatz der fairen Chancengleichheit „auf das Bildungssystem anwendet" bzw. anwenden soll (GaF, 83, Anm. 10), doch auf der Grundlage seiner spärlichen und verstreuten Kommentare zum Bildungssystem entsteht der Eindruck, dass Bildung eben nur *ein* möglicher Gegenstand der Chancengleichheit ist, von dem nicht deutlich wird, welchen Stellenwert er einnimmt.

Die gerade herausgearbeitete Konzeption der fairen Chancengleichheit lautet, wenn man sie auf den Gegenstand der Bildung anwendet, folgendermaßen: Für ähnlich Begabte und Motivierte sollte die „Möglichkeit, sich das Wissen und Können seiner Kultur anzueignen" (TdG, 93) nicht von ihrer sozialen Schicht abhängen. Unter „Bildungschance" wird also eine Chance im Sinne der Möglichkeit verstanden, kulturelles Wissen und Können zu erwerben, die letztlich in einen Bildungsabschluss mündet. Ähnlich Begabte und Motivierte, so lautet die Forderung nach fairer Chancengleichheit in der Bildung, sollen unabhängig von ihrer sozialen Herkunft dieselben Chancen auf den Erwerb von Wissen und der entsprechenden Bildungszertifikate haben. Anders formuliert, besteht Chancengleichheit in der Bildung nach Rawls darin, dass jeder durch Bildung die gleiche Möglichkeit haben soll, seine jeweiligen Talente und Fähigkeiten zu entfalten und einen entsprechenden Bildungsabschluss zu erwerben. Da sich Personen hinsichtlich ihrer natürlichen Fähigkeiten unterscheiden, werden sich unter der Bedingung der Chancengleichheit auch ihre Bildungsabschlüsse unterscheiden.

Das Argument, warum Chancengleichheit in Bezug auf Ämter und Positionen voraussetzt, dass Chancengleichheit in der Bildung realisiert wird, erwähnt Rawls nicht explizit. Es lässt sich folgender Maßen rekonstruieren: Wenn Ämter und Positionen allen Bürgern unabhängig von ihrer sozialen Herkunft offenstehen sollen, dann müssen alle unabhängig von ihrer sozialen Herkunft die Chance haben, die für die Ämter und Positionen notwendigen Qualifikationen zu erwerben, d.h. ihr Zugang zu den relevanten Bildungsabschlüssen muss ebenfalls (bzw. sogar: zu allererst!) unabhängig von ihrer sozialen Herkunft offenstehen.

Wird das Chancenprinzip nur auf Ämter und Positionen, und nicht auf das Bildungssystem, angewandt, läuft es Gefahr, auf das Verdienstprinzip reduziert zu werden, das besagt, das Ämter und Positionen an diejenigen vergeben werden sollen, die für diese Ämter am besten

---

*auch* dadurch erreicht werden, „daß man sich mehr um die Begabten kümmert" (ebd.). Das Differenzprinzip bedeutet nicht in erster Linie, daß man natürliche Nachteile ausgleicht, sondern vielmehr „die Verteilung der natürlichen Gaben in gewisser Hinsicht als Gemeinschaftssache betrachtet und in jedem Falle die größeren sozialen und wirtschaftlichen Vorteile aufteilt, die durch die Komplementaritäten dieser Verteilung ermöglicht werden" (ebd.). Gosepath hingegen meint, dass Eliteförderung (unter den Bedingungen der Ressourcenknappheit) zugunsten der Förderung der weniger Begabten verboten werden müsste, wenn faire Chancengleichheit einen Vorrang vor ökonomischer Umverteilung hat (Gosepath 2004, 446). Doch Rawls ist gerade deshalb anderer Meinung, da er das Chancenprinzip nicht auf *natürliche*, sondern nur soziale Faktoren anwendet (vgl. dazu auch Meyer 2011, 167f.).

geeignet sind. Doch dies würde auf eine bloße Unparteilichkeitsforderung bei der Vergabe von Ämtern hinauslaufen und hätte nichts mit fairer Chancengleichheit zu tun, die den Einfluss sozialer Herkunft auf den Zugang zu Ämtern verringern möchte.

Obgleich Rawls den systematischen Stellenwert der gleichen Bildungschancen nicht herausstreicht, kann die These vom Primat der Bildung in Bezug auf Chancengleichheit an seine Äußerungen anknüpfen. In der folgenden Passage wird deutlich, dass für Rawls gleiche Bildungschancen zu den zwei Voraussetzungen gehören, die für die Realisierung fairer Chancengleichheit in Bezug auf Ämter und Positionen erfüllt sein müssen:

> Damit das Prinzip [der fairen Chancengleichheit, CB] seinen Zweck erfüllt, müssen bestimmte Anforderungen an die Grundstruktur gestellt werden [...]. Ein freies Marktsystem muß in einen Rahmen politischer und rechtlicher Institutionen eingebettet werden, die den langfristigen Trend ökonomischer Kräfte so regeln, daß übermäßige Konzentrationen von Eigentum und Vermögen verhindert werden, insbesondere solche Formen der Konzentration, die wahrscheinlich zu politischer Vorherrschaft führen. Außerdem muß die Gesellschaft unter anderem *gleiche Bildungschancen für alle* durchsetzen, einerlei, wie hoch das Einkommen der jeweiligen Familie ist [...] (GaF, 79f., H.v.m.).[6]

Neben der Forderung nach gleichen Bildungschancen nennt Rawls die Verhinderung „übermäßiger Konzentrationen von Eigentum und Vermögen“, die notwendig ist, um Chancengleichheit in Bezug auf Ämter zu realisieren. Auch wenn dies zweifellos ein relevanter Punkt ist, ist er im Vergleich mit der Forderung nach gleichen Bildungschancen sekundär, denn er ist rein *negativ*, also bloß auf die Verhinderung bestimmter Umstände ausgerichtet. Die Forderung nach gleichen Bildungschancen ist demgegenüber *positiv* in dem Sinn, dass sie nicht nur Konzentration von Macht verhindern möchte, sondern auch auf die Befähigung von denjenigen drängt, die in der Gesellschaft schlechter gestellt sind. In diesem Sinn traut Rawls Bildung auch zu, den „Abbau von Klassenschranken“ (TdG, 93f.) zu bewirken.[7]

## 3. Grenzen der Realisierung fairer Chancengleichheit

Die Frage, wie Chancengleichheit realisiert werden kann, stellt sich demnach primär in Bezug auf Chancengleichheit in der Bildung. Ich werde im Folgenden untersuchen, inwiefern Rawls' Theorie eine Antwort auf diese Frage geben kann. Dabei werde ich insbesondere in den Blick

---

[6] Ähnlich auch in *Politischer Liberalismus*: Die Unterschiede hinsichtlich der moralischen und intellektuellen Fähigkeiten werden „durch soziale Praktiken reguliert, die sich auf die Qualifikation für Positionen und den freien Wettbewerb vor einem Hintergrund fairer Chancengleichheit beziehen, was faire Bildungschancen und die Regulierung von Einkommens- und Vermögensunterschieden durch das Differenzprinzip einschließt“ (PL, 279).

[7] Es ist Rawls' Chancenprinzip vorgeworfen worden, es beinhalte eine interne Widersprüchlichkeit, da es sich sowohl auf den Arbeitsmarkt bezieht (der insbesondere für die Realisierung formaler Chancengleichheit zuständig ist), als auch auf das Bildungssystem gerichtet ist (das gleiche Startchancen und damit Unabhängigkeit von der sozialen Herkunft ermöglichen soll). Der Bezug auf unterschiedliche Geltungsbereiche lasse sich jedoch nicht in einem Prinzip vereinigen (vgl. Richards 1997, 267). Mir leuchtet es jedoch nicht ein, warum der Bezug auf unterschiedliche Geltungsbereiche ein Nachteil oder sogar ein interner Widerspruch sein sollte. Das Chancenprinzip muss eben von verschiedenen Institutionen umgesetzt werden: Institutionen des Bildungswesens sind für die bestmögliche Entwicklung der natürlichen Fähigkeiten zuständig, während Institutionen des Arbeitsmarktes kontrollieren müssen, ob rechtliche Gleichheit im Zugang zu Ämtern garantiert ist (so auch Wallimann 2008, 24). Ich würde dem Vorwurf nur insofern zustimmen, dass von Rawls nicht deutlich genug gesagt wird, dass sich Chancengleichheit auf beide Institutionen (des Arbeitsmarktes und des Bildungssystems) beziehen muss und dementsprechend die Institutionen in beiden Kontexten notwendig sind, um das Prinzip zu verwirklichen.

nehmen, wie für Rawls die Frage nach der Realisierung von Chancengleichheit zu seiner Unterscheidung zwischen idealer und nicht-idealer Theorie steht. Es ist Rawls' idealer Theorie in neuerer Zeit vorgeworfen worden, dass sie aufgrund ihrer Abstraktionen ein *verzerrtes* Bild der Wirklichkeit liefert (Mills 2005) und *keinen Leitfaden* für drängende Gerechtigkeitsprobleme unter realen, d.h. „nicht-idealen" Umständen zu bieten hat („guidance critique"[8]). Ich möchte zunächst zeigen, dass diese Vorwürfe in Bezug auf das Problem der Chancengleichheit berechtigt sind, aber anschließend Rawls' ideale Theorie in ergänzter Form verteidigen.

Nach Rawls ist der Grund, warum Chancengleichheit (sowohl in Bezug auf Ämter und Positionen als auch auf Bildung) nur schwer zu realisieren ist, der Einfluss der Familie:

> [D]er Grundsatz der fairen Chancen [lässt sich, CB] nur unvollkommen durchführen, mindestens solange es die Familie in irgendeiner Form gibt. Der Grad der Entwicklung und Betätigung natürlicher Fähigkeiten hängt von allen möglichen gesellschaftlichen Verhältnissen und klassengebundenen Einstellungen ab. Selbst die Bereitschaft zum Einsatz, zur Bemühung, die im üblichen Sinne verdienstvoll ist, hängt noch von günstigen Familienumständen und gesellschaftlichen Verhältnissen ab (TdG, 94).

Das ist eine ungünstige Prognose für das Chancenprinzip. Dieses fordert, dass jede Person *unabhängig von ihrer sozialen Herkunft* dieselben Chancen (auf Bildung) haben soll. Die Faktoren, die laut Rawls legitimen Einfluss auf die Verteilung der Chancen nehmen dürfen, sind natürliche Fähigkeiten und Leistungsbereitschaft bzw. Motivation. Das Chancenprinzip geht jedoch davon aus, dass diese beiden Faktoren Variablen sind, die ihrerseits von der sozialen Herkunft unabhängig sind. Rawls meint nun, dass der Einfluss der Familie diese Unabhängigkeit untergräbt: Selbst die beiden Faktoren der natürlichen Fähigkeiten und Leistungsbereitschaft tragen durch den Einfluss der Familie auch den Stempel der sozialen Herkunft. In Bezug auf natürliche Fähigkeiten hängt die „Entwicklung und Betätigung" dieser natürlichen Fähigkeiten vom Einfluss der Familie ab. Selbst unter der Annahme, dass es natürliche Anlagen für Fähigkeiten und Talente gibt, lässt sich im Allgemeinen kaum unterscheiden, welcher Anteil an einer Fähigkeit natürlich gegeben und welcher später entwickelt worden ist. Somit stellt die Tatsache, dass die Entwicklung einer Fähigkeit stark von sozialen Faktoren abhängig ist, einen gewichtigen Einwand gegen die Annahme dar, mit „natürlichen Fähigkeiten" eine Variable gefunden zu haben, die von sozialer Herkunft unabhängig ist, sodass sie legitimen Einfluss auf die Verteilung von Chancen haben darf. Genau dasselbe gilt für die Variable der Leistungsbereitschaft bzw. Motivation.

## 4. Das Problem der Chancengleichheit im Spiegel der Unterscheidung zwischen idealer und nicht-idealer Theorie

Ich möchte im Folgenden aufzeigen, inwiefern das Problem der Realisierbarkeit von Chancengleichheit ein Problem für Rawls' ideale Theorie darstellt. Dazu ist es nötig, die Unterscheidung zwischen idealer und nicht-idealer Theorie zu skizzieren: Nach Rawls spaltet sich seine Theorie der Gerechtigkeit in zwei Teile auf, einen idealen und einen nicht-idealen (vgl. z.B. TdG, 25, 277f., 337, 387). Der ideale Teil enthält die Gerechtigkeitsgrundsätze für eine „wohlgeordnete Gesellschaft unter günstigen Umständen", in der sich alle Bürger vollständig an die Prinzipien halten, d.h. „vollständige Konformität" mit den Gerechtigkeitsgrundsätzen herrscht (TdG, 277). Im Gegensatz dazu beschäftigt sich nicht-ideale Theorie mit Grundsätzen „unter weniger günstigen Umständen" (ebd.). Diese umfassen zwei verschiedene Arten von Umständen: Erstens greift nicht-ideale Theorie dann,

[8] Vgl. für eine Darstellung der *guidance critique* Valentini 2009, die allerdings nicht in eigener Stimme diesen Vorwurf an Rawls erhebt.

wenn sich Menschen nicht an die Gerechtigkeitsgrundsätze halten, d.h. vollständige Konformität nicht gegeben ist. So befasst sich nicht-ideale Theorie mit der „Theorie der Strafe [...], des gerechten Krieges und der Kriegsdienstverweigerung, des zivilen Ungehorsams und des militanten Widerstands“ (TdG, 387). Der andere Fall, in dem nicht-ideale Theorie greift, ist der unverschuldeter ungünstiger Umstände, wie z.B. angesichts „natürlicher Beschränkungen und geschichtlicher Zufälligkeiten“ (TdG, 277f.).

Das Problem der Realisierbarkeit von Chancengleichheit angesichts des familiären Einflusses ist offenkundig ein Problem für nicht-ideale Theorie im zweiten Sinn. Denn der soziale Einfluss der Familie, der faire Chancengleichheit unterläuft, hat nichts damit zu tun, dass sich Menschen nicht an Gerechtigkeitsgrundsätze halten, sondern gehört zu den „ungünstigen Bedingungen“, von denen ideale Theorie abstrahiert.[9]

Im Folgenden werde ich zeigen, inwiefern diese Abstraktion zu einem *verzerrten* Bild der Wirklichkeit beiträgt und inwiefern die *Leitfunktion* der idealen Theorie bei dem speziellen Problem der Chancengleichheit problematisch ist.

### *4.1. Vorwurf der Verzerrung*

Die Abstraktion vom Einfluss der Familie führt zu einem verzerrten Bild der Wirklichkeit, insofern die Rede von *natürlichen* Fähigkeiten und *natürlicher* Leistungsbereitschaft und Motivation fälschlicher Weise als unproblematisch dargestellt wird. Nur unter der Annahme, dass einige Menschen „von Natur aus“ in bestimmten Hinsichten talentierter sind als andere oder aufgrund ihres naturgegebenen Charakters eher bereit und motiviert sind, ihre Fähigkeiten zu entwickeln und einzusetzen, ist Rawls' Chancenprinzip überzeugend. Denn nur unter dieser Annahme rechtfertigt das Chancenprinzip Ungleichheiten in den beruflichen Aussichten, indem es diese an die individuellen Entscheidungen der Personen bindet, wobei die Entscheidungen aufgrund von nahezu unveränderlichen natürlichen Gegebenheiten getroffen werden. Doch wird die Abstraktion vom Einfluss der Familie fallen gelassen, ändert sich das Bild drastisch: Es wird deutlich, dass die Erziehung und das Vorbild der Eltern so starken Einfluss auf die Entwicklung von Fähigkeiten der Kinder und deren Einstellung zum Lernen und Arbeiten hat, dass der Anteil des „Natürlichen“ ins nahezu Unkenntliche zusammenschrumpft. Selbst wenn dies nicht „bis auf Null“ geschieht, führt doch die Berücksichtigung des familiären Einflusses dazu, mit viel größerer Vorsicht auf das Argument zurückzugreifen, eine Person sei eben begabter als eine andere und deshalb in einer besseren beruflichen Position.[10] Es fragt sich, ob das Chancenprinzip nicht sinnlos bzw. intrinsisch unerfüllbar wird, wenn seine Anwendbarkeit es von einer idealisierten Annahme (der Existenz rein natürlicher Fähigkeiten) abhängt.

Meines Erachtens trifft der Vorwurf der Verzerrung einen wahren Kern, ist aber nicht vernichtend für das Chancenprinzip, da die Annahme natürlicher Fähigkeiten nicht schlichtweg *falsch* zu sein scheint. Man kann davon ausgehen, dass es natürliche Unterschiede der Begabungen gibt. Chancengleichheit lässt sich daher weiterhin als Gleichheit der Aussichten *bei gleicher Begabung* verstehen. Die genannte Kritik sollte jedoch dazu führen, den Begriff der „natürlichen Fähigkeiten“ mit der gebotenen Vorsicht zu verwenden, was auf Grundlage der Idealtheorie nicht notwendig wäre.

---

[9] Allerdings gehört das Problem der Chancengleichheit nicht zu den Fragen, die Rawls als Gegenstände der nicht-idealen Theorie aufzählt. Dazu gehört „unter anderem die Theorie der Strafe und der ausgleichenden Gerechtigkeit, des gerechten Krieges und der Kriegsdienstverweigerung, des zivilen Ungehorsams und des militanten Widerstands“ (TdG, 387).

[10] Bourdieu und Passeron machen darauf aufmerksam, dass der Verweis auf „natürliche Begabung“ „hypothetisch bleibt, solange sich der unterschiedliche schulische Erfolg auf andere Ursachen zurückführen läßt“ (Bourdieu/Passeron 1971, 40). Der Verweis auf natürliche Fähigkeiten birgt die Gefahr, die Bildungsunterschiede aufgrund sozialer Privilegien in Naturgegebenheiten umzudeuten und Ungleichheit auf diese Weise zu legitimieren (vgl. Bourdieu/Passeron 1971, 45).

### *4.2 Vorwurf des fehlenden Leitbildes*

Wenden wir uns dem Einwand zu, dass ideale Theorie keine hilfreichen Leitlinien für nicht-ideale Umstände liefern kann. Das folgende Dilemma zeigt, dass dieser Vorwurf zumindest teilweise berechtigt ist. Der Einfluss der Familie auf natürliche Fähigkeiten und Motivation lässt zwei Möglichkeiten offen, um mit dem Problem der Chancengleichheit umzugehen: Entweder man strebt es an, den Einfluss der Familie auszugleichen, oder man versucht die Nachteile von Personen mit sozial schlechter gestellten Herkunftsfamilien zu kompensieren. Letzteres hieße dann, dass Chancengleichheit zwar nicht bzw. unvollkommen realisiert wird, aber die schlechter Gestellten von den Bessergestellten profitieren können. In die Richtung des zweiten Vorschlags weist Rawls, indem er sagt:

> Hier ist das Differenzprinzip relevant, denn wenn ihm Genüge getan wird, fällt es Menschen mit geringeren Chancen leichter, die von der Familie und sonstigen sozialen Bedingungen auferlegten Zwänge zu akzeptieren (GaF, 251).

Das Dilemma besteht darin, dass beide Möglichkeiten in Spannung zu jeweils einer Rawlsschen Annahme stehen: Der erste Vorschlag – den Einfluss der Familie zu verringern – läuft Gefahr, eine Einschränkung der Grundfreiheiten der Familie mit sich zu bringen, und dies widerspricht dem Vorrang des ersten Gerechtigkeitsgrundsatzes vor dem zweiten. Der zweite Vorschlag zielt darauf ab, Aspekten des Ausgleichs für Nachteile mehr Gewicht einzuräumen als der Herstellung von Chancengleichheit, und schwächt den Vorrang des Chancenprinzips vor dem Differenzprinzip.

Die Unsicherheit bezüglich des letzten Punktes gesteht Rawls zu, wenn er sagt:

> Manche glauben, daß der lexikalische Vorrang des Grundsatzes der fairen Chancengleichheit vor dem Differenzprinzip zu ausgeprägt ist; eine abgeschwächte Vorrangstellung oder eine schwächere Form des Chancenprinzips seien besser und stünden im Grunde eher in Einklang mit manchen Grundgedanken der Konzeption der Gerechtigkeit als Fairneß. Derzeit weiß ich nicht, welche Lösung hier die beste ist, und damit möchte ich schlicht meine Unsicherheit zu Protokoll geben. Wie das Chancenprinzip bestimmt und gewichtet werden sollte, ist eine überaus schwierige Frage, und es kann durchaus sein, daß eine der genannten Alternativen besser ist (GaF, 252, Anm. 44).

Rawls räumt hier die Möglichkeit ein, dass das Differenzprinzip doch stärker gewichtet werden könnte, als es der Vorrang des Chancenprinzips unter idealen Umständen erlauben würde. Um zu verstehen, was bei der Frage der Rangordnung der Prinzipien auf dem Spiel steht, rekapitulieren wir kurz Rawls' Gründe für den lexikalischen Vorrang[11] des Chancenprinzips vor dem Differenzprinzip.

Eine Antwort im Rawlsschen Sinne stützt sich meines Erachtens *erstens* auf die Idee der moralischen Gleichheit der Personen, *zweitens* auf die politische Idee einer Gesellschaft, in der die als gleich aufgefassten Bürger auch in gleicher Weise sozial, politisch und ökonomisch teilhaben können, und *drittens* auf die Relevanz von Ämtern und Positionen für die Selbstachtung.

Der erste Punkt, die moralische Gleichheit der Personen, ist eng mit der Idee verknüpft, dass es nicht von Geburt an feststehen soll, ob ein Lebenslauf einer Person besser oder schlechter verläuft als der einer anderen. Wenn Personen moralisch *gleich* sind, dann sollten ihre Lebensaussichten sich nicht aufgrund von Faktoren (wie der sozialen Herkunft) unterscheiden, für die sie nichts können.

---

[11] Dass ein Gerechtigkeitsprinzip lexikalischen Vorrang vor einem anderen hat, bedeutet, dass es nicht zugunsten einer Umsetzung des nachgeordneten Gerechtigkeitsprinzips verletzt werden darf. Ein Prinzip kommt erst dann zum Tragen, „wenn die ihm vorgeordneten entweder voll erfüllt oder nicht anwendbar sind" (TdG, 62).

Hand in Hand mit der Idee der Gleichheit der Personen geht, zweitens, die Idee, dass auf dieser Grundlage eine Gesellschaft realisiert werden soll, in der alle Personen gleichermaßen in sozialer, politischer und ökonomischer Sicht *teilhaben* können. Eine solche Gesellschaft ist nach Rawls einer Gesellschaft, in der die schlecht Gestellten Versorgung erhalten, vorzuziehen. Zwar würde Rawls den Bedürftigen, „die durch Zufälle oder Pech Verluste erleiden", durchaus Beistand versichern (GaF, 217), aber diese Versorgung sollte in einer Gesellschaft, die sich als „faires System der Kooperation zwischen Bürgern" versteht, „die als freie und gleiche Personen gesehen werden" (GaF, 218), der Ausnahmefall bleiben.

Der dritte Aspekt, der den Vorrang der Chancengleichheit motiviert, ist die Rolle von Bildung (und beruflicher Tätigkeit) für die Selbstachtung, die Rawls als das „vielleicht wichtigste Grundgut" bezeichnet (TdG, 479). Bildung und die Ausübung einer beruflichen Position bedeutet nach Rawls „Selbstverwirklichung in Form der Erfüllung gesellschaftlicher Pflichten mit Können und Hingabe, einer der Hauptformen des menschlichen Wohles" (TdG, 105). Rawls weist damit auf den Punkt hin, dass Ämter und Positionen nicht nur mit gewissen „äußeren Vorteilen" (TdG, 105) einhergehen – dabei denkt er wohl an ein entsprechendes finanzielles Einkommen – sondern auch mit der Möglichkeit der Selbstverwirklichung. Es lässt sich argumentieren (vgl. Wallimann 2008, 16f.), dass die Ausübung von Ämtern und Positionen das Grundgut der Selbstachtung fördert: Nach Rawls' Konzeption der Person besitzen Personen eine Konzeption des Guten und wollen ihr Leben in Übereinstimmung mit den Grundsätzen praktischer Rationalität führen. Dabei streben sie danach, ihre Fähigkeiten einzusetzen und finden mehr Befriedigung, je komplexer und besser entwickelt diese sind (vgl. TdG, 480). Dementsprechend ist der Zugang zu Ämtern und Positionen notwendig, um die persönliche Konzeption des Guten zu verwirklichen und damit eine notwendige Bedingung der Selbstachtung.

Darüber hinaus lässt sich auf die gesellschaftliche Konstitution der Selbstachtung verweisen. Eine Person, die ein Amt kompetent ausübt, erfährt gesellschaftliche Anerkennung, die ihre Selbstachtung stärkt. In diese Richtung weisen Rawls' Bemerkungen, „daß man nur dann vom Wert seiner Bemühungen überzeugt sein kann, wenn sie von den Mitmenschen geschätzt werden" (TdG, 480) und dass „Gruppenbindungen" die Selbstachtung stärken, „da sie ein Versagen weniger wahrscheinlich machen, und, wenn es doch vorkommt, dem Selbstzweifel entgegenwirken" (TdG, 481).

Alle drei Punkte sind nur dadurch zu verwirklichen, dass man dem Chancenprinzip Vorrang vor dem Differenzprinzip einräumt. Der Vorrang des Chancenprinzips basiert demnach auf wichtigen Grundideen der Rawlsschen Theorie und kann nicht einfach aufgegeben werden.

Das Problem lautet zusammengefasst: Rawls' ideale Theorie beansprucht, mit dem ersten und zweiten Gerechtigkeitsgrundsatz und den Vorrangregeln die Prinzipien für eine „vollkommen gerechte[...] Grundstruktur" (TdG, 277) zu liefern. Insbesondere beinhaltet das Chancenprinzip die Forderung nach Chancengleichheit unabhängig von der sozialen Herkunft, wobei dieses Prinzip dem der Garantie der Grundfreiheiten nach- und dem Differenzprinzip vorgeordnet sein soll. Doch die Gerechtigkeitsgrundsätze einschließlich der Vorrangregeln können nur vollständig umgesetzt werden, wenn vom Einfluss der Familie weitgehend abstrahiert wird. Ideale Theorie sagt demnach zwar, *dass* Chancengleichheit eine Forderung der Gerechtigkeit ist, aber *auf welche Weise* sie angesichts der realen Bedingungen umgesetzt werden kann, bleibt offen.

## 5. Realisierung der Chancengleichheit zwischen idealer und nicht-idealer Theorie

Die gerade geäußerte Kritik macht deutlich, dass Rawls' ideale Theorie der Ergänzung durch Regeln einer nicht-idealen Theorie bedarf. Die Notwendigkeit der Ergänzung resultiert aus

dem oben beschriebenen Dilemma, weder die Grundfreiheiten der Familie einschränken, noch dem Differenzprinzip den Vorrang einräumen zu wollen. Die einzige Möglichkeit, Chancengleichheit vollumfänglich zu verwirklichen, bestünde darin, die Familie „abzuschaffen", was nach Rawls das Ideal der Chancengleichheit zwar verlangt, wofür aber „im Gesamtzusammenhang der Gerechtigkeitstheorie" (TdG, 555) wenig spricht, da diese Maßnahme eindeutig die Grundfreiheiten der Familie untergräbt. Rawls selbst expliziert nicht, wie weit die Bemühungen um Chancengleichheit gehen sollen, obgleich er diese Frage stellt, sondern verweist wiederum nur darauf, dass durch das ausgleichende Differenzprinzip Ungleichheiten in den Chancen ‚leichter hinzunehmen' seien (TdG, 556).

Doch die Unmöglichkeit, Chancengleichheit *vollständig* zu realisieren, bedeutet noch nicht, dass das Differenzprinzip den Vorrang erhalten soll. Nach Rawls bedeutet „lexikalischer Vorrang", dass das nachgeordnete Prinzip erst dann zum Tragen kommen darf, „wenn die ihm vorgeordneten entweder voll erfüllt oder nicht anwendbar sind" (TdG, 62). Die entscheidende Frage bezüglich des Verhältnisses zwischen Chancen- und Differenzprinzip lautet also, wann das Chancenprinzip als „nicht anwendbar" erklärt werden kann.

Zur Beantwortung dieser Frage muss geklärt werden, wie weit Chancengleichheit verwirklicht werden kann, ohne die Grundfreiheiten der Familie anzugreifen. Dass es eine Einschränkung der Grundfreiheiten von Eltern ist, wenn man ihnen ihre Kinder im Alter von drei Jahren wegnimmt, würden wohl die meisten unterschreiben. Ob es jedoch auch die Grundfreiheiten beschneidet, wenn der Besuch einer Vorschule für Fünfjährige für verpflichtend erklärt würde, ist sicherlich diskutierenswert. Was zu den Grundfreiheiten gehört und was nicht, entscheidet also wesentlich darüber, wie stark der Einfluss der Familie ausgeglichen und damit die Realisierung der Chancengleichheit befördert werden kann. Welche konkreten Maßnahmen Chancengleichheit verwirklichen können, hängt von den jeweiligen gesellschaftlichen Bedingungen ab. Diese Maßnahmen müssen in empirischen Studien auf ihre Wirksamkeit hin überprüft werden.

Es fragt sich, was die Quellen der normativen Regeln zur Bestimmung der familiären Grundfreiheiten sind. Muss man, um diese Fragen zu beantworten, Rawls' ideale Theorie lediglich „anwenden" ohne sie durch andere normative Regeln zu ergänzen? Das wäre beispielsweise der Fall, wenn es darum ginge, ob sich das Gehalt eines Managers gegenüber dem eines Müllmanns noch weiter erhöhen dürfte. Dann müsste man prüfen, ob gemäß dem Differenzprinzip die Gehaltserhöhung des Managers auch dem Müllmann (und anderen schlecht gestellten Gesellschaftsmitgliedern) zugute kommt, z.B. durch progressive Besteuerung. Dies wäre eine „Anwendung" der idealen Theorie, da keine weiteren normativen Regeln notwendig wären, um die Frage zu entscheiden. Das ist jedoch im Fall der Realisierung von Chancengleichheit nicht ohne Weiteres möglich, denn um zu prüfen, ob eine konkrete Maßnahme die Grundfreiheiten der Familie verletzt, muss der Begriff der familiären Grundfreiheiten erst einmal inhaltlich bestimmt werden. Die Aufgabe ist demnach eine *Interpretation* des Begriffs der familiären Grundfreiheiten und eine Bestimmung der notwendigen und wirksamen Maßnahmen, die den jeweiligen konkreten Missständen hinsichtlich der Realisierung von Chancengleichheit angemessen sind.

Diese Aufgabe scheint nicht in den Zuständigkeitsbereich der idealen Theorie zu fallen, sondern vielmehr eine Aufgabe der öffentlichen Auseinandersetzung zu sein. Gerade in einer pluralistischen Gesellschaft können die Meinungen darüber auseinandergehen, welche die schützenswerten Grundfreiheiten der Familie sind. Wenn hinsichtlich dieser Frage ein Konflikt besteht, müssen die verschiedenen Parteien in einen Prozess der öffentlichen Rechtfertigung einsteigen, d.h. die

> anderen durch öffentlichen Vernunftgebrauch überzeugen, d.h. durch Formen des Denkens und Schließens, die politischen Grundfragen angemessen sind, sowie durch

> Berufung auf Überzeugungen, Gründe und politische Werte, deren Anerkennung seitens der anderen ebenfalls vernünftig ist (GaF, 56).

Die durch öffentliche Diskussion erlangten Regeln zur inhaltlichen Bestimmung der familiären Grundfreiheiten und der effektiven Maßnahmen zur Umsetzung von Chancengleichheit gehören insofern nicht zur idealen Theorie, als sie nicht direkt aus dieser abgeleitet sind, sondern unter konkreten gesellschaftlichen Bedingungen im Rahmen einer öffentlichen kritischen Auseinandersetzung ausgehandelt werden müssen.

Der Nachteil eines solchen Vorgehens ist, dass es womöglich in vielen Situationen keine eindeutigen Ergebnisse gibt und keine vollständige Zustimmung zu einer Lösung erwartet werden kann. Doch gerade dieser Umstand lässt sich auch als Vorteil gegenüber einer idealtheoretischen, eindeutigen Lösung sehen: Gerade angesichts pluralistischer Auffassungen über die Rolle der Familie würde eine „ideale" Lösung höchstwahrscheinlich an den tatsächlichen Vorstellungen und Bedürfnissen der Betroffenen vorbeigehen und als aufoktroyiert empfunden werden.

Die Antwort auf meine Ausgangsfrage, welche Aussagen die Rawlssche Theorie in Bezug auf die *Realisierung* von Chancengleichheit machen kann, lautet zusammengefasst: Offenbar kann Rawls' ideale Theorie im Sinne seiner Gerechtigkeitsgrundsätze keine Leitfunktion übernehmen, wenn es um die konkrete Interpretation der Grundfreiheiten und die Bestimmung der jeweiligen Maßnahmen zur Herstellung von Chancengleichheit geht. Doch deshalb wird die Leitfunktion nicht völlig verneint, denn das Prinzip der Chancengleichheit gibt weiterhin an, wonach bei der Realisierung von Chancengleichheit überhaupt gestrebt werden soll: Nach gleichen Chancen auf Ämter und Positionen – und deshalb insbesondere: auf Bildung – bei gleicher Begabung, unabhängig von gesellschaftlichen Zufälligkeiten. Auch die Vorrangregeln behalten weiterhin ihre normative Kraft, denn auch unter nicht-idealen Bedingungen muss versucht werden, Chancengleichheit zu realisieren, ohne die Grundfreiheiten der Familie einzuschränken.[12]

**Claudia Blöser**

Institut für Philosophie
Goethe-Universität Frankfurt
Bloeser@em.uni-frankfurt.de

## Literatur

Boettcher, J. 2009: „Race, Ideology, and Ideal Theory", *Metaphilosophy*, 40, 2, 237–259.

Bourdieu, P. und J.–C. Passeron 1971: *Die Illusion der Chancengleichheit*. Stuttgart: Klett.

Gosepath, S. 2004: *Gleiche Gerechtigkeit: Grundlagen eines liberalen Egalitarismus*. Frankfurt: Suhrkamp.

Meyer, K. 2011: *Bildung*. Berlin: de Gruyter.

Mills, C. W. 2005: „Ideal Theory as Ideology", *Hypathia*, vol. 20, no.3, 165–184.

Rawls, J. 1979: *Eine Theorie der Gerechtigkeit*. Frankfurt: Suhrkamp. (kurz: TdG)

Rawls, J. 1998: *Politischer Liberalismus*. Frankfurt: Suhrkamp (kurz: PL)

Rawls, J. 2003: *Gerechtigkeit als Fairneß. Ein Neuentwurf*. Frankfurt: Suhrkamp. (kurz: GaF)

[12] Für wertvolle Anregungen und Hinweise zu früheren Fassungen dieses Textes danke ich Corinna Mieth und Christian Neuhäuser.

Schaub, J. 2010: „Ideale und/oder nicht-ideale Theorie – oder weder noch? Ein Literaturbericht zum neuesten Methodenstreit in der politischen Philosophie“, *Zeitschrift für Philosophische Forschung*, Band 64, 3, 393–409.

Sher, G. 1988: „Qualifications, Fairness and Desert“, in N. Bowie (Hrg.): *Equality of opportunity*. London: Westview, 113–127.

Richards, J. 1997: „Equality of Opportunity“, *Ratio X*, 253–279.

Rössler, B. (Hrg.) 1993: *Quotierung und Gerechtigkeit. Eine moralphilosophische Kontroverse*, Frankfurt/New York: Campus.

Simmons, A. J. 2010: „Ideal and Nonideal Theory“, *Philosophy and Public Affairs* 38, 5–36.

Valentini, L. 2009: „On the Apparent Paradox of Ideal Theory“, *Journal of Political Philosophy*, 17, 3, 332–355.

Wallimann, I. 2008: „Warum faire Chancengleichheit“ in: C. F. Gethmann: Lebenswelt und Wissenschaft. XXI. Deutscher Kongress für Philosophie, Sektions-CD, 2008.

# Liberalismus, Handlungsfreiheit und Autonomie

Christine Bratu

Häufig wird behauptet, normativer Kerngedanke des Liberalismus sei ein individuelles Recht auf Freiheit, d.h. es ginge dem Liberalismus um den Schutz der individuellen Handlungsfreiheit. Ich möchte zeigen, dass diese Annahme für den Kontext der politischen Philosophie nicht zutrifft. Denn der Liberalismus setzt sich nur insofern für die Handlungsfreiheit der Bürgerinnen und Bürger gegenüber dem Staat ein, als diese notwendig ist für deren Autonomie. Das Legitimitätskriterium, das der Liberalismus an staatliches Handeln anlegt, sichert den Bürgerinnen und Bürgern diejenigen Handlungsalternativen, die sie brauchen, um autonom handeln zu können. Um für diese These zu argumentieren, werde ich in einem ersten Schritt darlegen, was man unter individueller Handlungsfreiheit in deskriptiver Hinsicht verstehen sollte (1 und 2). Vor dem Hintergrund dieses Freiheitsbegriffes müssen staatliche Handlungen immer als Einschränkungen der individuellen Handlungsfreiheit der Bürgerinnen und Bürger verstanden werden, wie ich in einem zweiten Schritt (3) darstellen möchte. Schließlich werde ich zeigen, dass der Liberalismus nicht jede Einschränkung der Handlungsfreiheit der Bürgerinnen und Bürger durch den Staat verbietet, sondern ihnen nur den Handlungsspielraum für autonomes Handeln lässt (4). Doch zudem wird deutlich werden, dass dies nicht als Mangel des Liberalismus verstanden werden muss. Denn tatsächlich ist uns vernünftigerweise nicht an allen Instanzen unserer individuellen Handlungsfreiheit gelegen.

## 1. Kriterien für einen angemessenen Freiheitsbegriff

Schon vielfach ist der Versuch unternommen worden, eine befriedigende Auffassung von Handlungsfreiheit zu etablieren[1], also davon, wann man sagen sollte, dass es einer Person A möglich ist, eine Handlung x zu vollziehen. Doch dieser Versuch ist bisher nicht geglückt, da gegen jede Explikation zahlreiche Einwände vorgebracht wurden. Dies spricht in den Augen vieler Autorinnen und Autoren dafür, die Suche danach, was wir mit "Handlungsfreiheit" *eigentlich* meinen, insgesamt als sinnlos anzusehen. So stellt sich bspw. für Leif Wenar die Frage,

> [W]hat could be the warrant for reaching into polysemic concept such as freedom and privileging – simply as a conceptual matter – one of its many conceptions? How could one sense of "freedom" be "freedom as such"? [...] The danger of this semantic privileging is that it creates a kind of tunnel vision, blocking out the many complexities of ordinary usage that analysis has revealed. (Wenar 2008: 46)[2]

Doch obwohl Wenar Recht hat mit der Feststellung, dass es schwierig ist, ein angemessenes Verständnis von Handlungsfreiheit als solcher zu gewinnen, bleibt er hierfür eine Erklärung schuldig. Meiner Ansicht nach ist der Grund folgender: Wenn wir versuchen, uns darüber klar zu werden, was wir mit "Handlungsfreiheit" meinen, legen wir verschiedene Kriterien an. Doch diese Kriterien sind nicht ohne weiteres miteinander vereinbar, so dass eine Auffassung

[1] Bereits Isaiah Berlins klassischer Aufsatz kann so verstanden werden, vgl. Berlin 1996 [1969].

[2] Zu einer ähnlichen Diagnose kommen auch Ian Carter, Matthew Kramer und Hiller Steiner: "Like other ideas that figure in our political thinking [...] the idea of freedom is complex, not simple. For freedom is a concept that comprises many aspects or dimensions, each of which is at least somewhat open to rival interpretations. [...] Any particular conception of freedom consists in some permutation of these rival dimensional interpretation."[2] (Carter/ Kramer/ Steiner 2007: xvii f.)

von Freiheit, die gemäß einem ersten Kriterium angemessen ist, gemäß einem zweiten als inadäquat angesehen werden muss. Tatsächlich scheint es keinen Freiheitsbegriff zu geben, der allen relevanten Maßstäben vollständig gerecht wird. Aber ich möchte zeigen, dass eine Auffassung im Vergleich zu anderen dennoch überzeugt.[3]

Dabei sollten wir an eine sinnvolle Auffassung von Freiheit drei Anforderungen stellen. Zum einen sollte sie *mit unseren begrifflichen Intuitionen weitestgehend übereinstimmen* und Fällen gerecht werden, die wir als paradigmatisch empfinden (im Folgenden Kriterium I). Eine Auffassung von Handlungsfreiheit, nach der Individuen als frei anzusehen sind, denen wir dies intuitiv keinesfalls zusprechen würden (etwa Menschen, die im Gefängnis sind oder in Sklaverei leben), ist verfehlt. Zwar muss eine adäquate Auffassung von Freiheit nicht in allen Fällen mit unseren Intuitionen übereinstimmen, sondern kann diese vereinzelt modifizieren, wenn dadurch unser Gebrauch des Begriffs kohärent wird. Aber letztlich sollte die philosophische Auffassung von Freiheit mit unseren lebensweltlichen Intuitionen in einem Überlegungsgleichgewicht stehen.

Neben der Intuitivität eines zu etablierenden Freiheitsbegriffs sollte auch dessen wissenschaftliche Fruchtbarkeit beachtet werden. So ist eine Auffassung von Freiheit, die mehr Forschungsperspektiven eröffnet und den Begriff für weitere Debatten anschlussfähig macht, angemessener als eine, die das Nachdenken über Freiheit von anderen wissenschaftlichen Diskursen isoliert. Insbesondere sollte die Auffassung *kreativ* sein (im Folgenden Kriterium K). D.h. eine gute Explikation von Freiheit achtet darauf, dass mit "Freiheit" nicht etwas bezeichnet wird, für das wir bereits über einen anderen Begriff verfügen (vgl. Wendt 2009: 25).

Schließlich ist der Tatsache Rechnung zu tragen, dass „Freiheit" für uns intuitiv *positiv besetzt* ist (im Folgenden Kriterium P). Dies belegen nicht nur politische Pamphlete spätestens seit dem Amerikanischen Unabhängigkeitskrieg und der Französischen Revolution, sondern zahllose Bücher und Filme, in denen Freiheit als hohes Gut gefeiert wird. Eine sinnvolle Auffassung des Konzeptes muss dieses so explizieren, dass die positive Konnotation erhalten bleibt. Insofern trifft folgender Hinweis David Millers zu:

> [W]hat precisely is this thing, freedom or liberty, for which so many have fought? What do we mean when we say that a person or a society is free or enjoys liberty? Wrestling with this problem may involve, at times, arguing in a somewhat abstract way. But a correct answer must be such that a student in Tiananmen Square could have recognized it as a description of what he was fighting for. (Miller 1991: 2)

Natürlich sind diese Kriterien nicht unabhängig von einander: Eine Auffassung, die P nicht erfüllt, erfüllt auch I nicht – denn intuitiv halten wir Freiheit für wertvoll. Ebenso wenig erfüllt ein Konzept, das K nicht gerecht wird, P. Denn wenn sich Freiheit auf ein anderes Konzept reduzieren lässt, warum wird dann für Freiheit und nicht für dieses andere Konzept gekämpft? Dennoch möchte ich die Kriterien auseinanderhalten um zeigen zu können, in welcher Hinsicht ein vorgeschlagenes Konzept von Freiheit überzeugt oder eben nicht.

Dabei lassen sich die vorgeschlagenen Konzepte grundsätzlich in drei Kategorien einteilen: In Anlehnung an Charles Taylor kann man Handlungsfreiheit entweder als "*opportunity*-" oder aber als "*exercise-concept*" verstehen (vgl. Taylor 1991 [1979]: 143f.). Die Klasse der Freiheitsauffassungen, nach denen Freiheit in der Möglichkeit ("opportunity") etwas zu tun besteht, kann man wiederum aufteilen in *negative* und *positive* Auffassungen von Freiheit.

[3] Die Auffassungen von Freiheit, die ich im Folgenden diskutiere, sind nicht alle denk-möglichen. Aber da es diejenigen sind, die in der Literatur am häufigsten vertreten werden, sollte ein Vergleich dieser Positionen doch aufschlussreich sein.

## 2. Eine angemessene Auffassung von Handlungsfreiheit

Alle negativen Auffassungen von Freiheit beinhalten folgenden Kerngedanken: *Eine Person A ist genau dann frei zu einer Handlung x, wenn A nicht an der Handlung x gehindert, d.h. wenn die Handlung x für A nicht unmöglich gemacht wird.* Dieser Kerngedanke ist für besagte Auffassungen insofern namensgebend, als Freiheit "nach der eigentlichen Bedeutung des Wortes die Abwesenheit äußerer Hindernisse" (Hobbes 2000 [1651]: 99) ist. D.h. A ist genau dann frei zu x, wenn etwas nicht der Fall ist, nämlich wenn ihr x nicht unmöglich gemacht wird. Für die umfassendste Version von negativer Freiheit $NF_1$ stellt dies bereits eine vollständige Explikation dar. Doch ausgehend von obigen Kriterien sollte man sich diesem Urteil nicht anschließen:

Zum einen wird $NF_1$ bestimmten grundlegenden begrifflichen Intuitionen und damit Kriterium I nicht gerecht. Denn das Urteil "A ist frei zu x" impliziert, dass A auch die Möglichkeit hat, x zu tun. Damit dies der Fall ist, muss eine sinnvolle Explikation von Freiheit aber fordern, dass die Handlung x (bzgl. derer behauptet werden soll, dass A zu ihr frei ist) eine ist, die A tatsächlich offensteht. Anderenfalls könnte man behaupten, dass Anna dazu frei sei, bis zum Mond zu springen, wenn sie nicht von anderen Personen daran gehindert wird. Weil $NF_1$ zulässt, dass Individuen als frei zu Handlungen bezeichnet werden können, die ihnen in einem gehaltvollen Sinne gar nicht möglich sind, entspricht $NF_1$ auch nicht der Forderung P und wird den positiven Konnotationen, die wir mit Freiheit verbinden, nicht gerecht. Denn warum sollte man bspw. in einem politischen Kampf die Möglichkeit erstreiten wollen, an bestimmten Handlungen nicht mehr gehindert zu werden, wenn man diese Handlungen ohnehin nicht durchführen kann?

Insofern stellt es eine Verbesserung dar, wenn $NF_2$ zu obigem Kerngedanken noch die Bedingung hinzufügt, dass die Handlung x eine sein muss, welche A in einem gehaltvollen Sinne offensteht. So gilt nach $NF_2$: *Eine Person A ist genau dann frei zu einer Handlung x, wenn (1) die Handlung x A nicht durch ein Hindernis H unmöglich gemacht wird und (2) die Handlung x für A möglich ist, wenn H unterbleibt.*

Doch auch diese Auffassung kann nicht überzeugen. Einerseits werden wiederum wichtige begriffliche Intuitionen außer Acht gelassen. Denn intuitiv scheinen wir A nur dann als frei bezeichnen zu wollen, wenn sie unbehelligt ist durch Hindernisse, die von anderen Menschen verursacht wurden. Wir würden bspw. kaum sagen wollen, dass Anna frei dazu ist, geradeaus zu gehen, weil sie keine Bäume daran hindern. Die Rede von Unfreiheit und Freiheit erscheint nur dort begrifflich angemessen, wo signalisiert werden soll, dass sich Menschen wechselseitig behindern oder eben nicht. Ebenfalls gegen $NF_2$ spricht, dass diese Auffassung nicht begrifflich kreativ ist und damit Bedingung K verletzt. Fordert man nämlich nicht explizit, dass die Hindernisse, die zu Freiheitseinschränkungen führen, von anderen Menschen verursacht sein müssen, wird das Prädikat "frei zu x" synonym mit "x können". Nach dieser Lesart ist A immer dann frei zu x, wenn sie kein Hindernis H gleich welcher Art von x abhält, d.h. wenn sie x tun kann. Doch obwohl ein angemessener Freiheitsbegriff der Tatsache Rechnung trägt, dass "frei zu x" "x können" impliziert, sollte die andere Implikationsrichtung und damit die Äquivalenz der beiden Begriffe vermieden werden. Anderenfalls verliert der Begriff der Freiheit gegenüber dem des Könnens seine Pointe (vgl. Wendt 2009: 25).

Diese Einsicht versucht $NF_3$ in folgender Auffassung von Freiheit umzusetzen: *Eine Person A ist genau dann frei zu einer Handlung x, wenn (1) die Handlung x A nicht durch andere Personen unmöglich gemacht wird und (2) die Handlung x für A möglich ist, wenn eine Hinderung durch andere Personen unterbleibt.* Diese Auffassung ist zum einen intuitiv und entspricht damit Kriterium I. Denn nach ihr kann man nicht davon sprechen, dass Anna frei dazu ist, bis zum Mond zu springen, oder dass Anna durch Bäume unfrei gemacht wird. Zudem charakterisiert $NF_3$ paradigmatische Fälle korrekt. Denn nach dieser Auffassung sind

Menschen im Gefängnis oder in Sklaverei unfrei. Zum anderen ist $NF_3$ begrifflich kreativ, da so verstanden „Freiheit“ und „Können“ keine Synonyme mehr sind. Doch hinsichtlich des Kriteriums P kann $NF_3$ scheinbar nicht überzeugen. Denn gemäß dieser Auffassung erscheinen plötzlich sehr viele Handlungen als Freiheitseinschränkungen. Wenn Anna bspw. den Raum betritt und sich auf den einzigen Stuhl setzt, macht das dem ebenfalls anwesenden Ben die Handlung „sich auf den einzigen Stuhl im Raum setzen“ unmöglich. Tatsächlich muss jede Handlung, die eine Person im Beisein anderer vollzieht, als Freiheitseinschränkung gewertet werden: Nimmt Anna einen bestimmten Punkt im Raum ein, so macht sie es dadurch anderen Anwesenden unmöglich, sich selbst an diesem Platz aufzuhalten, und diese somit zu dieser Handlung unfrei. Aber wer würde schon für die Möglichkeit streiten wollen, sich auf einen Stuhl setzen oder einen bestimmten Punkt im Raum einnehmen zu können? Versteht man Freiheit in Sinne von $NF_3$ droht man sie zu banalisieren, so dass sie kaum noch als positiv besetzt anzusehen ist.

Man könnte versuchen, $NF_3$ folgendermaßen zu verteidigen: Vielleicht bringt $NF_3$ nicht in jedem Fall angemessen zum Ausdruck, was wir mit “Freiheit” meinen, aber doch zumindest in politischen Fragen. Wovor sollen uns denn die klassischen Freiheitsrechte wie Religions-, Meinungs- und Versammlungsfreiheit schützen, wenn nicht davor, dass uns der Staat Handlungen erschwert oder unmöglich macht, die uns grundsätzlich offenstehen? Doch folgendes Beispiel von Taylor zeigt, dass $NF_3$ den positiven Konnotationen, die Freiheit haben sollte, auch im politischen Kontext nicht gerecht wird:

> Consider the following defence of Albania as a free country. We recognize that religion has been abolished in Albania, whereas it hasn’t been in Britain. But on the other hand there are probably far fewer traffic lights per head in Tirana than in London. [...] Suppose an apologist for Albanian Socialism were nevertheless to claim that this country was freer than Britain, because the number of acts restricted was far smaller. After all, only a minority of Londoners practice some religion in public spaces, but all have to negotiate their way through traffic. [...] In sheer quantitative terms, the number of acts restricted by traffic lights must be greater than that restricted by a ban on public religious practice. So if Britain is considered a free society, why not Albania? (Taylor 1991 [1979]: 150f.)

Taylor zeigt, dass man, wenn man $NF_3$ akzeptiert, öfter Grund dafür gehabt hätte, in Großbritannien für Freiheit zu demonstrieren als in Albanien. Denn unter Magaret Thatcher wurden den Bürgerinnen und Bürgern häufiger Handlungen unmöglich gemacht als unter Enver Hoxha, selbst wenn die Handlungen, die durch Thatcher verhindert wurden, weniger relevant waren als die, die Hoxha verboten hat. Da aber selbst die Iron Lady die Repressivität von Hoxhas Regime kaum überboten hat, scheint $NF_3$ auch im politischen Kontext nicht das widerzuspiegeln, was wir mit Freiheit meinen.

Taylors Hinweis macht deutlich, dass uns nicht an allen Handlungen begründetermaßen etwas liegt, die wir ausführen können und durch deren Ver- oder Behinderung der Staat unsere Handlungsfreiheit (gemäß den Auffassungen $NF_1$ – $NF_3$) beschneiden würde. Manche Einschnitte in unsere Möglichkeiten nehmen wir billigend in Kauf, weil sich uns dadurch andere, wichtigere Möglichkeiten eröffnen. Sollte man sich bspw. wirklich empfindlich in seiner Freiheit eingeschränkt fühlen, wenn einem verboten wird, mit dem Auto auf der linken Straßenseite zu fahren, obwohl man dadurch in den Genuss eines funktionierenden Verkehrsablaufs kommt? Diese Überlegung macht sich die positive Auffassung von Freiheit PF zunutze, indem sie festlegt: “Positive liberty is the possibility of acting [...] in such a way as to take control of one’s life and realize one’s fundamental purposes.” (Carter 2012) D.h. PF ergänzt die negative Auffassung von Freiheit um die Bedingung, dass die Handlung x, zu der A frei sein soll, eine sein muss, die für A wichtig ist (“to take control of one’s life and realize one’s fundamental purposes”). Nach PF gilt also: *Eine Person A ist genau dann frei zu einer Handlung x, wenn (1) die Handlung x A nicht durch andere Personen unmöglich gemacht*

*wird, (2) die Handlung x für A möglich ist, wenn die Hinderung durch andere Personen unterbleibt, und (3) die Handlung x eine Handlung ist, die A wichtig ist.*[4]

Durch Bedingung (3) ist gewissermaßen trivialerweise sichergestellt, dass PF Kriterium P erfüllt. Denn frei kann eine Person nach dieser Auffassung ausschließlich zu Handlungen sein, an denen ihr gelegen ist. Dadurch ist gewährleistet, dass eine Person, die in ihrer Freiheit beschnitten wird, diesen Verlust betrauern wird. Zudem erfüllt PF das Kriterium K, da "Freiheit" aufgrund von Bedingung (2) nicht auf "Können" reduziert wird. Allerdings sind diese Vorteile nur um einen gewissen Preis zu haben: Angenommen, Anna wird von Ben in einem dunklen Verließ gefangen gehalten. Doch weil sich die Lebensbedingungen draußen wegen des zunehmenden Klimawandels radikal verschlechtert haben, ist Anna nicht daran gelegen, bspw. spazieren gehen oder verreisen zu können. Als Anhänger von PF kann man nun nicht behaupten, dass Anna unfrei dazu ist, spazieren zu gehen oder zu verreisen; doch ebenso wenig kann man sagen, dass sie zu diesen Handlungen frei ist. Denn da Anna weder am einen noch am anderen gelegen ist, erfüllt keine dieser Handlungen die dritte Bedingung von PF. Beide Handlungstypen gehören also nicht zu der Art von Handlungen, bzgl. derer sich die Frage nach Freiheit überhaupt stellt. Aber würde man nicht vielmehr sagen wollen, dass eine gefangene Person unfrei ist, ganz unabhängig davon, ob sie daran etwas auszusetzen hat oder nicht (vgl. Berlin 1996 [1969]: 215-221)? Eine Auffassung von Freiheit, die dies nicht formulieren kann, verstößt gegen unsere fundamentalen begrifflichen Intuitionen und damit gegen das oben geforderte Kriterium I.

Alle Auffassungen von Freiheit, die ich bisher dargestellt habe, verstehen Freiheit als opportunity-concept, "where being free is a matter of what we can do, of what is open to us to do, whether or not we do anything to exercise these options." (Taylor 1991 [1979]; 144) D.h. obwohl sie sich in Details unterscheiden, besteht für sie die Handlungsfreiheit eines Individuums darin, dass diesem Handlungsmöglichkeiten offenstehen. Doch man könnte Freiheit auch im Sinne eines exercise-concepts verstehen, wonach gilt: "[O]ne is free only to the extent that one has effectively determined oneself and the shape of one's life." (Taylor 1991 [1979]; 143) Freiheit besteht nach dieser Auffassung nicht in einer Möglichkeit, sondern im Ergreifen derselben, denn *Person A ist genau dann frei in Handlung x, wenn (1) A x vollzieht und (2) x eine selbstbestimmte Handlung As ist.*

Wie PF hat diese Auffassung den Vorteil, dass sie dem Kriterium P gerecht wird. Denn wenn Freiheit darin besteht, selbstbestimmt zu leben, so ist Freiheit zweifellos etwas, was uns wichtig ist und sein sollte. Doch mit PF teilt Freiheit im Sinne eines exercise-concepts nicht nur Vorteile, sondern auch Schwächen. Denn auch diese Auffassung kann nicht adäquat mit der eingesperrten Anna, die ihr Verlies nicht verlassen will, umgehen. Sofern es Anna nämlich wirklich vorzieht, eingesperrt zu sein, bestünde ihre Freiheit nach dieser Auffassung gerade darin, im Kerker zu verbleiben. Noch grundlegender kann man fragen, ob Freiheit für uns intuitiv wirklich im Vollziehen einer bestimmten Handlung besteht und nicht doch (wie es Auffassungen im Sinne eines opportunity-concepts behaupten) darin, dass die betreffende Handlungsoption nicht von anderen Personen blockiert wird. Natürlich impliziert der Vollzug einer bestimmen Handlung, dass man zu dieser frei gewesen ist, da Freiheit Können

---

[4] Nach Jahren des Streits darum, ob man Freiheit negativ oder positiv, also – so die gängigen Schlagworte – als *Freiheit von* oder als *Freiheit zu* verstehen sollte, hat Gerald MacCallum dafür argumentiert, dass man Freiheit ohnehin als dreistellige Relation auffassen sollte: "[...] freedom is thus always *of* something (an agent or agents), *from* something, *to* do, not do, become, or not become something; it is a triadic relation." (MacCallum 1991 [1967]: 102). Diese Bemerkung ist korrekt: Sowohl negative (in der Version $NF_3$) als auch positive Freiheit entsprechen dieser begrifflichen Vorgabe, da Freiheit jeweils darin besteht, dass A nicht *von* einer anderen Person darin gehindert wird, x *zu* tun. Als relevanter Unterschied bleibt aber bestehen, dass positive Freiheit bestimmte Anforderungen an x stellt und dadurch den Aspekt stärker betont, wozu man frei sein kann. Da negative Freiheit bzgl. x keinerlei Einschränkungen macht, rückt der "frei von"-Aspekt in den Vordergrund. Zu der Frage, welches Konzept angemessener ist, können MacCallums Überlegungen keinen Beitrag leisten.

voraussetzt; aber es scheint ein begrifflicher Kurzschluss zu sein, die Freiheit zu einer Handlung mit deren Durchführung gleichzusetzen. D.h. Freiheit als exercise-concept kann Kriterium I nicht entsprechen. Und auch der Forderung K wird diese Auffassung nicht gerecht. Denn wie meine obige Formulierung nahelegt, verfügen wir bereits über einen Begriff, der die Kernüberlegung dieser Freiheitsauffassung besser zum Ausdruck bringt. Eigentlich würden wir nämlich sagen wollen, dass eine Person, die frei ist im Sinne des excercise-concepts, eben selbstbestimmt ist.

Diese Darstellung verschiedener gängiger Auffassungen von Freiheit scheint die Sorge zu bestätigen, die ich anfangs formuliert habe: Ein Verständnis von Freiheit, das allen Kriterien gerecht wird, die wir an dieses Konzept herantragen, gibt es prima facie nicht. Allerdings erlaubt der Vergleich, zwei Auffassungen hervorzuheben. Denn im Gegensatz zu den anderen Positionen konfligieren $NF_3$ und PF jeweils nur mit einem der drei geforderten Kriterien.

Wenn man nun abschließend diese beiden Auffassungen nochmals betrachtet, wird deutlich, dass $NF_3$ angemessener als PF ist. PF hatte die kontraintuitive Implikation, dass ein in Ketten liegender Mensch nicht als unfrei zu all den Handlungen bezeichnet werden kann, die ihm aufgrund seiner Gefangenschaft zwar unmöglich, zudem aber gleichgültig sind. Auf diese begriffliche Schwierigkeit können Anhänger von PF nichts entgegnen, sie müssen sie schlicht akzeptieren. Das Problem von $NF_3$ war dagegen, dass eine Person nach dieser Auffassung von anderen unfrei zu bestimmten Handlungen gemacht wird, ohne dass diese Beschneidung ihrer Handlungsoptionen kritisch zu bewerten wäre. D.h. nach dieser Auffassung von Freiheit ist nicht jeder Freiheitsspielraum schützenwert und nicht jede Freiheitseinbuße kritikwürdig. Dieses Ergebnis erstaunt, da „Freiheit“ für uns positiv besetzt ist. Doch hierauf kann man mit dem Hinweis antworten, dass „Freiheit“ auch dann noch positiv besetzt sein kann, wenn uns die Freiheit des Individuums meistens, wenngleich nicht in jedem Fall schützenswert erscheint. Anders formuliert: Dass Handlungsfreiheit schätzenswert und also „Freiheit“ positiv besetzt ist, lässt sich auch ausgehend von $NF_3$ behaupten. Denn dies kann der Fall sein, selbst wenn nur bestimmte Freiheitspielräume zu schützen sind, während andere durchaus eingeschränkt werden können, ohne dass dies zu kritisieren sei. Insofern lässt sich abschließend festhalten, dass $NF_3$ alle erforderlichen Kriterien erfüllt, so dass ich dafür plädiere, deskriptive Handlungsfreiheit im Folgenden im Sinne von $NF_3$ zu verstehen.

## 3. Staatliches Handeln und die Freiheit der Bürgerinnen und Bürger

Vor dem Hintergrund dieses Freiheitsbegriffes lässt sich nun zeigen, dass staatliche Handlungen *immer* eine Einschränkung der Handlungsfreiheit der Bürgerinnen und Bürger darstellen. Dabei sollen unter „staatlichen Handlungen“ die Handlungen staatlicher Institutionen verstanden werden, also etwa die Gesetzgebung durch das Parlament oder die Ausführung der Gesetze durch die Exekutive. Allerdings werde ich die Gesetzgebung als paradigmatischen Fall staatlichen Handelns verstehen. Denn wann immer eine staatliche Institution handelt, tut sie das auf der Grundlage eines gesetzgeberischen Aktes (vgl. Raz 1988 [1986]: 3).

Aber stellen gesetzgeberische Akte tatsächlich eine Einschränkung der Handlungsfreiheit der Bürgerinnen und Bürger dar? Dies könnte man zum einen deswegen bezweifeln, weil es Gesetze den Individuen nicht strictu sensu unmöglich machen Handlungen auszuführen: Zwar besagt die Straßenverkehrsordnung, dass Anna nicht über eine rote Ampel fahren soll, doch dadurch wird Anna nicht daran gehindert, eine rote Ampel zu überfahren. Dies wird sie zwar unter Umständen teuer zu stehen kommen, aber dazu in der Lage ist sie nach wie vor. Zum anderen sollte man zugestehen, dass bestimmte Handlungsspielräume für die Bürgerinnen und Bürger erst durch staatliche Handlungen eröffnet werden. So kann Anna

nur innerhalb eines Staates bzw. eines Staatensystems Außenministerin werden. Denn dieses Amt wäre sinnlos, gäbe es keinen Staat, der gegenüber anderen Staaten repräsentiert werden muss. Ebenso – wenngleich nicht rein begrifflich – gilt, dass etwa größere Maßnahmen zur sozialen Umverteilung materieller Güter und die Etablierung eines verbindlichen Rechtssystems nur innerhalb von Staaten dauerhaft durchzuführen sind, so dass Anna erst durch den Staat in den Genuss dieser Vorteile kommt. Warum also sollte man staatliche Handlungen als Einschränkungen und nicht vielmehr als Ermöglichungen ansehen?

Um zu verstehen, inwiefern das Erlassen von Gesetzen den Bürgerinnen und Bürgern Handlungen unmöglich macht, möchte ich zuerst eine andere Klasse von Handlungen betrachten, nämlich das Aussprechen von Drohungen. Drohungen haben im Allgemeinen die folgende Struktur: Eine Person A setzt Person B davon in Kenntnis, dass, wenn B die Handlung y tut, A die Handlung x vollziehen wird. Darüber hinaus ist für Drohungen kennzeichnend, dass durch die Ankündigung, dass das Ausführen von y durch B das Ausführen von x durch A nach sich ziehen wird, die Attraktivität von y für B sinkt. Wenn Anna Ben ankündigt, dass sie nie wieder mit ihm reden wird, wenn Ben nicht pünktlich am verabredeten Treffpunkt erscheint, so spricht Anna damit eine Drohung gegenüber Ben aus, vorausgesetzt die Ankündigung, dass Anna nie wieder mit ihm reden wird, lässt die Option zu spät zu kommen in Bens Augen weniger erstrebenswert erscheinen.

Doch selbst wenn durch die Drohung die Attraktivität von y für B stark sinkt – unmöglich wird y dadurch für B nicht. Ben kann die Handlung “zur Verabredung mit Anna zu spät kommen” immer noch ausführen, wenn auch nur um einen gewissen Preis. Vermittelt die Betrachtung von Drohungen also nicht denselben Eindruck wie zuvor die von gesetzgeberischen Handlungen, nämlich dass nur in einem übertragenen Sinne davon gesprochen werden kann, dass beide Handlungen unmöglich machen? Ich denke nicht: Denn während Ben die einfache Handlung “zur Verabredung mit Anna zu spät kommen” trotz Annas Drohung noch offen steht, wird ihm durch die Drohung die *komplexe Handlung* “zur Verabredung mit Anna zu spät kommen und immer noch Annas Freund sein” unmöglich. Hat Anna ihre Drohung einmal ausgesprochen, wird dadurch die genannte komplexe Handlung zwar wahrscheinlich zum ersten Mal von Ben als eine mögliche Handlung erwogen, aber zudem wohl als diejenige erkannt, die er eigentlich ausführen möchte. Doch gleichzeitig wird diese komplexe Handlung durch die Drohung aus Bens Menge möglicher Handlungen gestrichen: In dieser Welt wird er nicht mehr zu spät zu einer Verabredung mit Anna kommen und dennoch deren Freund bleiben können. Dies ist der für Drohungen kennzeichnende Effekt: “Drohungen [...] verhindern, machen unfrei zu komplexen Handlungen.” (Wendt 2009: 43)

Meiner Ansicht nach sollte man das Erlassen von Gesetzen seitens des Staates analog zum Aussprechen einer Drohung gegenüber den Bürgerinnen und Bürgern verstehen. Denn Gesetze sind typischerweise zwangsbewehrt, d.h. ihre Übertretung wird bestraft. Erlässt der Staat ein Gesetz, so kündigt er damit implizit an, dass die Bürgerinnen und Bürger zur Rechenschaft gezogen werden, falls sie sich nicht an dieses Gesetz halten. Setzt man zudem voraus, dass die Strafen für Gesetzesübertretungen so gewählt sind, dass sie Gesetzesverstöße unattraktiv machen, so entspricht das Erlassen eines Gesetzes dem Aussprechen einer Drohung. Das Erlassen eines Gesetzes macht also bestimmte komplexe Handlungen für die Bürgerinnen und Bürger unmöglich. Sobald die Straßenverkehrsordnung in Kraft getreten ist, kann Anna bspw. nicht mehr die komplexe Handlung ausführen „in Anwesenheit der Polizei über eine rote Ampel fahren und dafür nichts zahlen“. D.h. dadurch, dass ein Gesetz erlassen wurde, hat der Staat Anna eine komplexe Handlung unmöglich gemacht, die ihr ohne diesen gesetzgeberischen Akt offen gestanden hätte – und sie damit nach obiger Darstellung in ihrer Freiheit eingeschränkt.

Doch selbst wenn Gesetze und damit staatliches Handeln bestimmte komplexe Handlungen der Bürgerinnen und Bürger unmöglich machen: Wie begegnet man obigem Einwand, dass

andere Taten – etwa das Kandidieren für ein politisches Amt oder das Erheben einer Anklage vor Gericht – durch staatliche Handlungen erst ermöglicht werden? Hierauf kann man mit dem Hinweis reagieren, dass jede staatliche Handlung durch Steuergelder ermöglicht wird. Denn es sind Steuergelder, aus denen Beamtenstellen finanziert und mit denen Gesetzestexte gedruckt werden, ebenso wie es Steuergelder sind, die im Rahmen sozialer Ausgleichszahlungen umverteilt werden. Doch das Erheben von Steuern seitens des Staates stellt natürlich eine Einschränkung der Handlungsfreiheit der Bürgerinnen und Bürger dar. Denn ihnen wird die Handlungsalternative genommen, selbst zu entscheiden, was mit einem Teil ihres Geldes geschieht (vgl. Gaus 2003: 147).[5]

## 4. Das liberale Prinzip und die Freiheit der Bürgerinnen und Bürger

Bisher habe ich zweierlei geleistet: In 1 und 2 habe ich für eine bestimmte Auffassung von Handlungsfreiheit argumentiert und in 3 habe ich gezeigt, dass man – ausgehend von dieser Auffassung – feststellen muss, dass staatliche Handlungen die Freiheit der Bürgerinnen und Bürger einschränken. In diesem Teil meiner Ausführungen möchte ich nun zuerst auf ein kurioses Faktum hinweisen, nämlich darauf, dass es dem Liberalismus offenbar gar nicht darum geht, die Freiheit der Bürgerinnen und Bürger umfassend vor Eingriffen durch den Staat zu schützen. Denn obwohl diese Strömung der politischen Philosophie die Freiheit sogar im Namen trägt, setzt sie sich tatsächlich nur für den Schutz *bestimmter* Handlungsspielräume der Bürgerinnen und Bürger ein. Abschließend will ich dafür argumentieren, dass man dies nicht als Schwäche des Liberalismus betrachten sollte. Denn Handlungsfreiheit ist nicht uneingeschränkt schützenswert und der Liberalismus versteht es, einen wichtigen Handlungsspielraum der Bürgerinnen und Bürger zu sichern – nämlich denjenigen, den sie benötigen um autonom zu handeln.

Um diese Überlegung auszuführen, möchte ich zuerst ein Verständnis von Liberalismus vorschlagen. Dabei betrachte ich den Liberalismus wie angemerkt als Strömung der politischen Philosophie und damit als Theorie, die sich mit der Frage staatlicher Legitimität und also damit beschäftigt, was ein Staat tun darf und was nicht. Die Antwort des Liberalismus auf diese Frage besteht meiner Ansicht nach in folgendem *liberalen Prinzip*: *Eine staatliche Handlung x ist genau dann legitim und darf vollzogen werden, wenn x vor allen Bürgerinnen und Bürgerinnen gerechtfertigt ist.* Dabei ist eine Handlung x vor einer Bürgerin A genau dann gerechtfertigt wenn (1) mindestens eine Tatsache T konklusiv für x spricht, (2) A um T weiß und (3) A vernünftigerweise darauf festgelegt ist, T als konklusiven Grund für x anzusehen. Den Maßstab, der bestimmt, ob A vernünftigerweise darauf festgelegt ist, T als konklusiven Grund für x anzusehen, bilden dabei die weiteren Gründe, über die A rationaliter verfügt. Wenn also aus dem Gesamt von As guten Gründen folgt, dass T konklusiv für x spricht, so sollte A T als konklusiven Grund für x akzeptieren. Diese Auffassung von Rechtfertigung lässt einerseits zu, dass den Bürgerinnen und Bürgern eines Staates nicht immer unmittelbar bewusst ist, was vor ihnen gerechtfertigt ist. So kann es vorkommen, dass Anna nicht überblickt, worauf sie rationaliter festgelegt ist, weil sie relevante Fakten vergessen hat oder einen Zusammenhang nicht erkennt. Insofern muss es sich nicht notwendigerweise in der expliziten Zustimmung der Bürgerinnen und Bürger zu einer Handlung manifestieren, wenn diese vor ihnen gerechtfertigt ist. Das liberale Prinzip fordert also eine *hypothetische* Zustimmung. Andererseits lässt sich nach dieser Auffassung nur

[5] Es ist denkbar, dass ein Staat Handlungen vollzieht, die nicht auf die Erhebung von Steuergeldern zurückzuführen sind und insofern nicht auf eine Einschränkung der Bürgerinnen und Bürger zurückgehen. Dies könnte etwa der Fall sein, wenn der Staat den Bau eines Museums durchführt, das ausschließlich durch die Spenden einer großzügigen Milliardärin finanziert wird. Da Fälle wie dieser nicht allzu häufig auftreten dürften, werde ich sie im Folgenden vernachlässigen.

ausgehend vom einzelnen Individuum ermitteln, worauf dieses rationaliter festgelegt ist. Denn es sind die Gründe, über die A bereits rationaliter verfügt, die festlegen, welche weiteren Gründe A akzeptieren sollte. Dass Ben rationaliter darauf festgelegt ist, die Tatsache T als konklusiven Grund für eine staatliche Handlung x anzusehen, impliziert also nicht notwendigerweise, dass T auch vor Anna für x spricht. Dies ist nur dann der Fall, wenn Anna die Gründe, in deren Licht T für Ben konklusiv für x spricht, begründetermaßen teilt. Insofern ist die Auffassung von Rechtfertigung, die dem liberalen Prinzip zugrunde liegt, *nicht-monologisch*.

Vielleicht erscheint diese Charakterisierung des Liberalismus zu abstrakt. Stellt der Liberalismus nicht viel konkretere Forderungen an einen legitimen Staat, wie etwa, dass er seinen Bürgerinnen und Bürgern gleiche maximale Freiheitsrechte einräumen und zudem dafür Sorge tragen muss, dass sich ökonomische Ungleichheiten zugunsten der gesellschaftlich am schlechtesten Gestellten auswirken und deren Chancen auf soziale Positionen nicht beeinträchtigen (vgl. Rawls 2003 [1971]: 52-3 und 72)? Doch gerade die Theorie John Rawls', die prima facie als Beispiel für solche konkreten Forderungen herangezogen werden kann, zeigt, dass diese nicht den normativen Kern des Liberalismus ausmachen. Denn nach Rawls sollte sich der liberale Staat nur deswegen an genannten Grundsätzen orientieren, weil sie im Urzustand gewählt werden würden. D.h. beide Grundsätze müssen dem Kriterium der Urzustandswahl entsprechen und es ist dieses übergeordnete Kriterium, an dem letztlich die Legitimität staatlichen Handelns hängt. Dabei hat der Urzustand den Effekt, Eigeninteresse-orientierte sowie an partikularen Vorstellungen vom Guten ausgerichtete Abwägungen unmöglich zu machen und die potentiell Wählenden damit zu Entscheidungen zu zwingen, die vor allen Menschen gerechtfertigt sind. Zu fordern, dass staatliches Handeln Gegenstand einer Urzustandswahl sein können muss, heißt also zu fordern, dass staatliches Handeln vor allen Menschen und damit insbesondere vor allen Bürgerinnen und Bürgern gerechtfertigt sein muss. Der normative Kern des Liberalismus ist also das liberale Prinzip.

Akzeptiert man das liberale Prinzip so wird deutlich, dass etwas nicht der Fall ist, was man gemeinhin annimmt. Gemeinhin geht man nämlich davon aus, dass der Liberalismus freiheitsorientiert ist, d.h. dass es ihm um den umfassenden Schutz der Handlungsfreiheit der Bürgerinnen und Bürger geht (vgl. Gaus und Courtland 2011). Doch wenn ein Individuum frei ist, wenn ihm von niemandem Handlungen unmöglich gemacht werden, zu denen es eigentlich in der Lage wäre; wenn staatliches Handeln aber immer zur Folge hat, dass dem Individuum Handlungen unmöglich gemacht werden, die ihm ohne staatliche Intervention offengestanden hätten; dann müsste der Liberalismus alle Handlungen des Staates unterbinden, wollte er die individuelle Handlungsfreiheit umfassend schützen. Die einzige Ausnahme wären solche staatlichen Handlungen, die selbst auf die Verhinderung weiterer Freiheitseinschränkungen abzielen, wie etwa polizeiliche Maßnahmen, die Bürgerinnen und Bürger vor Freiheitseinschränkungen durch andere Bürgerinnen und Bürger schützen. Denn in diesem Falle würde gelten, was Immanuel Kant in der *Metaphysik der Sitten* festhält:

> Der Widerstand, der dem Hindernis einer Wirkung entgegengesetzt wird, ist eine Beförderung dieser Wirkung und stimmt mit ihr zusammen. [...] Folglich: wenn ein gewisser Gebrauch der Freiheit selbst ein Hindernis der Freiheit nach allgemeinen Gesetzen (d.i. unrecht) ist, so ist der Zwang, der diesem entgegengesetzt wird, als *Verhinderung* eines *Hindernisses der Freiheit* mit der Freiheit nach allgemeinen Gesetzen zusammen stimmend, d.i. recht [...]. (Kant 1985 [1797/98]: A34,35/ B35)

Aber das liberale Prinzip verbietet durchaus nicht alle staatlichen Handlungen bis auf diejenigen, die selbst wieder dem Freiheitsschutz dienen. Vielmehr sind nach diesem Prinzip bestimmte staatliche Handlungen legitim obwohl sie, wie jede staatliche Handlung, die Einschränkung der Handlungsfreiheit der Bürgerinnen und Bürger mit sich bringen – nämlich diejenigen, die vor allen Bürgerinnen und Bürgern gerechtfertigt sind. Insofern dient

das liberale Prinzip offenbar nicht dazu, die Handlungsfreiheit der Individuen gegenüber dem Staat umfassend zu wahren, und der Liberalismus erscheint nicht freiheitsorientiert zu sein.

Doch wie meine Ausführungen zur Handlungsfreiheit in 2 gezeigt haben, muss man dies nicht als Versäumnis des Liberalismus betrachten. Denn man sollte nicht hinter die Einsicht zurückfallen, dass nicht jede Instanz individueller Handlungsfreiheit schützenswert ist. Ginge es dem Liberalismus um den umfassenden Schutz individueller Handlungsfreiheit, würde er staatliche Handlungen für illegitim erklären, die in den Augen aller Bürgerinnen und Bürger sinnvoll sein sollten (wie etwa das Einrichten einer Straßenverkehrsordnung), um stattdessen Freiheitsspielräume zu wahren, an denen niemandem vernünftigerweise gelegen sein kann (wie etwa nach Belieben Auto fahren zu können). Wenn man sich der Vorstellung verpflichtet fühlt, dass die individuelle Handlungsfreiheit zu schützen ist, sollte man sich also genauer gesagt der Vorstellung verpflichtet fühlen, dass die *relevanten* Freiheitsspielräume des Einzelnen zu schützen sind. Sofern das liberale Prinzip dazu dient, gerade diese Freiheitsspielräume der Bürgerinnen und Bürger vor Eingriffen durch Staat zu schützen, könnte man also den Liberalismus auch weiterhin als freiheitsorientiert bezeichnen – und das obwohl er nicht alle Instanzen individueller Handlungsfreiheit den Eingriffen des Staates entzieht.

Zusammenfassend lässt sich Folgendes feststellen: Das liberale Prinzip schützt nicht die gesamte Handlungsfreiheit der Bürgerinnen und Bürger vor Übergriffen durch den Staat. Denn jede staatliche Handlung muss als Einschränkung der individuellen Freiheit gewertet werden, das liberale Prinzip verbietet aber nicht rundheraus alle staatlichen Handlungen. Allerdings ist individuelle Handlungsfreiheit nicht insgesamt schützenswert, sondern nur bestimmte Aspekte oder Teile von dieser. Falls das liberale Prinzip gerade die relevanten Freiheitsspielräume schützt, spricht daher nichts dagegen, den Liberalismus weiterhin als der Freiheit verpflichtet anzusehen.

Doch welche Bereiche der individuellen Handlungsfreiheit der Bürgerinnen und Bürger werden durch das liberale Prinzip geschützt? Um dies zu beantworten, möchte ich zuerst klären, was es eigentlich bedeutet, dass eine staatliche Handlung x vor einer beliebigen Person A gerechtfertigt ist. Denn gemäß dem liberalen Prinzip sind gerade diejenigen staatlichen Handlungen legitim, die diese Eigenschaft aufweisen. Dabei ist zu berücksichtigen, dass A durch jede staatliche Handlung zumindest einer Handlungsalternative beraubt wird, die ihr vor der Handlung des Staates offenstand, nämlich ungestraft gegen das zu verstoßen, was durch das neu erlassene Gesetz vorgeschrieben wird. Wenn nun eine staatliche Handlung in As Augen gerechtfertigt ist, dann bedeutet das, dass in As eigenen Augen die besseren Gründe dafür sprechen, einer bestimmten Alternative beraubt zu werden als weiterhin die Möglichkeit zu dieser zu haben. Doch wenn nach A die besseren Gründe für die Verhinderung einer ihr möglichen Handlung als für das Durchführen dieser Handlung sprechen, so kann die fragliche Handlung nicht konklusiv vor A gerechtfertigt gewesen sein. Wenn es bspw. in den Augen Annas gerechtfertigt ist, dass die Straßenverkehrsordnung eingeführt wird, so ist offensichtlich in Annas Augen gerechtfertigt, dass sie in Zukunft nicht mehr die Möglichkeit haben wird, straflos nach Gutdünken Auto zu fahren. Aber das kann nur der Fall sein, wenn es in Annas Augen alles in allem betrachtet keinen zwingenden Grund gab, nach Gutdünken Auto zu fahren. Gemäß dem liberalen Prinzip darf ein Staat also nur in solchen Fällen aktiv werden und den Bürgerinnen und Bürgern durch seine Aktivität Instanzen ihrer Handlungsfreiheit rauben, in denen die einzuschränkenden Handlungen nicht konklusiv vor den Bürgerinnen und Bürgern gerechtfertigt waren. An Handlungen, die auszuführen sie alles in allem zwingenden Grund haben, dürfen die Bürgerinnen und Bürger dagegen nicht gehindert werden.

Vor dem Hintergrund des Autonomiebegriffs von Kant lässt sich diese letzte Einsicht folgendermaßen re-formulieren: *Nach Kant handelt eine Person A genau dann autonom in einer Handlung x, wenn (1) x alles in allem betrachtet vor A gerechtfertigt ist und (2) A*

*durch die Einsicht, dass x alles in allem betrachtet vor A gerechtfertigt ist, zu A motiviert wird* (vgl. Kant 1999 [1785]: 433). D.h. für Kant ist es eine notwendige Bedingung für die Autonomie einer Person, dass sie diejenigen Handlungen vollzieht, zu denen sie alles in allem Grund hat. Aber es sind gerade diese Handlungen – die Handlungen, zu denen sie alles in allem betrachtet Grund haben – an denen die Bürgerinnen und Bürger gemäß dem liberalen Prinzip nicht gehindert werden dürfen. Was der Liberalismus mittels des liberalen Prinzips wahren will, ist also nicht die individuelle Handlungsfreiheit der Bürgerinnen und Bürger tout court, sondern deren Möglichkeit, autonom zu handeln. Der Freiheitsspielraum, den Liberalismus sichert, ist der Raum individueller Autonomie.

## 5. Ausblick

Insgesamt haben meine Ausführungen dreierlei gezeigt: Dass man eine Person A genau dann als frei zu einer Handlung x ansehen sollte, wenn A nicht von anderen Personen an x gehindert wird, wobei A grundsätzlich in der Lage dazu sein muss, x auszuführen; dass staatliche Handlungen vor dem Hintergrund dieses Freiheitsbegriffs immer Freiheitseinschränkungen für die Bürgerinnen und Bürger darstellen; und dass der Liberalismus nicht darum bemüht ist, die Freiheit der Bürgerinnen und Bürger als Ganze zu schützen, sondern lediglich deren Möglichkeit zu autonomem Handeln sichern will.

Was noch nicht gezeigt wurde, ist zum einen, ob der Liberalismus damit gerade denjenigen Teil der individuellen Handlungsfreiheit der Bürgerinnen und Bürger schützt, an dem diesen gelegen sein sollte, d.h. ob der Liberalismus den relevanten Teil individueller Handlungsfreiheit schützt. Um dies zu zeigen, müsste ich ausführen, was gut daran ist, autonom zu handeln – was allerdings den Rahmen dieser Untersuchung sprengen würde. Ebenfalls offen lassen muss ich an dieser Stelle, mit wie vielen und welchen staatlichen Handlungen und damit mit wie vielen und welchen Einschränkungen ihrer Handlungsfreiheit die Bürgerinnnen und Bürger eines liberalen Staates zu rechnen haben. Denn auch diese Überlegung – d.h. die Überlegung, welche staatlichen Handlungen vor allen Bürgerinnen und Bürgern zu rechtfertigen sind – würde zu viel Raum einnehmen, sofern man sie überhaupt ohne genauere Kenntnis des zu beurteilenden Staates bzw. der zu beurteiltenden Bürgerschaft anstellen kann.

Aber zumindest einem Einwand möchte ich an dieser Stelle noch entgegentreten: Der eine oder die andere könnte die Tatsache, dass der Liberalismus den Bürgerinnen und Bürgern nur die Möglichkeit zu autonomem und das heißt zu vor ihnen alles in allem gerechtfertigtem Handeln einräumt, als ersten Schritt in die Tugenddiktatur verstehen. Sollte es den Bürgerinnen und Bürgern wirklich nur noch frei stehen, das zu tun, wozu sie Grund haben? Haben wir bspw. alles in allem betrachtet Grund dazu, unsere Wochenenden faul im Bett zu verbringen – und darf uns ein liberaler Staat, sofern das nicht der Fall ist, diese Handlung verbieten? Diese Sorge lässt sich meiner Ansicht nach dadurch entkräften, dass man sich bspw. an John Stuart Mills Ausführungen in „On Liberty“ erinnert (vgl. Mill 2008 [1859]: Kapitel 2 und 3). Dort weist Mill darauf hin, dass wir guten Grund dazu haben, den Menschen das Machen bestimmter Fehler – nämlich solcher, durch die niemand außer die Handelnde selbst geschädigt wird – nicht zu verbieten. Denn nur aus selbst gemachten Erfahrungen kann der Mensch lernen und ein individuelles Leben entwickeln; und nur auf der Grundlage vieler individueller Erfahrungen und Lebensentwürfe kann Wissen erworben werden und gesellschaftlicher Fortschritt stattfinden. D.h. unter Umständen ist es alles in allem gerechtfertigt, die Bürgerinnen und Bürger selbst herausfinden zu lassen – und zwar auch mittels trial and *error* –, ob sie alles in allem betrachtet Grund dazu haben, sich aus dem Bett zu quälen oder eben nicht.

**Christine Bratu**

Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft
Ludwig-Maximilians-Universität München
christine.bratu@lrz.uni-muenchen.de

## Literatur

Berlin, I. 1996 [1969]: „Zwei Freiheitsbegriffe“, in *Freiheit. Vier Versuche*, Frankfurt a.M.: Fischer, 197-256.

Carter, I. 2012: „Positive and Negative Liberty“ in Edward N. Zalta (Hrsg.): *The Stanford Encyclopedia of Philosophy (Spring 2012 Edition)*, URL = <http://plato.stanford.edu/archives/spr2012/entries/liberty-positive-negative/>

Carter, I., M. Kramer und H. Steiner 2007: „General Introduction“, in *Freedom. A Philosophical Anthology*, Malden et al.: Blackwell, xvii-xxi.

Gaus, G. 2003: „Liberal Neutrality: A Compelling and Radical Principle“, in: S. Wall und G. Klosko (Hrsg.): *Perfectionism and Neutrality*, Lanham et al.: Rowman & Littlefield Publishers, 137-165.

Gaus, G. und S. Courtland 2011: „Liberalism“, in Edward N. Zalta (Hrsg.): *The Stanford Encyclopedia of Philosophy (Spring 2011 Edition)*, URL = <http://plato.stanford.edu/archives/spr2011/entries/liberalism/>.

Hobbes, T. 2000 [1651]: *Leviathan oder Stoff, Form und Gewalt eines kirchlichen und bürgerlichen Staates*. Frankfurt a.M.: Suhrkamp.

Kant, I. 1999 [1785]: *Grundlegung zur Metaphysik der Sitten*, Hamburg: Meiner.

Kant, I. 1985 [1797/98]: *Die Metaphysik der Sitten*, Frankfurt a.M.: Suhrkamp.

MacCallum, G. C. 1991 [1967]: „Negative and Positive Freedom“, in D. Miller (Hrsg.): *Liberty*, Oxford et al.: Oxford University Press, 100-122.

Mill, J.S. 2008 [1859]: *On Liberty and other writings*, Cambridge (MA) et al.: Cambridge University Press.

Miller, D. 1991: „Introduction“, in *Liberty*, Oxford et al.: Oxford University Press, 1-20.

Rawls, J. 2003 [1971]: *A Theory of Justice*, Cambridge/ London: Belknap.

Raz, J. 1988 [1986]: *The Morality of Freedom*, Oxford et al.: Clarendon.

Taylor, C. 1991 [1979]: „What's Wrong with Negative Liberty?“ in D. Miller (Hrsg.): *Liberty*, Oxford et al.: Oxford University Press, 141-162.

Wenar, L. 2008: „The Meanings of Freedom“, in L. Thomas (Hrsg.): *Contemporary Debates in Social Philosophy*, Malden et al.: Blackwell, 43-53.

Wendt, F. 2009: *Libertäre politische Philosophie*, Paderborn: Mentis.

# Im Namen der Autonomie? Eine kritische Untersuchung des liberalen Paternalismus am Beispiel von Maßnahmen des kognitiven Enhancements

Rebecca Gutwald

Im vorliegenden Beitrag möchte ich einen kritischen Blick auf den sog. liberalen Paternalismus werfen. Der Rechtfertigungsstrategie über die liberale Forderung nach Respekt für Autonomie sollten meiner Ansicht nach Grenzen gesetzt werden, auch wenn Respekt für Autonomie vermeintlich gewahrt wird. Kritisch zu betrachten sind vor allem die Arbeiten von Cass Sunstein und Richard Thaler sowie einigen Medizinethikern, die einen Autonomie-fördernden bzw. verbessernden Paternalismus unterstützen. Dieser erscheint, so meine Argumentation, vor allem fragwürdig, wenn wir neue, technische Möglichkeiten betrachten, um Autonomie und Entscheidungen (vermeintlich) zu verbessern, insbesondere Maßnahmen des sog. „Cognitive Enhancements". Der Einsatz bzw. Förderung dieser Maßnahmen scheint eine logische Konsequenz aus der Befürwortung eines weichen Paternalismus im Stile von Sunstein/Thaler. Meine These ist jedoch, dass dies nicht der liberalen Grundidee des Freiheitsschutzes entspricht: nicht alles, was Autonomie befördert, muss auch getan werden, selbst wenn die Risiken gering sind. Vielmehr muss dem Menschen auch die Möglichkeit eingeräumt werden, eigene Fehler zu machen und Irrtümer zu begehen. Mein Anliegen ist hier vor allem, eine kritische Frage gegenüber Theoretikern wie Sunstein/Thaler aufzuwerfen, nämlich, was aus ihrem verbessernden, weichen Paternalismus folgt und inwiefern dieser noch als „liberal" aufgefasst werden kann.

Die Arbeiten von Cass Sunstein und Richard Thaler haben dazu beigetragen, Paternalismus in einem politisch-liberalen System wieder salonfähig zu machen.[1] Sunstein/Thaler bezeichnen die von Ihnen definierte und befürwortete Form des wohlwollenden Eingriffs in die Freiheit von Menschen als „weichen" bzw. „libertären" Paternalismus. „Libertär", da die Betroffenen des Eingriffs noch die Möglichkeit haben, sich frei anders zu entscheiden – zumindest steht eine gewisse Möglichkeit dazu offen. Das allgemeine Ziel solcher Maßnahmen ist demnach, das Wohlergehen von Menschen zu schützen bzw. zu fördern, ohne die liberale Grundforderung nach Respekt für Selbstbestimmung zu verletzen.

In dem vorliegenden Beitrag möchte ich einen kritischen Blick auf eben jenen „libertären" Paternalismus werfen. Der Rechtfertigungsstrategie über die liberale Forderung nach Respekt für Autonomie sollten meiner Ansicht nach Grenzen gesetzt werden. Exemplarisch zeigt sich dies an einem extrem gewählten Beispiel: wenn wir neue, technische Möglichkeiten betrachten, um Autonomie und Entscheidungen (vermeintlich) zu verbessern, insbesondere Maßnahmen des sog. „cognitive enhancement", könnten diese durch den libertären Paternalismus à la Sunstein/Thaler durchaus gerechtfertigt werden. Meine These ist: nicht alles, was Autonomie respektiert bzw. sogar befördert, ist auch rechtfertigbar, weil es uns auf eine ‚slippery slope' führen kann. Mit dieser These soll Sunstein/Thalers Theorie nicht widerlegt werden, sondern vielmehr ein gewisses Unbehagen aus ethischer Sicht ausgedrückt werden. Hier soll es also um die Grenzen des weichen Paternalismus gehen, die bisher wenig bis gar nicht untersucht sind.

[1] vgl. Sunstein C., Thaler R. 2008

## 1. Paternalismus

Es hat sich in der philosophischen Diskussion eingebürgert, betreffend Paternalismus zwei Unterscheidungen zu treffen. Zum einen unternimmt man eine Trennung zwischen der *deskriptiven* Frage der Definition von paternalistischem Verhalten und seiner *normativen* Rechtfertigung. Die allgemeine Paternalismusdefinition wird also nur als beschreibend aufgefasst: sie impliziert noch nicht, ob das definierte Verhalten legitim ist oder nicht.[2]

Zum anderen wird als Basis zur Rechtfertigung zwischen *weichem* und *hartem* Paternalismus unterschieden, wobei letzterer aus liberaler Sicht als illegitim gekennzeichnet wird, da er die Freiheit des Betroffenen in ungebührlicher Weise einschränkt. Im Folgenden werde ich auf die allgemeine Definitionsproblematik eingehen und im zweiten Schritt auf die Unterscheidung hart/weich.

### *1.1 Definition*

Die Etymologie des Begriffs „Paternalismus" legt nahe, dass es sich um eine Art „väterliches" Verhalten gegenüber einem Kind bzw. einem Subjekt handelt, das dem Handelnden nicht gleichwertig gegenüber steht. Folgendes Beispiel von John Stuart Mill illustriert diese Art des Eingriffs (etwas vereinfacht):

> Ein Mann will auf eine Brücke gehen, die stark beschädigt ist. Man darf laut Mill den Mann aufhalten, notfalls mit Gewalt, um ihn davon abzuhalten bzw. zumindest zu informieren, dass die Brücke beschädigt ist, und er ins Wasser fallen könnte.

Es wird also durch einen (noch näher zu fassenden) Eingriff die Freiheit bzw. Autonomie eines Einzelnen beschränkt, um das Wohl dieser Person zu schützen oder zu verbessern.

Über diesen Kern hinaus ist es aber nicht unbedingt klar, was alles unter paternalistisches Verhalten fällt. Es muss sich beispielsweise nicht unbedingt um eine aktive Intervention wie im eben genannten Beispiel handeln: möglich ist auch die Manipulation einer Entscheidung durch Verschweigen oder Lüge, um den Entscheidenden sanft zu seinem Wohl zu führen.[3] Beispielsweise war es früher häufig der Fall, dass ein Arzt einem todkranken Patienten die terminale Diagnose verschwiegen hat, um ihm die Aufregung zu ersparen. Ebenso kann das Ziel von paternalistischen Handeln sein, die Freiheit bzw. Autonomie eines Individuums zu ermöglichen bzw. zu fördern.[4]

Eine Definition von Paternalismus, welche diese verschiedenen Arten von Verhalten einschließt, muss – so meine Ansicht – weit und allgemein gefasst sein.[5] Mein Vorschlag ist folgender (auch im Hinblick auf die folgende Unterscheidung von hart/weich): Paternalismus ist eine Art von wohlwollender Kontrolle, die über ein Individuum ausgeübt wird, ohne dessen gegenwärtigen Präferenzen zu achten.[6]

Diese Art der Definition mag auf den ersten Blick etwas zu vage erscheinen, sie wird aber durch die Unterscheidung hart/weich weiter präzisiert, auf welche ich im Folgenden eingehen werde.[7]

---

[2] vgl. Feinberg 1989

[3] vgl. Clarke 2002

[4] vgl. Beauchamp, Childress 2001

5 Im Gegensatz zu anderen Definitionsversuchen, wie etwa denen von Donald vanDeVeer (1986), welche die Definition immer weiter verzweigen und eine Sammlung von Verhaltensweisen explizit aufnehmen. (vgl. Clarke 2002)

[6] vgl. Gutwald 2010

[7] Einen guten Überblick zur Unterscheidung zwischen hartem und weichem Paternalismus bietet Dworkin G. 2010

### *1.2 Harter Paternalismus*

Ziehen wir wiederum das o.g. Millsche Beispiel heran: Stellen wir uns vor, der Paternalist hält den Mann davon ab, auf die Brücke zu gehen und sich selbst zu verletzen (evtl. sogar absichtlich). Dieser „harte" Paternalist würde damit auch eingreifen, wenn der Mann sich selbst verletzen oder gar töten will.

Harter Paternalismus kann damit – so wie man Paternalismus klassischerweise kennt– als wohlwollende Kontrolle über das Verhalten eines Individuums definiert werden, um dessen Wohl zu schützen; dies geschieht ggf. auch gegen die autonom gefassten Präferenzen bzw. Lebenspläne eines Individuums. Ein bekanntes Beispiel für dieses Verhalten kommt aus der Medizin: früher wurden bei Zeugen Jehovas Bluttransfusionen durchgeführt, obwohl diese solche Eingriffe aus religiösen Gründen ablehnten und auch nicht umgestimmt werden konnten. In diesem Fall würde gegen die Autonomie des Individuums gehandelt.[8]

### *1.3 Weicher Paternalismus*

Weicher Paternalismus hingegen wird als Gegenpol zum harten Paternalismus als Autonomie-respektierend eingestuft – zumindest in letzter Konsequenz. Wieder bezogen auf obiges Beispiel von Mill bedeutet dies: der weiche Paternalist zielt nicht darauf ab, den Mann vom Überqueren der Brücke abzuhalten, weil er sich damit schaden könnte. Vielmehr will er den Mann unterstützen, seine wahren, autonomen Wünsche zu bewahren. Laut Mill ‚will der Mann nichts ins Wasser fallen'.[9] Es ist damit legitim, ihn aufzuhalten, um ihn über den Zustand der Brücke zu informieren, damit er auf dieser Basis eine wohlinformierte und damit autonome Entscheidung treffen kann – und zwar die Brücke zu überqueren oder nicht. Wie er sich autonom entscheidet, ist für den weichen Paternalisten nicht mehr von Belang.[10]

Weicher Paternalismus ist damit auf den Respekt für Autonomie gerichtet: paternalistische Eingriffe sind dann *legitim* bzw. sogar *gefordert*, wenn die Autonomie der Person nicht verletzt wird bzw. geschützt oder befördert wird.

Weicher Paternalismus wird damit – und das ist zentral für das folgende – meist angewandt, wenn ein Defizit in der aktualen Autonomie des Individuums vorliegt: wenn ihm, wie im vorliegenden Fall, wichtige Informationen fehlen, wenn er temporär aufgrund von Stress, Druck etc. nicht zur Entscheidung in der Lage ist oder seine Autonomie in sonstiger Weise beeinträchtigt wird. So gesehen sind, hier wieder an Beispielen aus der Medizin gut zu illustrieren, Informationspflichten oder Beratungsgespräche auch paternalistische Eingriffe, sogar wenn sie die Autonomie eines Individuums letztlich erhöhen.[11]

### *1.4 Rechtfertigungsfrage*

Für viele Liberale bzw. Libertäre wie Sunstein/Thaler ist mit der Unterscheidung zwischen *hart* und *weich* die Rechtfertigungsfrage entschieden. Weicher Paternalismus hat zum Ziel, die Autonomie eines Individuums zu respektieren oder gar zu fördern.

Damit scheint für viele Liberale klar, welche Art von Paternalismus legitim ist. Der weiche Paternalismus wird als eine gerechtfertigte Form anerkannt, weil er die Autonomie eines Individuums *respektiert*, erst *herstellt* oder – noch mehr – *fördert*. Insofern kann Paternalismus sogar in einen beträchtlichen Autonomiegewinn münden: wenn beispielsweise durch kompulsive Aufklärungsarbeit eine voll informierte – und damit augenscheinlich bessere – Entscheidung getroffen werden kann. *Weicher Paternalismus* ist also ein

[8] vgl. Beauchamp, Childress 2011

[9] vgl. Mill 1859

[10] vgl. Häyry 2002

[11] vgl. ebenda

Paternalismus unter liberalen Vorzeichen, der wesentlich weniger unproblematisch scheint als ein harter.

Sunstein/Thaler sowie weitere Vertreter eines Ansatzes der unten sog. *Beschränkten Rationalität*[12] gehen hier aber noch weiter. An einigen Stellen befürworten sie sogar eine Pflicht zu wohlwollenden, weichen Eingriffen, welche dem Paternalisierten immer noch eine Möglichkeit lassen, sich anders zu entscheiden. Da diese weich im eben genannten Sinne bleiben, seien sie legitim. Diese generelle Rechtfertigung möchte ich auf Basis einer Analyse verschiedener Ziele von paternalistischen Eingriffen im Folgenden in Frage stellen.

## 2. Beschränkte Rationalität und Paternalismus

Sunstein/Thaler beziehen ihre Paternalismusstrategie auf Untersuchungen aus der ökonomischen Forschung der sog. „Beschränkten Rationalität“[13]. Die dahinter stehenden Überlegungen sind folgende: Man weiß bzw. glaubt aus entsprechenden empirischen Untersuchungen zu wissen, dass die Entscheidungen von Menschen nicht vollständig rational bzw. häufig nicht genau festgelegt sind. So zeigt sich u.a., dass Menschen nicht gerade selten zirkuläre Präferenzen oder Inkonsistenzen besitzen. Zudem weiß man durch die Untersuchungen von Entscheidungsverhalten, dass sich Menschen in bestimmten Situationen – und auch wenn es um ihr Wohl geht – häufig noch keine feste Meinung gebildet haben. So lassen sie z.T. sehr stark von bestimmten Vorbedingungen und Umständen beeinflussen. Sie verletzen damit die normativen Anforderungen an die Rationalität der *rational-choice*-Theorie[14] aus der Ökonomie, insbesondere die *Vollständigkeit*, die *Konsistenz* oder die *Transitivität* von Präferenzen. Folgende Beispiele illustrieren dies[15]:

> *Default-Option*: Man kann bei der Präsentation von Optionen, z.B. A und B, eine bestimmte Standardeinstellung A als gegeben vorlegen und die Möglichkeit lassen, dass Menschen sich noch abweichend für B entscheiden können. Ein Staat kann z.B. als Standard festlegen, dass Menschen automatisch krankenversichert sind, aber auch zugleich anbieten, dass man per Antrag aus der Versicherung austreten kann. Untersuchungen zeigen, dass von den meisten Menschen der Standard bevorzugt wird, vermutlich weil gemutmaßt wird, dass der vorgegebene Standard besser ist als eine andere, optionale Lösung – oder weil es einfach aufwändiger ist, vom Standard abzuweichen. Für das Versicherungsbeispiel gilt also: die meisten Menschen bleiben versichert.
>
> *Framing Effekt*: Menschen lassen sich in ihrer Entscheidung stark davon beeinflussen, wie Fragen/Bedingungen formuliert sind (z.B. ist, wenn bei der Entscheidung für einen medizinischen Eingriff entscheidend, ob die Risiken in Sterbe- oder Überlebensrate ausgedrückt werden).
>
> *Endowment-Effekt*: Menschen verzichten weniger gern auf eine Sache, welche sie bereits besitzen, als auf eine, die sie noch nicht erhalten haben (auch wenn sie ihnen zusteht). Es macht beispielsweise einen Unterschied, wie viele Menschen sich für die Einzahlung in die Rentenversicherung entscheiden, ob Rentenbeiträge automatisch eingezogen werden oder sie nach Erhalt des Gehalts abgezogen werden.

Durch die genannten Effekte bleiben Menschen in ihren Entscheidungen hinter den Anforderungen der ökonomischen Rationalitätstheorie zurück. Aus ethischer Sicht kann

---

[12] vgl. Sunstein, Thaler 2003

[13] vgl. ebenda

[14] vgl. Hargreaves-Heap et. al. 1992

[15] vgl. Sunstein, Thaler 2003

ebenso auf Basis bestimmter normativer Autonomietheorien Ansprüche gefordert werden, dass der Mensch die relevanten Aspekte seines Lebens und seines Wohls weitgehend selbst bestimmt und nicht durch Manipulation von außen.[16] Dies zeigt sich besonders, wenn man die Autonomiedefinition stark mit dem Begriff der Rationalität verschränkt und fordert, dass eine autonome Entscheidung zu einem starken Grad auch rational sein muss.[17]

Sunstein/Thaler nehmen diese Defizite im menschlichen rationalen Entscheiden zum Anlass, um eine, wie sie betonen, legitime Version eines Paternalismus zu konstruieren. Sie schlagen z.B. vor, die Speisenfolge in der Kantine so zu arrangieren, dass Menschen eher die gesunden Lebensmittel wie Salat und Obst bevorzugen. Ein anderes von ihnen empfohlenes Szenario besteht darin, dass Arbeitgeber in den USA eine Kranken- bzw. Rentenversicherung als Standardoption im Arbeitsvertrag festlegen. Die Betroffenen bekommen eine opt-out-Lösung, mit der sie sich auch gegen die Versicherung entscheiden können. Die Untersuchungswerte zeigen aber, wie eben gerade beschrieben, dass die meisten Menschen versichert bleiben werden.

Diese Form von Paternalismus scheint in der Regel sehr weich und eher unproblematisch. Sie ist libertär, weil Menschen sich immer anders entscheiden könnten. Indem man z.B. Informationen in bestimmter Weise formuliert oder die Entscheidungssituation anpasst, kann man Menschen dazu bringen, etwa „bessere“ Entscheidungen über ihre Zukunft zu treffen. Beispielsweise kann man mehr Menschen auf sanfte Art[18] dazu bringen in eine Renten- oder Krankenversicherung einzutreten.

Ebenso kann man – wie es häufig in medizinethischen Diskussionen und auch bei Sunstein/Thaler anklingt[19] – paternalistische Eingriffe dazu nutzen, Autonomiedefizite wie die eben genannten zu beheben, z.B. dadurch, dass man die Menschen besser informiert, Einsichten vermittelt oder Irrtümer im rationalen Entscheiden korrigiert. Man kann also Fehler, die Menschen in ihrer freien rationalen Entscheidung machen, durch paternalistische Regelungen ausmerzen.[20]

### *2.1 Eine Unterscheidung in den Zielen von paternalistischen Eingriffen*

Da es in den genannten Fällen von Paternalismus für einen – zumindest einigermaßen autonomen Menschen – immer die Möglichkeit gibt, sich anders zu entscheiden, werden Strategien dieser Art derzeit von den meisten liberalen Theoretikern als legitimer Paternalismus eingestuft. Ich möchte hier jedoch für ein differenzierteres Bild plädieren, um legitime Arten von weichem Paternalismus von – auch unter liberalen Gesichtspunkten – zweifelhaften Formen des weichen Paternalismus unterscheiden.

Als Basis für meine kritischen Anmerkungen will ich den eben gezeichneten weichen Paternalismus anhand der Zielsetzung in zwei grundlegende Kategorien zu unterscheiden:

> *(1) Wohl-orientierter weicher Paternalismus*: Autonomieunsicherheiten bzw. Defizite werden genutzt, um das Wohl von Menschen zu verbessern, ohne sie vollends ihrer Entscheidungsmöglichkeit zu berauben.
>
> *(2) Autonomie-orientierter weicher Paternalismus*: Autonomiedefizite sollen behoben werden. Die Förderungen bzw. Herstellung der Autonomie des Individuums ist das Ziel – z.B. durch Informationsgabe.

---

[16] vgl. Christman 2011

[17] vgl. Gutwald 2010

[18] Sunstein/Thaler sprechen hier von “nudge”, vgl. Sunstein C., Thaler R. 2008

[19] vgl. Beauchamp T.L., Childress J.F. 2001

[20] vgl. Wear 1983

Mein Argument ist nun, dass beide Formen dieses Paternalismus – im Gegensatz zu dem was Sunstein/Thaler et al behaupten – auf liberaler Grundlage nur eingeschränkt rechtfertigbar sind – allerdings aus verschiedenen Gründen und mit unterschiedlichen Grenzen. Mit anderen Worten: es kommt nicht nur darauf an, dass Autonomie in irgendeiner Form bzw. Definition gewahrt wird, sondern auch wie und wie weit die Maßnahmen gehen.

Es ist also, so meine Behauptung, nicht jeder weiche Paternalismus echt liberal. Umgekehrt ist meiner Ansicht nach kritisch, ob evtl. auch harter Paternalismus im Liberalismus Platz hat[21]. Das ist aber Gegenstand einer anderen Untersuchung, die hier nicht unternommen werden kann.

Meine Kritikpunkte bauen hier aufeinander auf. Zunächst möchte ich (1) aus liberaler Perspektive kritisch untersuchen, woraus sich Konsequenzen für (2) ergeben, welches seinerseits aus anderen Gründen im Liberalismus eingeschränkt werden sollte.

### *2.2 Wohl nicht ohne Autonomie denken*

Um (1) zu begründen, möchte ich auf den zentralen liberalen Wert der Freiheit und der Autonomie von Bürgern in einem entsprechenden Staat verweisen. Das oberste Ziel des liberalen Staates besteht darin, dem Bürger die Möglichkeit zu einem autonomen Leben zu geben und die entsprechenden Voraussetzungen zu sichern. Dies kann natürlich konkret in unterschiedliche Ausgestaltungen münden.

Im Lichte dieses liberalen Grundgedankens erscheint manche der von Sunstein/Thaler vorgeschlagenen Maßnahmen aus meiner Sicht Unbehagen hervorzurufen – zumindest bei genauerer Analyse. So sind Fälle denkbar, in denen die Unsicherheiten bzw. Defizite in der Autonomie in einem gewissen Sinn *ausgenutzt* werden, um das Wohlergehen von Menschen nach einem bestimmten Bild zu formen. Man weiß also darum, dass Menschen sich in einem bestimmten Setting eher für A anstatt für B entscheiden. Diesen Effekt kann man geschickt dafür nutzen, sie zu ihrem Besten zu „führen“, ohne dass die Menschen darum wissen. Solche Maßnahmen sind nicht automatisch manipulativ, bergen aber die Möglichkeit der Manipulation und Intransparenz, was sie aus meiner Sicht fragwürdig macht. So soll im liberalen Staat der mündige Bürger für sich selbst entscheiden – Freiheit von Manipulation und eine gewisse Transparenz sind hier Voraussetzung.

Im „Geiste“ entspricht diese Art von Paternalismus zudem viel mehr dem harten Paternalismus, der die Autonomie des Individuums offen zur Seite legt. Zwar wird die Autonomie nicht offen übergangen, kann jedoch subtil hintergangen werden. Eine andere mögliche Zielsetzung wäre, die Autonomie zu befördern oder herzustellen. Insofern entspricht die Ratio hinter den genannten Eingriffen nicht dem liberalen Grundgedanken, Autonomie soweit als irgend möglich zu respektieren.

Dass es in den von Sunstein/Thaler vorgeschlagenen Fällen immer noch eine opt-out-Lösung gibt, ist nur ein eingeschränkt plausibles Argument: die Erkenntnisse der beschränkten Rationalität zeigen gerade, dass die autonomen Menschen diese gerade nicht nutzen bzw. durch subtile Effekte davon weggesteuert werden, sie benutzen zu wollen. Damit werden bestimmte Optionen als weniger attraktiv ausgezeichnet, was eine Manipulation der autonomen Entscheidung impliziert.

Damit möchte ich im Übrigen nicht sagen, dass harter Paternalismus und wohlwollende Sorge für den Bürger gänzlich abzulehnen sind. Es kann gute Gründe geben, um manche Eingriffe durchaus zu befürworten: entweder aus gewichtigen Gründen der Grundversorgung der Bürger oder weil man die Voraussetzungen der Autonomie erst schaffen muss (etwa durch Bildung). Dies ist eine sehr feine Linie, die auf Basis einer komplexen normativen

[21] Was er bei bestimmten Maßnahmen zumindest in der Realität eines Sozialstaats hat, man denke an das Verbot harter Drogen oder an die Krankenversicherungspflicht in Deutschland.

Grundlage gezogen werden muss. Dies kann hier leider nicht weiter diskutiert werden. Diese Überlegungen genügen aber, so meine ich, um Zweifel an dem von Sunstein/Thaler als so unschuldig bezeichneten Paternalismus anzumelden. Meiner Ansicht nach reicht damit die Unterscheidung zwischen weichem und hartem Paternalismus nicht aus, um die Rechtfertigungsfrage vollends zu entscheiden.

### *2.3 Autonomie-Förderung*

Aber auch die unkritische Förderung von Autonomie kann in Frage gestellt werden. Verdeutlichen lässt sich dies an Maßnahmen des sog. kognitiven Enhancements, welche in den letzten Jahren in der Ethik kontrovers diskutiert wurden. Ich möchte diese Maßnahmen als Möglichkeiten der Autonomierweiterung interpretieren und fragen, ob Vertreter eines Paternalismus beschränkter Rationalität diese Möglichkeit empfehlen könnten bzw. sogar als wünschenswert betrachten müssten. Damit wird meiner These nach eine *slippery slope* eröffnet, so dass auch die Autonomieförderung aus liberaler Einsicht beschränkt werden soll. Dies gelingt, so mein Argument, wenn man grundlegende normative Begriffe des Liberalismus, insbesondere den der Autonomie, im Sinn eines Schwellenbegriffs interpretiert.

#### *2.3.1 Enhancement: Methoden und Auswirkungen*

Als Enhancement-Maßnahmen möchte ich solche medizinischen Interventionen verstehen, welche nicht auf das Heilen von Krankheiten gerichtet sind, sondern von Gesunden gewählt werden, um bestimmte kognitive Kapazitäten und Fähigkeiten des menschlichen Gehirns über das „normale“ Maß zu steigern, etwa die Merk- oder Konzentrationsfähigkeit, die Informationsaufnahme oder das Selbstbewusstsein allgemein.[22] Es kann davon gesprochen werden, dass solche Maßnahmen unter Umständen einen Autonomiegewinn nach sich ziehen: man kann mehr Informationen aufnehmen, ist konzentrierter, freier von Zweifel etc.[23]

In Bezug auf Eingriffe des Enhancements wird die Frage nach Paternalismus vor allem in der klassischen Weise des Verbots der Selbstschädigung gestellt:[24] Es scheinen aber bestimmte *weich* paternalistische Argumente, welche die Förderung von Autonomie zum Ziel haben, den Einsatz von Enhancement als legitim bzw. sogar wünschenswert zu befürworten. Dies ist so eben in Fällen, wenn die genannten Fähigkeiten gesteigert werden können.

Dies wirkt auf den ersten Blick kontraintuitiv, ergibt sich meiner Ansicht nach jedoch, wenn man den Gedanken des Autonomie-fördernden Paternalismus zu Ende denkt: Indem bestimmte Medikamente unsere Aufmerksamkeit oder Merkfähigkeit positiv beeinflussen, steigern sie auch Fähigkeiten und Möglichkeiten, die für den Grad und die Ausprägung der personalen Autonomie relevant sind. Die kritische Frage lautet hier: Wenn es also legitim ist, die Autonomie von Menschen durch Maßnahmen wie Korrektur von Rechenfehlern, Informationsgabe und Bildungsmaßnahmen zu steigern, wieso sollten wir dann hier stoppen und nicht noch den Schritt wagen, Autonomie medikamentös so zu verbessern, dass z.B. mehr Informationen aufgenommen werden, wir besser rechnen können etc.? Wie eben

---

[22] Die schwierig zu fassenden Begriffe dieser Definitionen, wie „normal“ oder „Krankheit“ können hier im Einzelnen nicht ausführlicher diskutiert werden. Für die gegenwärtige Diskussion genügt jedoch die Feststellung, dass Enhancement-Maßnahmen bestimmte, grundlegende kognitive Leistungen des menschlichen Gehirns verändern.

[23] Ich möchte hier nicht untersuchen, ob Enhancement an sich gerechtfertigt ist oder verboten werden soll. Was man aber zumindest behaupten kann, ist, dass Enhancement weitreichend in die menschliche Lebensführung eingreift und daher zumindest diskussionswürdig ist. Ich erkenne freilich an, dass es gewichtige Gründe gibt, am Enhancement an sich zu zweifeln, die ich hier aber nicht aufführen kann.

[24] vgl. Heilinger 2010

dargestellt gibt es der gegenwärtigen Paternalismusdiskussion inhärente Argumente, die dies als legitim zulassen bzw. die dies positiv bewerten.

Meine Behauptung ist nicht, dass der weiche Paternalismus, der Enhancement einsetzt, illegitim ist, weil sich eine eventuelle Illegitimität von Enhancement-Maßnahmen überträgt. Mein folgendes Argument ist vielmehr, dass *weitreichende* autonomiefördernde Maßnahmen *generell* illegitim sind. Das Enhancement ist das Extrembeispiel, an welchem sich zeigen lässt, warum man den autonomiefördernden Paternalismus beschränken soll.

### *2.3.2 Paternalismus des Enhancements*

Meiner Ansicht nach lässt sich die Kontroversität des Enhancement darauf zurückführen, dass es menschliche Fähigkeiten und die Bedingungen von Autonomie stark beeinflusst und weitreichend erweitert. Dies scheint umso problematischer, wenn jemand solche Maßnahmen für Dritte einsetzt, selbst wenn es um Autonomieerweiterung geht. Der Gedanke löst also per se großes Unbehagen aus.

Alle Maßnahmen, mit denen Dritte in die Autonomie von Menschen eingreifen, sollten jedoch ein ähnliches Unbehagen hervorrufen. Auch wenn sie letztlich die Autonomie eines Menschen steigern oder wieder herstellen, handelt es sich generell um Eingriffe in die Lebensführung eines selbst bestimmten Individuums. Hier sollte man auf die Grenzen achten. Diese Grenzen sind jedoch nicht immer klar zu ziehen. Daran liegt, dass der zentrale Begriff der Autonomie und ihre Voraussetzungen, aus liberaler Sicht unterschiedlich gedeutet werden können. Nur einer ist meiner Ansicht nach für die liberale Rechtfertigungsdiskussion des Paternalismus geeignet, nämlich der erste.

Die beiden Möglichkeiten, die u.a. auf Joel Feinberg zurückgehen, stellen sich wie folgt dar:

#### *2.3.2.1 Autonomie als Schwellenbegriff: „genug Autonomie ist genug“*

Autonomie kann als *Schwellenbegriff* aufgefasst werden, wie es häufig in der Medizinethik und Rechtsphilosophie der Fall ist. Es geht dabei um die Frage, wann einem Menschen die Kompetenz und v. a. das Recht zugeschrieben werden kann, eine Entscheidung für sich zu treffen.[25] Entscheidend ist, dass der Mensch damit in einem *hinlänglichen* Sinn Entscheidungen für sein Leben treffen kann. Ein Schwellenkonzept funktioniert also nach dem Motto alles-oder-nichts: Wenn der Zustand des Menschen eine gewisse Schwelle überschreitet, wird er in vollem Sinn als autonom angesehen. Dies gilt auch dann, wenn seine Entscheidungen unklug, irrational oder aus anderer Perspektive defizitär sind. Wichtig ist nur, dass sie seinen eigenen Wünschen entsprechen.[26]

Dieser Gedanke kann auf das Millsche Brückenbeispiel angewendet werden. Stellen wir uns vor, der Mann auf der Brücke würde sich trotz der Information über den Schaden an der Brücke entscheiden, über die Brücke zu gehen. Wenn wir die Schwelle der Autonomie an den Besitz dieser Information knüpfen, müssen wir ab diesem Zeitpunkt sagen, die Entscheidung sei autonom geschehen. Dies gilt auch, wenn der Mann darüber hinaus rationale Fehler macht, wie z. B. das Risiko zu gering einzuschätzen, es eilig zu haben, oder gar nicht weiter über die Folgen seines Handelns nachzudenken. Die Entscheidung, die Brücke zu überqueren, mag rational kritisierbar sein, kann aber nicht mehr als nicht-autonom eingestuft werden. Feinberg bemerkt in diesem Sinne: „In some contexts we may even want to permit choices that are quite substantially less than ‚fully voluntary' to qualify as ‚voluntary enough'".[27]

---

[25] vgl. Beauchamp, Childress 2001

[26] vgl. Gutwald 2010

[27] Feinberg 1989

Weicher Paternalismus hat hier durchaus einen legitimen Platz. Man sollte vor allem betrachten, wie weit wir die Autonomie eines Betroffenen unterstützen bzw. wiederherstellen sollen, wenn jemand in einer bestimmten Entscheidungssituation unter negativen Einflüssen leidet, die eine autonome Entscheidung behindern – und den Betreffenden unter die Schwelle drücken. Die restriktive, liberal geprägte Paternalismuskritik, welche John Stuart Mill folgt[28], würde meiner Ansicht nach in bestimmten Fällen sogar verpflichten, Autonomie (wieder) herzustellen. Das impliziert beispielsweise, jemand mit den grundlegenden Informationen zu versorgen, auch wenn er diese evtl. nicht haben möchte, damit er *überhaupt* eine autonome Entscheidung treffen kann. Ebenso ist nötig, den Menschen mit den grundlegenden materiellen und geistigen Voraussetzungen für eine autonome Entscheidung auszustatten – etwa durch Versorgung mit lebensnotwendigen Gütern und ein genügendes Maß an Bildung.

Jedoch sind dieser Art der Autonomie Grenzen gesetzt: Entscheidend für den Paternalisten ist hier, wann eine Person in ihrer Autonomie hinter dieser Schwelle zurückbleibt. In diesem Fall liegt *keine* Autonomie vor und ein weicher Paternalismus ist gerechtfertigt.

Diese Deutung von Autonomie als Schwellenbegriff hat den Vorteil, dass sie eine *prinzipielle* Grenze für Eingriffe setzt. Der Wert der so verstandenen Autonomie erfordert, Personen und deren Entscheidungen grundsätzlich zu respektieren und ihren Status als Wesen zu achten, die ihr Leben selbst bestimmen können. Dadurch wird eine private Sphäre garantiert, in der ein Mensch in Bezug auf sein eigenes Leben schalten und walten kann, wie er will – auch irrational, dumm oder irrtümlich. Dies entspricht der eingangs angesprochenen liberalen Idee, dass die Freiheit des Menschen grundlegend ist, und soweit als möglich respektiert werden muss, wenn nicht Dritte geschädigt werden.

Diese Position steht dem Enhancement generell kritisch gegenüber, auch wenn es sich auf die Autonomie richtet: Ziel ist es, den Menschen in die Lage zu versetzen, dass er selbstbestimmt entscheiden kann – darüber hinaus sollte er aber frei entscheiden können, wie und was er wählt. Dies kann auch beinhalten, eben nicht seine Fähigkeiten zu steigern, Fehler zu machen etc. Enhancement wäre hier also nur in dem sehr limitierten Bereich der Autonomieherstellung legitim.

#### *2.3.2.2 Autonomie als Ideal: „mehr ist besser“*

Man kann Autonomie allerdings auch graduell deuten. In diesem Sinn misst man die Fähigkeiten und Zustände eines Menschen an einem normativen Standard. Derjenige kann anhand dessen als mehr oder weniger autonom eingestuft werden. Dieses Verständnis von Autonomie ist insofern plausibel, als wir bei der Beurteilung von Entscheidungen häufig davon sprechen, eine Entscheidung sei mehr oder weniger X – z. B. klüger, dümmer, überlegter etc. So können wir von Menschen auch sagen, sie seien autonomer, weil sie ihre eigenen Wünsche besser verwirklichen als andere.

In der Diskussion um die o.g. *Bounded Rationality* wird Autonomie zudem mit Rationalität gleichgesetzt. Sunstein und Thaler beziehen sich, wie ich meine, häufig auf ein graduelles Konzept von Rationalität in Fällen, in denen sie annehmen, der Mensch solle in der Umsetzung seiner Präferenzen *maximal* rational sein. So beschreiben sie das Ideal eines perfekt rational agierenden Entscheiders mit vollständigen, transitiven und symmetrischen Präferenzen, dem wir in der Realität mehr oder weniger entsprechen können.

Autonomie in dieser Form ist also eng an das beschriebene Bild von Rationalität geknüpft. So kann man auch von mehr oder weniger autonomen Entscheidungen sprechen, gemessen an dem Standard für Rationalität, den man zugrunde legt.

Wiederum lässt sich der Brückenfall auch in diesem Sinn deuten: Wenn man vollständige Information als Kriterium für Autonomie annimmt, dann folgt, dass der Mann in Bezug auf

[28] vgl. Mill 1859

seine Entscheidung, die Brücke zu überqueren, *weniger* autonom ist, wenn ihm die Information über den Schaden an der Brücke fehlt.

Mit dieser graduellen Interpretation von Autonomie eröffnen sich hier zwei Probleme, die aus liberaler Sicht bedenklich sind, und daher Gründe liefern, um das graduelle Konzept als politisch-liberalen Begriff möglichst wenig einzusetzen.

Zum einen ergibt sich aus dem Millschen Liberalismus ein überzeugender Grund, warum wir Autonomie *nicht* graduell verstehen sollten. So ist es eine zentrale Forderung des Autonomieschutzes, dass jedem Menschen eine Sphäre zugestanden wird, in der er frei entscheiden kann und darf. Dieses Ziel ist aber durch die Auffassung von Autonomie als Schwelle wesentlich besser erreicht. Wenn wir Autonomie als graduell und perfektionistisch verstehen, werden viele Entscheidungen als defizitär gelten. Damit wird ein großer Raum für Paternalismus eröffnet, da jede Entscheidung, die offen für rationale Kritik ist, offen für einen potentiellen Eingriff ist, insbesondere wenn, wie die *Bounded Rationality*-Theorie behauptet, die meisten unserer Entscheidungen defizitär sind. Damit wird der Freiheitsraum, der im liberalen System garantiert werden soll, verkleinert.[29]

Zweitens wird die Engführung von Rationalität und Autonomie der philosophischen Idee von Autonomie nicht so gut gerecht wie das alternative Verständnis. Rationalität ist nur eine *Voraussetzung* von Autonomie. Das bedeutet aber nicht, dass sie die wichtigste Rolle spielen soll. Im Gegenteil, Rationalität kann, als graduelles Konzept verstanden, sogar für die Autonomie hinderlich sein. Stellen wir uns vor, neben uns stünde ein Roboter, der sich jedes Mal in unsere Handlungen einmischt, wann immer wir eine Entscheidung treffen, die nicht völlig unsere eigenen Präferenzen maximiert.[30] Auf diese Art mag die Rationalität einer Person verbessert werden, aber ihre Autonomie muss so nicht gefördert werden bzw. wird sogar darunter leiden. Zentral für Autonomie ist, wie Mill betont, die *Individualität* und *Selbstentwicklung* einer Person[31], d. h. die Führung eines *eigenen* Lebens. Dies kann und muss auch die Möglichkeit beinhalten, Fehler zu machen, aus diesen zu lernen bzw. die Konsequenzen daraus zu ziehen etc. Zu einer eigenen, authentischen Entscheidung gelangt man nämlich in manchen Fällen erst dadurch, dass man die Freiheit hat, sich zu irren, seine Fehler einzusehen und daraufhin, den eigenen Weg zu finden.

Gerade letztere Überlegung spricht meiner Ansicht nach dagegen, Autonomieförderung und insbesondere Enhancement als positiv zu bewerten. Folgt man dem „Spirit" der o.g. Argumente Sunstein/Thalers bzw. vieler Medizinethiker, scheint es durchaus gerechtfertigt, ja evtl. sogar gefordert, Enhancement zur Förderung von Autonomie einzusetzen. Sunstein/Thaler sprechen z.B. davon, Menschen zu „besseren" Entscheidungen zu führen. Autonomie bzw. die relevanten Fähigkeiten werden nicht als individuelle Selbstbestimmung betrachtet, sondern vielmehr als Ideal, dem die tatsächlichen Fähigkeiten und Möglichkeiten der Menschen möglichst entsprechen sollten. Wieso sollte aus dieser Perspektive der Gedanke fern liegen, dass die Fähigkeiten der Menschen sich generell steigern, um sich besser selbst bestimmen zu können? Wenn dieses Ziel durch Maßnahmen des kognitiven Enhancements besser zu erreichen ist, wieso sollten wir – angenommen die Vorteile überwiegen die Risiken – auf diese verzichten? Und sie auch nicht bei Menschen mit Defiziten einsetzen?

---

[29] Anzumerken ist, wie Markus Englerth es tut, dass andere Überlegungen ins Feld geführt werden können, die gegen einen Eingriff in die Autonomie sprechen, z. B. Überlegungen der Effizienz, das Vorhandensein milderer Mittel, oder der positiven Effekte, die bestimmte rationale „Fehler" haben können (vgl. Englerth 2007). Dies mag stimmen, kann das Problem aber nicht vollständig beheben, da dies keine prinzipielle Grenze festlegen kann, ab wann in das Leben von autonomen Personen nicht mehr eingegriffen werden darf.

[30] vgl. Arneson 1980

[31] vgl. Mill 1859

Obwohl Sunstein/Thalers Position aufgrund ihrer Betonung von menschlichen Fähigkeiten und Freiheit (es steht den Menschen ja immerhin noch frei, was sie wählen) von vielen liberal genannt werden kann, möchte ich dies aufgrund der eben genannten Argumente bezweifeln: zur Selbstbestimmung des Menschen gehört auch das Recht, Fehler zu machen bzw. selbst für sich herauszufinden, was das Beste für ihn ist. Durch den Einsatz von Enhancement und anderen vermeintlich autonomiefördernden Maßnahmen wird einem die Möglichkeit genommen, diese Art der Freiheit in Anspruch zu nehmen. Dem libertären Paternalismus steht aber nicht viel offen, um aus seiner libertären Perspektive diesen Maßnahmen Grenzen zu setzen. Daher halte ich diese Perspektive für fragwürdig.

## 3. Schlussfolgerung

Der liberale Respekt vor Autonomie verpflichtet uns also – obwohl dies Positionen wie die von Sunstein/Thaler nahe legen – nicht dazu, Maßnahmen des kognitiven Enhancements ebenso wie andere weitreichende Formen der Autonomieförderung aus weich paternalistischen Gründen zu befürworten oder gar zu befördern. Vielmehr sollte darauf geachtet werden, dass die Menschen in der Lage sind, ihr Leben selbst zu bestimmen inklusive Fehlern und Umwegen.

Wenn man sich die Frage nach Paternalismus in Bezug auf Enhancement auf eine andere Weise stellt als die übliche, stellt man fest, dass auch aus liberaler Sicht Vorsicht gegenüber dem weichen Paternalismus geboten ist: nicht alles, was man optimieren kann, muss auch verbessert werden – selbst wenn es sich um die Autonomie handelt. Damit möchte ich zwei prinzipielle Grenzen für den weichen Paternalismus vorschlagen:

A. Weicher Paternalismus soll möglichst nicht am Wohl des Menschen allein, sondern an dem Respekt seiner Autonomie und deren Wiederherstellung bzw. Beförderung orientiert sein.

B. Auch unter dem Autonomie-fördernden Paternalismus fallen aus liberaler Sicht kontroverse Eingriffe, die nicht schon gerechtfertigt sind, weil sie weich sind.

**Rebecca Gutwald**

Ludwig-Maximilians-Universität München, Fakultät für Philosophie, Wissenschaftstheorie und Religionswissenschaft, Lehrstuhl IV Prof. Nida-Rümelin (Staatsminister a.D.)
rebecca.gutwald@lrz.uni-muenchen.de

## Literatur

Arneson, R. J. 1980: „Mill Versus Paternalism“. In: *Ethics* 90 (4), S. 470–489.

Beauchamp, T.L; Childress, J.F 2001: *Principles of Biomedical Ethics*, Oxford: Oxford University Press.

Clarke, S. 2002: „A definition of paternalism“ in: *Critical Review of International Social and Political Philosophy* 5 (1), S. 81–91.

Christman, J. 2011: „Autonomy in Moral and Political Philosophy“, *The Stanford Encyclopedia of Philosophy* (Spring 2011 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/spr2011/entries/autonomy-moral/>.

Dworkin, G. 2010: „Paternalism“, *The Stanford Encyclopedia of Philosophy* (Summer 2010 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2010/entries/paternalism/>.

Englerth M., „Behaviorial Law and Economics – eine kritische Einführung“, in Engel, C. Englerth M. Lüdemann J. Spieker I. (Hrsg.) 2007: *Recht und Verhalten: Beiträge Zu Behavorial Law and Economics*, Tübingen: Mohr Siebeck GmbH KG.

Feinberg, J. 1989: *The Moral Limits of the Criminal Law: Volume 3: Harm to Self*, New York: Oxford University Press.

Gutwald, R. (2010): „Autonomie, Rationalität und Perfektionismus. Probleme des weichen Paternalismus im Rechtfertigungsmodell der Bounded Rationality“, in Fateh-Moghadam B., Vossenkuhl W., Sellmaier S. (Hrsg.) *Grenzen des Paternalismus*, Stuttgart: Kohlhammer, S. 94–125.

Häyry, H. 2002: *The Limits of Medical Paternalism*, London and New York: Routledge.

Heap, S.; Lyons, B.; Hollis, M.; Sugden, R.; Weale, A. 1992: *The Theory of Choice: A Critical Guide*, London: John Wiley & Sons.

Heilinger J. 2010: *Anthropologie und Ethik des Enhancements*, De Gruyter.

Mill, J.S 1859: *On liberty*, London: J. W. Parker and Son.

Sunstein C.; Thaler R. 2003: „Libertarian Paternalism is not an Oxymoron“ in: *University of Chicago Law Review*. 70, S. 1159–1202.

—, 2008: Nudge: *Improving Decisions About Health, Wealth and Happiness*, New Haven: Yale University Press.

VanDeVeer, D. 1986: *Paternalistic Intervention: The Moral Bounds of Benevolence*, Princeton: Princeton University Press.

Wear, S. 1983 „Patient autonomy, paternalism, and the conscientious physician“, in *Theoretical Medicine and Bioethics*, Vol 4, No. 3, 253-274.

# Zum Begriff des Kindeswohls: Ein liberaler Ansatz

Christoph Schickhardt

Der Begriff des Kindeswohls ist von großer sozialer Bedeutung. Er ist grundlegend für die elterliche Erziehung von Kindern sowie für die Regelung der Stellung des Kindes im Gesetz und den Umgang mit Kindern in der rechtlichen und sozialstaatlichen Praxis. Gleichzeitig stellt die inhaltliche Ausfüllung des Kindeswohlbegriffs eine enorme Herausforderung für die Ethik dar, insbesondere für eine liberal orientierte Ethik. Vom hier zugrunde gelegten liberalen Ansatz ausgehend wird argumentiert, dass Glück und personale Autonomie zentrale Bestandteile des Kindeswohls sein sollten. Es wird jeweils erörtert, was Glück bzw. personale Autonomie begrifflich bedeuten, wie sie sich als Kindeswohlgehalte moralisch begründen lassen und was notwendig oder förderlich dafür ist, dass sie sich in Kindern entwickeln. Ziel des vorliegenden Beitrags ist erstens eine Annäherung an die generelle Problematik des Kindeswohlbegriffs und zweitens die Unterbreitung eines Vorschlags zur inhaltlichen Bestimmung des Kindeswohlbegriffs.

## 1. Das Kindeswohl im deutschen Rechtssystem

Sowohl die herausragende soziale Bedeutung des Kindeswohlbegriffs als auch die Schwierigkeiten mit dem Kindeswohlbegriff für eine liberale sozialethische Position lassen sich anhand eines Blicks auf den Begriff des Kindeswohls im deutschen Rechtssystem veranschaulichen.

Das Bundesverfassungsgericht definiert das verfassungsrechtlich geschützte Elternrecht (GG Art. 6 Abs. 2) derart, dass das Kindeswohl die „oberste Richtschnur“ für die Stellung der Eltern gegenüber ihrem Kind darstellt (BVerfGE 72, 122/139) und dass das Elternrecht „wesentlich ein Recht im Interesse des Kindes“ ist (BVerfGE 72, 122/139; 75, 201/218). Im familienrechtlichen Teil des Bürgerlichen Gesetzbuchs (BGB), der Kinder betrifft, wird „das Kindeswohl“ etliche Male explizit erwähnt. Zusätzlich zu diesen Erwähnungen gibt es ein allgemeines Kindeswohlprinzip (§ 1697a BGB), das sich generell auf richterliche Entscheidungen bezieht, von denen Kinder betroffen sind. Das Kindeswohl bzw. die Gefährdung des Kindeswohls ist gesetzliche Grundlage für staatliche Eingriffe in das elterliche Sorgerecht, d.h. in die Beziehung zwischen Erziehungsberechtigten und dem Kind (§ 1666 BGB). Wenn z.B. staatliche Einrichtungen wie Jugendämter ein Kind aus seinem familiären Umfeld heraus und in Obhut nehmen, geschieht dies zur Vorbeugung von Kindeswohlgefährdungen, d.h. zum Schutze des Wohls des Kindes. Die Zahl derartiger Eingriffe in Deutschland ist in den letzten Jahren massiv angestiegen. Im Jahr 2011 gab es in Deutschland 38.000 Inobhutnahmen von Kindern durch den Staat (Statistisches Bundesamt).[1]

Angesichts der enormen Bedeutung des Kindeswohlbegriffs in der Rechtsordnung und Rechtspraxis wäre es zu erwarten, dass der Gesetzgeber den Begriff des Kindeswohls

[1] Die rechtssystematische, praktische und nicht zuletzt auch sozial-politische Bedeutung des Kindeswohlbegriffs wurde zuletzt auch im Urteil des Kölner Landgerichts zur Beschneidung deutlich; zum Kindeswohlbegriff im Zusammenhang der Beschneidungsproblematik siehe auch Schickhardt (2012b).

inhaltlich näher bestimmt und explizit klärt, worin konkret das Wohl von Kindern besteht bzw. was im sogenannten Interesse von Kindern liegt. Diese Erwartung wird allerdings enttäuscht. Das Kindeswohl ist ein – nicht nur im deutschen Gesetz – unbestimmter Rechtsbegriff.

Traditionell wurden von liberalen Positionen aus Bedenken und Einwände gegen eine (auch nur partielle) gesetzliche Festlegung des Kindeswohlbegriffs vorgebracht – wobei meines Erachtens ein einseitiges Verständnis des liberalen Werts der Freiheit als Basis diente. Die rechtspolitische Sichtweise, die bestimmten liberalen Bedenken und Einwänden zugrunde liegt, sowie die sich aus ihr ergebende Rolle des Kindeswohlbegriffs im Rechtssystem zeigen beispielhaft die Schwierigkeiten, die der Kindeswohlbegriff für den politischen Liberalismus birgt: Einerseits scheint vom Ideal des liberalen Rechtsstaates die Forderung ableitbar, dass sich der Staat aus der sensiblen Sphäre der Familie und der elterlichen Kindererziehung grundsätzlich heraushalten soll. Da eine inhaltliche Bestimmung oder Teilfestlegung des Kindeswohls darauf hinauslaufen würde, den Eltern mit staatlicher Autorität und potenzieller staatlicher Gewalt direkt oder indirekt (durch die Bindung des Elternrechts an das Kindeswohl) ein Erziehungsziel für ihr Kind vorzugeben, erscheint gemäß liberalen Idealen eine solche inhaltliche Bestimmung als inakzeptabel.

Andererseits anerkennt der Liberalismus jedoch auch generell den Wert der Rechtssicherheit und des Schutzes der Bürger vor staatlicher Willkür, wozu eben die Bestimmtheit von Rechtsbegriffen wesentlich beiträgt. Der Liberalismus hat speziell aus den Erfahrungen mit totalitären und autoritären Regimen und Staatssystemen die Lehre gezogen, dass sich ein vager Kindeswohlbegriff leicht durch den Staat missbrauchen lässt, z.B. um politisch oder sozial unliebsamen oder unbequemen Eltern unter Berufung auf ein rechtlich unbestimmtes Kindeswohl ihr Kind wegzunehmen und es der Erziehung und Obhut linientreuer Adoptiveltern oder staatlicher Einrichtungen zu überantworten. Die Missbrauchsgefahr durch den Staat ist umso größer, je unbestimmter und vager der Rechtsbegriff des Kindeswohls ist.

Es kommt somit in Teilen des liberalen Denkens zu der merkwürdigen und problematischen Konstellation, dass der Staat sich einerseits in möglichst großem Umfang aus der inhaltlichen Bestimmung des Kindeswohlbegriffs heraushalten soll, andererseits aber gerade die rechtliche Unbestimmtheit des Kindeswohlbegriffs als unerwünscht und bedrohlich wahrgenommen wird.

Eine völlige Aufgabe des Kindeswohlbegriffs stellt jedoch auch keine akzeptable Lösung für diese Problematik dar. Die Funktion, die dem Kindeswohlbegriff zukommt, ist grundsätzlich unabdingbar. Diese Funktion besteht darin, aus der Perspektive Dritter einen Ersatzwillen oder ein Hilfskonstrukt für Kinder zu formulieren, die noch nicht kompetent und mündig bestimmen und erkennen können, was in ihrem Interesse liegt, was gut (bzw. schädlich) für sie ist, und die sich nicht eigenverantwortlich selbst bestimmen können. Wir ziehen einem Kleinkind im Winter eine wärmende Jacke an, geben ihm gesundes Essen und passen im Straßenverkehr auf es auf, weil wir glauben, dass dies im Interesse des Kindes liegt und dass das Kind in den fraglichen Angelegenheiten nicht ausreichend mündig und kompetent für sich selbst sorgen kann.

Ein Verzicht auf den Begriff des Kindeswohls oder auf ein entsprechendes begriffliches Konstrukt kommt auch im gesetzlichen Kontext nicht infrage, solange man anerkennt, dass Kinder in ihrem Schicksal und Wohlergehen nicht voll umfänglich und bis zur letzten Konsequenz ihren Erziehungsberechtigten ausgeliefert sein sollten.

Stark vereinfachend und abstrahierend lassen sich drei verschiedene zeitliche Dimensionen unterscheiden, auf die sich das Kindeswohl beziehen kann: Ein Kind kann erstens bezüglich seiner in naher Zukunft liegenden Interessen inkompetent oder unmündig sein, zweitens bezüglich seiner mittelfristigen Interessen, z.B. wenn es bei einem Sechsjährigen darum geht,

was für ihn in drei Jahren gut sein wird, und drittens bezüglich seiner langfristigen Interessen, wenn es z.B. um die Frage geht, was hinsichtlich eines sechsjährigen Kindes getan (oder unterlassen) werden sollte, um das zukünftige Wohl des „Kindes" bzw. der Person zu fördern, in die sich das Kind entwickeln wird, z.B. im Alter von 16 oder 22 Jahren.

## 2. Glück und personale Autonomie als Bestandteile des Kindeswohls

Es stellt sich nun die Frage, worin das Kindeswohl konkret bestehen sollte. Von einem liberalen Standpunkt aus möchte ich zwei unterschiedliche Elemente als grundlegende Bestandteile und Güter des Kindeswohls vorschlagen: erstens *Glück* im Sinne von subjektivem Wohlfühlen und zweitens *personale Autonomie*.

Zu Glück und personaler Autonomie als mögliche Bestandteile des Kindeswohls stellen sich jeweils drei Fragen: a) Worin bestehen sie, was ist mit ihnen begrifflich gemeint? b) Aus welchen moralischen oder normativen Gründen sollten sie zentrale Bestandteile des Kindeswohls sein? c) Was ist faktisch, sozial und entwicklungspsychologisch notwendig und förderlich dafür, dass ein Kind glücklich ist und personale Autonomie entwickelt?

Versuchen wir in der gebotenen Kürze zunächst, Antworten auf die drei genannten Fragen mit Blick auf das Glück als Inhalt des Kindeswohls zu skizzieren. Mit dem Begriff des Glücks (a) ist das subjektive Glück, das subjektive Gefühl des Wohlbehagens, die persönlich empfundene Glückseligkeit (happiness) gemeint. Der Begriff ist gemäß der liberalen Tradition nicht objektiv bestimmt und lässt es grundsätzlich offen, worin das Glück einer konkreten Person im Einzelnen besteht. Mit dem Begriff des Glücks soll kein Modell eines definierten glücklichen Lebens vorgegeben werden. Jeder soll auf seine eigene Weise glücklich sein oder werden. Dies gilt im Grundsatz auch schon für sehr junge Kinder: Für ein bestimmtes Kind mag z.B. das Hören von Musik Ursache oder Quelle von Glücksgefühlen sein, für ein anderes nicht. Dem Glück entgegengesetzt sind Gefühle der Unlust wie Schmerzen, Angst, Einsamkeit, Hunger oder Kälte. Was ein Kind glücklich (bzw. unglücklich) macht, hängt also von dem individuellen Kind ab und kann auch schon bei sehr jungen Kindern zumindest teilweise anhand der nicht-verbalen Expressivität eines Kindes, z.B. anhand von Schreien als Ausdruck des Unbehagens, erkannt werden.

Was ein individuelles Kind im Einzelnen glücklich macht, hängt grundsätzlich subjektiv vom Kind ab. Dagegen handelt es sich bei der These, dass das derart offen gelassene individuelle Glück zentraler Bestandteil des Kindeswohls sein sollte, um eine objektive Bestimmung des Kindeswohls aus der Perspektive Dritter mit einem „objektiven" bzw. allgemeinen moralischen Geltungsanspruch. Was die moralisch-normative Begründung dieses Anspruchs angeht (b), so kann hier nur in äußerster Kürze darauf verwiesen werden, dass es erstens intuitiv naheliegend ist, dass individuelles Glück für Kinder und Menschen allgemein grundsätzlich ein Gut darstellt. Es ist demnach gut und wünschenswert, dass ein Kind glücklich ist, und schlecht und zu vermeiden, dass ein Kind unglücklich ist. Zweitens stellt das subjektive Glück nicht allein für eine spezifische, sondern für unterschiedliche moralphilosophische Ausrichtungen ein anzustrebendes Gut bzw. einen Wert dar. Drittens lässt sich gerade bei moralisch unreifen Kleinkindern, die sich nicht kompetent für andere Güter entscheiden können, mit Verweis auf die Beweislast und auf eine gewisse „natürliche" Tendenz im Kinde argumentieren: Es bringt eine geringere Beweis- oder Rechtfertigungslast mit sich, wenn ein Kind glücklich ist und wenn man das Glück des Kindes anstrebt und verfolgt; dagegen ist die moralische Beweis- und Rechtfertigungslast größer, wenn mit einem Kind in einer bestimmten (erzieherischen) Weise umgegangen wird, die einen anderen Wert als das Glück des Kindes verfolgt, z.B. eine besonders hohe sportliche oder intellektuelle

Leistungsfähigkeit, und die das Kind unglücklich macht, z.B. weil es zu hartem Training oder Lernen gezwungen wird.[2]

Im Anschluss an die begriffliche Bestimmung des Glücks und an die Nennung einiger moralischer Gründe für das Glück als Kindeswohlgehalt stellt sich jetzt die Frage, was Kinder und Menschen brauchen, um glücklich zu sein und um ihr persönliches Glück möglichst umfangreich zu erreichen und zu verwirklichen (c). Zur Beantwortung dieser Frage kann auf eine Kombination und Integration verschiedener Strategien zurückgegriffen werden. Je nach Entwicklungsstand kann zum einen das Kind gefragt werden oder es können Rückschlüsse aus der kindlichen Expressivität und Individualität gezogen werden. Zum anderen müssen die Mängel des Kindes an Reife und Kenntnissen bezüglich der Frage, was es als Individuum – insbesondere mittelfristig und langfristig – benötigt bzw. benötigen wird, um sein persönliches Glück zu finden, mit einem Rückgriff auf Theorien von objektiven Gütern kompensiert werden. Dafür kommt z.B. die Theorie der Grundgüter in Betracht, die John Rawls für die hypothetische Entscheidungssituation der Menschen im Urzustand und hinter dem Schleier des Nichtwissens entwirft (Rawls 1979 S. 83f., S. 112ff.). Derart verstandene Grundgüter geben keinem Menschen ein definiertes Modell des glücklichen Lebens vor, aber sind vernünftigerweise und erwartungsgemäß für unterschiedliche persönliche und individuelle Lebensgestaltungen und das Streben nach Glück eines Jeden nützlich und hilfreich. Dies gilt z.B. für Gesundheit, Rechte, grundlegende Fähigkeiten, Bildung, Selbstrespekt, Chancen und Optionen.

Was nun personale Autonomie als zweiten zentralen Bestandteil des Kindeswohls betrifft, so ist auch hier zunächst die begriffliche Bedeutung zu klären, d.h. die Frage, was mit dem Begriff der personalen Autonomie überhaupt gemeint ist (a). Wenig überraschend tun sich Philosophen schwer damit, sich auf eine allseits akzeptierte Bedeutung des Begriffs der personalen Autonomie zu einigen, sodass es diesbezüglich eine rege philosophische Debatte gibt. Nichtsdestotrotz lässt sich grundsätzlich – und im Gegensatz zum subjektiven Glücksbegriff – von personaler Autonomie ein allgemeiner und objektiver Begriff inhaltlich definieren, der auf verschiedene Menschen angewandt werden kann – ähnlich wie z.B. der Begriff der körperlichen Fitness: Auch von körperlicher Fitness haben wir nicht exakt eine genau definierte Auffassung, und dennoch glauben wir, dass wir uns prinzipiell von Fitness einen objektiven Begriff bilden können, sodass der Wahrheitsgehalt der Aussage, ob bzw. wie sehr eine Person fit ist, nicht von der subjektiven Meinung dieser Person abhängt, sondern davon, wie sehr die Person objektive Eigenschaften aufweist, die die Bedeutung des Begriffs der Fitness ausmachen.

Ohne im Detail auf die komplexe Debatte über die begriffliche Bedeutung personaler Autonomie eingehen zu wollen, sei hier im Sinne einer ersten Annährung Folgendes angeführt: Personal autonom ist derjenige, der sich ungefähr der Art seiner Handlungen, seines Lebens und seines Wertesystems bewusst ist und diese in einem gewissen Umfang reflektiert und selbst gestaltet. Einen nur sehr geringen Grad an personaler Autonomie weist auf, wer sich in seinem Denken, Empfinden, Handeln oder in seinem Wertesystem nicht wiedererkennt oder wer darin von (unbewussten) Dynamiken und Kräften beherrscht und „fremdbestimmt“ wird, z.B. von Manipulationen, Indoktrinationen, Neurosen oder gewissen fundamentalen persönlichen Abhängigkeiten und Schwächen.

---

[2] Auch die personale Autonomie, die im vorliegenden Aufsatz als zweiter zentraler Bestandteil des Kindeswohls postuliert wird, steht bezüglich ihres Verhältnisses zum Glück unter besonderer Beobachtung: Ist die Förderung und Entwicklung von personaler Autonomie in einem Kind für das Glück des Kindes abträglich (*trade-off*), gleichgültig oder sogar förderlich (*win-win*)? Wenn die Förderung personaler Autonomie dem Glück abträglich ist, muss sie besonders gut begründet werden und es stellt sich die Frage, in welchem Umfang personale Autonomie auf Kosten des Kindesglücks gefördert werden darf und soll. Generell handelt es sich hier um die Frage der Verhältnisse, gegenseitigen Auswirkungen und der Gewichtung unterschiedlicher Bestandteile des Kindeswohls.

Hinsichtlich der moralischen Begründung der Forderung, dass personale Autonomie ein zentraler Bestandteil des Kindeswohls sein sollte (b), stellt sich auf einer grundsätzlichen Ebene die Frage, ob personale Autonomie einen intrinsischen oder einen extrinsischen Wert darstellt oder beides. Als intrinsischer Wert ist personale Autonomie ein Wert an sich, der um seiner selbst willen erstrebens- und wünschenswert bzw. „gut" ist. Personale Autonomie als extrinsischen Wert anzusehen bedeutet hingegen, dass der Wert der Autonomie abgeleitet ist und von den wünschenswerten und „guten" Konsequenzen herrührt, die durch personale Autonomie ermöglicht oder gezeitigt werden. Als derartige wünschenswerte und wertstiftende Konsequenzen kommen z.B. Selbstrespekt oder soziale Selbständigkeit, aber auch Glück im obigen Sinne infrage. Auf der Grundlage und im Rahmen eines ethischen Liberalismus ist es prinzipiell ein Wert an sich, dass Kinder und Erwachsene personale Autonomie entwickeln und besitzen und ihre Wertesysteme und Vorstellungen vom guten Leben in personaler Autonomie reflektieren und gestalten. Dem Wert der personalen Autonomie entgegengesetzt ist es, wenn Menschen Wertüberzeugungen und einem bestimmten Modell vom guten Leben nachgehen, die ihnen von Dritten autoritär und ohne Raum für eigenständiges kritisches Bewusstsein und Gestalten vorgegeben oder oktroyiert wurden.

Einem Kritiker, der in dem Verständnis von personaler Autonomie als einem intrinsischen Wert einen (allzu) perfektionistischen Liberalismus sieht, mit dem im Namen einer vermeintlichen Freiheit Kindern, Jugendlichen und Bürgern generell ein bestimmter Wert, nämlich personale Autonomie, aufgezwungen wird, kann Folgendes entgegnet werden: Es ist ein moralisch-axiologischer Unterschied, ob sich ein junger Erwachsener, der im Rahmen seines Heranwachsens einen gewissen Grad an personaler Autonomie entwickelt hat, in Ausübung seiner Autonomie, z.B. unter Rückgriff auf seine erworbenen Fähigkeiten zu selbständigem Denken und Handeln, dafür entscheidet, sein Leben einem bestimmten Wert zu widmen, z.B. dem finanziellen Erfolg oder der Kunst; oder ob er einen bestimmten Wert oder eine bestimmte Lebensform verfolgt, weil er seit seiner Kindheit durch eine einseitige Erziehung oder gar durch Indoktrination, die der Ausbildung personaler Autonomie entgegengesetzt waren, darauf festgelegt wurde. Analog ist es ja auch ein Unterschied, ob jemand Bauer oder Philosoph wird, weil er es aufgrund einer autoritären und unfreien Staatsordnung werden muss oder ob er sich ohne staatlichen Zwang für einen dieser Berufe entscheidet.

Personale Autonomie als Bestandteil des Kindeswohls erscheint als die beste Lösung für die Problematik, welche durch die (negative) liberale Kernthese entsteht, dass es nicht gut, sondern vielmehr zu vermeiden ist, dass ein Mensch deshalb eine bestimmte Lebensform ergreift, weil diese ihm vorgegeben wurde ohne die Möglichkeit, sie selbständig zu überdenken und gegebenenfalls eine alternative Lebensform zu konzipieren und zu verfolgen.[3]

Wenn man anerkennt, dass personale Autonomie zu Recht ein zentraler Bestandteil des Kindeswohls sein sollte, dann stellt sich nun die Frage, wodurch und wie die Entwicklung und Ausbildung personaler Autonomie in Kindern ermöglicht und gefördert wird (c). Damit ein Kind Autonomie entwickelt und ein- und ausübt, ist es u.a. nützlich und förderlich, dass das Kind zu eigenständigem Denken und Handeln ermutigt wird, dass es im kleinen und größeren Rahmen die Möglichkeit hat, Entscheidungen zu treffen und Optionen auszuwählen und auszuprobieren, dass es ein angemessenes Selbstbewusstsein und Bewusstsein der

[3] Diese These zu Wert und Funktion der personalen Autonomie knüpft an die in der philosophischen Debatte um Liberalismus, Perfektionismus, Autonomie und eine „offene Zukunft" (Feinberg 1980) präsente Position an, der zufolge personale Autonomie kein Wert wie jeder andere ist. Ob personale Autonomie deshalb als *ein* Wert einer eigenen *Kategorie* von Werten, der Werte zweiter Ordnung, einzustufen ist, wie von Colburn (2010) systematisch vorgeschlagen und von Meyer (2011) kritisiert, kann hier offen gelassen werden.

eigenen Fähigkeiten und Wirksamkeit erlangt und dass es – mit Wilhelm von Humboldt (2006 S. 22) gesprochen – mit mannigfaltigen gesellschaftlichen Situationen, Anreizen und Modellen konfrontiert wird, z.B. mit verschiedenen Charakteren, Berufen und Lebenswegen, und nicht in gesellschaftlicher Einseitigkeit und Einöde oder unter erdrückendem Konformitätsdruck aufwächst.

Abschließend ist anzumerken, dass die vorliegenden Überlegungen nur einen kleinen Diskussionsbeitrag zur Frage der inhaltlichen Bestimmung des Kindeswohlbegriffs bieten können und dass viele andere komplexe Fragestellungen rund um das Kindeswohl gänzlich unerörtert bleiben. Dies gilt beispielsweise auch für die Frage, ob und in welchem Umfang der widerstrebende Wille eines Kindes berücksichtigt werden muss, wenn es um das Wohl des Kindes geht. Im Sinne des hier vertretenen Ansatzes müssen der aktuelle Wille oder Unwille bzw. die aktuelle Gefühlslage und eventuelle Unlust eines Kindes berücksichtigt und in Rechnung gestellt werden, wenn es darum geht, etwas für das in mittelfristiger oder später Zukunft liegende Kindeswohl zu tun (oder zu unterlassen). Das Glück der Kindheit – als ein zentraler Kindeswohlbestandteil – darf nicht einfach dem späteren Glück im Alter der Reife geopfert werden.[4]

**Christoph Schickhardt**

Otto-Friedrich-Universität Bamberg
www.christoph-schickhardt.de
mail@christoph-schickhardt.de

## Literatur

Bundesverfassungsgericht (BVerfGE) 1951ff.: *Entscheidungen des Bundesverfassungsgerichts*, Hrg. von den Mitgliedern des Bundesverfassungsgerichts, Tübingen: Mohr.

Colburn, B. 2010: "Anti-Perfectionisms and autonomy", Analysis 70, 246-256.

Feinberg, J. 1980: "The child's right to an open future", in W. Aiken und H. Lafollette (Hrg.): *Children's rights, parental authority, and State Power*, 1980, Totowa: Rowman.

Humboldt, W. v. 2006: *Ideen zu einem Versuch, die Grenzen der Wirksamkeit des Staates zu bestimmen*, Stuttgart: Reclam.

Meyer, K. 2011: "Political liberalism and the value of autonomy", XXII. Deutscher Kongress für Philosophie, 11.-15. September 2011, München.

Rawls, J. 1979: *Eine Theorie der Gerechtigkeit*, Frankfurt am Main: Suhrkamp.

Schickhardt, C. 2012a: *Kinderethik. Der moralische Status und die Rechte der Kinder*, Münster: Mentis.

– 2012b: „Ein Kompromissvorschlag. Die Beschneidung lässt sich nicht nach Recht oder Unrecht beurteilen. Deswegen sollte man sie erst nur symbolisch vollziehen", in: Frankfurter Allgemeine Sonntagszeitung, 12.08.2012, S. 11.

Statistisches Bundesamt: http://de.statista.com/statistik/daten/studie/12982/umfrage/inobhutnahmen-minderjaehriger-durch-jugendaemter/ (14.01.2013).

[4] Die Problematik, ob und in welchem Umfang das Wohl eines Kindes *gegen* den (völlig oder teilweise inkompetenten) Willen des betroffenen Kindes geschützt oder verwirklicht werden darf oder soll, lässt sich dem Themengebiet des Paternalismus gegenüber Kindern zuordnen. Die Problematik der Gewichtung (und „Verrechnung") des Wohls eines Kindes in verschiedenen Lebensphasen, also z.B. die Gewichtung des aktuellen Wohls gegenüber dem langfristigen, ist eine weitere wesentliche und komplexe Fragestellung rund um das Kindeswohl; zu ihr und zum Paternalismus gegenüber Kindern siehe auch Schickhardt (2012a).

# Erbschaftssteuern, Obduktionen und die postmortale Konfiszierung von Organen

Christoph Schmidt-Petri

Ich werde in diesem Text folgendermaßen argumentieren: Wenn wir Erbschaftssteuern als legitim ansehen und auch solche Obduktionen als legitim ansehen, die nicht aufgrund des erklärten Willen des Verstorbenen durchgeführt werden, dann sollten wir auch eine Konfiszierungslösung für Organe von Hirntoten befürworten. Ich erörtere zahlreiche Einwände, die die jeweiligen Analogien anzweifeln aber nicht stichhaltig sind. Meine Argumentation gilt jedoch nur, wenn wir das Hirntodkriterium als Todeskriterium akzeptieren oder zumindest Organtransplantationen als grundsätzlich zulässig ansehen.

## 1. Einleitung

### *1.1 Zustimmung und Widerspruch zur Organspende*

In den vergangenen Jahren wurde in Deutschland heftig um die Neuregelung der Organspende gestritten. Das Transplantationsgesetz von 1997 sah eine sogenannte ‚erweiterte Zustimmungslösung' vor: Organe können Verstorbenen nur dann explantiert (und danach transplantiert) werden, wenn eine explizite Zustimmung erfolgt ist. Diese kann entweder vom Verstorbenen selbst stammen, etwa in Form eines entsprechend ausgefüllten Organspendeausweises, einer mündlichen oder sonstigen Erklärung oder in Ermangelung einer solchen Erklärung, von seinen Angehörigen, die sich bei ihrer eigenen Entscheidung zuerst noch mit dem sogenannten „mutmaßlichen Willen" des Verstorbenen befassen müssen.[1] Da mit diesem Verfahren anscheinend weder alle Personen versorgt werden können, bei denen eine Organtransplantation medizinisch indiziert ist, noch alle Personen zum Spenden bewegt werden können, die grundsätzlich einer Organspende positiv gegenüberzustehen scheinen,[2] wird inzwischen vielerorts eine Widerspruchslösung favorisiert, wie sie beispielsweise in Österreich praktiziert wird, sich in der deutschen Debatte aber nicht durchsetzen konnte: Organe sollen auch dann entnommen werden können, wenn der Verstorbene einer Organentnahme zwar nicht explizit zugestimmt, ihr aber zumindest nicht explizit widersprochen hat; bei einer ‚erweiterten' Regelung wäre auch hier der Wille der Angehörigen zu berücksichtigen.

Aus zwei Gründen ist es sozialpolitisch von sehr großer Bedeutung, welche dieser beiden Regelungen favorisiert wird. Erstens legen sie den potentiellen Organspendern unterschiedliche Handlungslasten auf. Da Meinungsbildung und anschließende Meinungsäußerung – völlig unabhängig davon, wie sie ausfallen mag – stets mehr Aufwand bedeutet als Untätigkeit, ist davon auszugehen, dass nicht jeder Bürger seinen tatsächlichen

[1] Auf dieses Verfahren gehe ich hier nicht weiter ein, obwohl ich es für sehr problematisch halte (Vgl. Schmidt-Petri 2012).

[2] Wie sich in Umfragen wiederholt gezeigt hat. Die Bundeszentrale für gesundheitliche Aufklärung (BZgA) nennt eine repräsentative Umfrage aus dem Jahr 2010, bei der „74 Prozent der 14- bis 75-Jährigen bereit [waren], nach ihrem Tod zu spenden. Doch nur 25 Prozent der Befragten sind in Besitz eines Organspendeausweises", siehe http://www.organspende-info.de/information/gesetz-und-studien/aktuelles, letzter Zugriff: 31.1.2013.

Willen erklären wird oder gar eine explizite Meinung zur Problematik ausbilden wird. Dies aber hat *ceteris paribus* zur Konsequenz, dass bei einer Widerspruchslösung das Organaufkommen höher sein dürfte als bei einer Zustimmungslösung. Im Falle von Untätigkeit werden so nämlich die Organe gespendet, wohingegen Untätigkeit bei einer Zustimmungslösung dazu führt, dass die Organe mit ins Grab genommen werden.[3]

*Ceteris* dürften aber kaum *paribus* sein, denn noch wichtiger ist, und dies ist der zweite Punkt, dass jegliche gesetzliche Regelung eine mehr oder weniger starke Setzung einer Handlungsnorm bedeutet. Eine Zustimmungslösung kann beispielsweise als Signal gedeutet werden, dass die Organspende eine Entscheidung von einer solchen Tragweite ist, dass zur Wahrung einer adäquaten und gültigen Zustimmung eine aktive und damit bewusste Handlung des Bürgers gefordert ist, bei der u.U. sogar (ähnlich der derzeit gültigen Regelung beim Schwangerschaftsabbruch) eine gewisse Aufklärungspflicht seitens des staatlichen Gemeinwesens besteht. Eine Widerspruchslösung hingegen signalisiert tendenziell, glaube ich, dass die Entscheidung nicht von solcher Tragweite ist. Einer potentiellen Organexplantation wird damit weniger Bedeutung beigemessen. Vielleicht signalisiert eine Widerspruchslösung auch, dass jemand eigentlich nur in ungewöhnlichen Ausnahmefällen das Widerspruchsrecht überhaupt wahrnehmen wollen dürfte – ähnlich der Verweigerung der Annahme von lebensrettenden Bluttransfusionen, wie sie z.B. von den Zeugen Jehovas praktiziert wird.

Dieser zweite Grund verstärkt also den ersten noch. Einige der Personen, die bei der Zustimmungslösung nicht zustimmen würden, und zwar aus Überzeugung und nicht aufgrund des Hangs zur Untätigkeit, werden bei einer Widerspruchslösung dennoch nicht widersprechen. Denn sie werden alleine aufgrund der Einführung der Widerspruchslösung ihre Meinungen ändern. Die Widerspruchslösung signalisiert ihnen nämlich mehr oder weniger implizit, dass in ihrer politischen Gemeinschaft eigentlich die Organspende als Regelfall angesehen wird, sie also zur Minderheit gehören.

Es ist daher eine Illusion zu glauben, es könnte eine Regelung geben, die gewissermaßen moralisch neutral *keinen* moralischen Druck aufbaut. Zu entscheiden ist nur, in welche Richtung der Druck gehen soll und viel wichtiger, wie stark er sein soll. Eine Widerspruchslösung baut deutlich mehr Druck zur Organspende auf als eine Zustimmungslösung und dürfte also weitaus höhere Organentnahmequoten mit sich bringen als eine Zustimmungslösung.

Die lange Zeit politisch möglich erscheinende aber letztlich für Deutschland ebenfalls verworfene ‚Erklärungslösung', bei der eine Zustimmung zur Spende noch erforderlich gewesen wäre, eine wie auch immer lautende Erklärung aber verbindlich hätte eingefordert werden können, hätte den Druck in Richtung Organspende verstärkt. Die nun beschlossene Regelung, die sogenannte ‚Entscheidungslösung', bei der man sich überraschenderweise weder erklären noch entscheiden muss, sich nun aber entscheiden *soll* und dazu persönlich aufgefordert wird, entspricht im Wesentlichen der seit 1997 geltenden erweiterten Zustimmungslösung. Sie soll aber, so scheint es mir, langfristig dazu beitragen, die freiwillige Organspende als Normalfall anzusehen und könnte sogar ohne Namensänderung zu einer Erklärungslösung mutieren.[4]

### *1.2 Meine Thesen*

Allen bisher diskutierten Ansätzen liegt die Annahme zugrunde, dass ein Mensch über seinen Körper auch über seinen Tod hinaus verfügen kann. In diesem Text möchte ich diese Annahme hinterfragen und letztlich ablehnen. Ich werde mich für die vermutlich als grotesk

[3] Für einige Zahlen hierzu vgl. Johnston und Goldstein 2003.

[4] Welche dann, wenn man noch weiter spekulieren möchte, wiederum ohne Namensänderung zu einer Widerspruchsregelung mutieren könnte.

anmutende *Konfiszierungslösung* aussprechen. Ich werde also dafür plädieren, dass Organe für potentielle Empfänger von so großer Bedeutung sein können, dass unter bestimmten – sehr selten eintretenden – Umständen auch die Entnahme *gegen* den erklärten Willen des Verstorbenen bzw. seiner Angehörigen moralisch zulässig sein kann. Verfügungen über den eigenen Körper über den Tod hinaus sind damit nicht schlechthin unzulässig, aber sie sind auch nicht immer zulässig.

Da diese Position erfahrungsgemäß als völlig abwegig angesehen wird, werde ich hier probieren, mich nicht auf mir eigene philosophisch vielleicht dubiose, sicherlich aber immer kontroverse moralphilosophische Prämissen oder Theorien abzustützen, sondern versuchen aufzuzeigen, dass die Konfiszierungslösung auch von Überlegungen unterstützt wird, die primär auf eine Kohärenz weit verbreiteter moralischer Überzeugungen abzielen. Ich glaube, die Konfiszierungslösung entspricht in vielen Aspekten der gängigen Praxis – nur nicht der Praxis der Organspende natürlich. Ich werde daher zwei m.E. hinreichend ähnliche Konstellationen als Analogien anführen, mit denen kaum jemand in Deutschland grundsätzliche Schwierigkeiten hat: erstens das Erheben von Erbschaftssteuern und zweitens die Anordnung von Obduktionen. In beiden Fällen findet im Zweifelsfall auch gegen den Willen des Verstorbenen bzw. seiner Angehörigen ein Zugriff auf sein Vermögen bzw. seinen Körper statt. Beide Regelungen werden zumeist als völlig unproblematisch empfunden. Ich glaube, dass unter diesen Umständen auch Organkonfiszierungen als unproblematisch empfunden werden sollten.

Im nächsten, dem zweiten Abschnitt werde ich jedoch kurz ein Argument für die postmortale Konfiszierung von Organen darlegen, das ohne den erwähnten Analogieschluss auskommt. Dieses Argument ist hier nur als Anregung für weitere Überlegungen gedacht und soll nicht die Last der Schlussfolgerungen tragen. Sicherlich muss es, bevor es ernsthaft in Betracht gezogen werden kann, noch deutlich verfeinert werden. Da es so bereits in der Literatur existiert werde ich es hier nicht umfassend untersuchen, sondern nur darstellen und kurz erläuternd kommentieren.

Im dritten Abschnitt werde ich dann die Analogie zwischen Erbschafssteuern und postmortaler Organkonfiszierung beleuchten. Ich werde kurz die Analogie darstellen und auf einige wichtige Einwände eingehen.

Ein wichtiger Unterschied zwischen Organkonfiszierung und Erbschaftssteuern besteht sicherlich darin, dass bei der postmortalen Organkonfiszierung, nicht aber bei der Erbschaftssteuer, auf den *Körper* des Verstorbenen zugegriffen wird. Die Existenz von Erbschaftssteuern alleine kann daher nicht plausiblerweise ausreichen, um durch einen Analogieschluss postmortale Organkonfiszierungen zu legitimieren.[5] Daher beschreibe ich im vierten Abschnitt die derzeit in Deutschland zur Anwendung kommende Praxis der postmortalen Körperbearbeitung im Rahmen von Obduktionen. Diese werden deutlich häufiger vorgenommen als gemeinhin vermutet, nämlich knapp 40.000-mal pro Jahr. Nicht alle dieser Obduktionen geschehen ohne Berücksichtigung des Willen des Verstorbenen oder seiner Angehörigen, aber es sind doch immerhin über 16.000 Fälle pro Jahr (vgl. Brinkmann et. al. 2002). Bisher wird diese Praxis nicht ernsthaft hinterfragt, was angesichts der Diskussionen um die zahlenmäßig viel weniger bedeutsamen Organtransplantationen (ca. 1.200 Organspender pro Jahr[6], auf jeden Organspender kommen also knapp 33 Obduzierte, davon 13 Zwangsobduzierte) überrascht. Ich werde wiederum die Analogie präsentieren, um dann einige wichtige Einwände zu diskutieren.

Obduktionen finden stets an Verstorbenen statt, die herz-kreislauf-tot sind, wohingegen Organtransplantationen an *hirntoten* Spendern vorgenommen werden. Diesen Unterschied,

[5] Fabre (2006) diskutiert zwar Erbschaftssteuern, nicht aber Obduktionen, Herschenov und Delaney (2009) hingegen beschäftigen sich nur mit angeordneten Obduktionen.

[6] Siehe Deutsche Stiftung Organtransplantation (2012, S. 13).

der durchaus als von großem Belang angesehen werden kann, beleuchte ich im fünften Abschnitt. Hier komme ich letztlich zu dem Schluss, dass wir angesichts der immer noch schwelenden Debatte zum Hirntod zwei Möglichkeiten haben: entweder akzeptieren wir die momentane Gesetzeslage, der zufolge ein Hirntoter ‚richtig' tot ist, um es etwas salopp auszudrücken – dann müssten wir aber den vorausgegangenen Argumenten folgen und die postmortale Organkonfiszierung befürworten. Oder aber wir entscheiden uns, z.B. um dieser doch unangenehmen Schlussfolgerung entgehen zu können, zwar Herz-Kreislauf-Tote nicht aber Hirntote als ‚richtig' tot zu kategorisieren. Wenngleich diese Option das Problem der Konfiszierung lösen würde, da wohl niemand die erzwungene Tötung zum Zwecke der Organgewinnung befürworten möchte, wirft sie doch andererseits ähnlich unangenehme Probleme auf. Denn sie zu akzeptieren würde eben gerade bedeuten, dass Organentnahmen zu Transplantationszwecken an Hirntoten nicht aber an Toten im engeren Sinne vorgenommen würden, die Spender also durch den Vorgang getötet würden; mithin würde es bedeuten, dass es sich (bei einer Zustimmungslösung) um eine Tötung auf Verlangen handelt.

Diese Schlussfolgerung, die die derzeitige Praxis der Organtransplantation vollends ad absurdum führen würde, lässt sich leicht vermeiden, indem man Hirntote einfach wie bisher als ‚richtig' tot betrachtet. Damit ergibt sich nun aber doch, wie ich abschließend im sechsten Abschnitt zusammenfassen werde, dass die postmortale Organkonfiszierung tatsächlich zu befürworten ist.

### *1.3. Grenzen der Argumentation*

Gleich zu Anfang möchte ich kurz die Grenzen meiner Argumentation darlegen. Erstens betrachte ich nur bestimmte menschliche Organe, nicht aber menschliches Gewebe. Die genaue Definition dieser Konzepte möchte ich den Medizinern überlassen. Beispielsweise ist die Haut eines Menschen zwar umgangssprachlich auch ein Organ, für Zwecke der Transplantationsmedizin zählt sie jedoch als Gewebe (wie u.a. auch die Knochen, die Blutgefäße, die Augenhornhaut und die Herzklappen). Als Organe gelten nur das Herz, die Lunge, die Bauchspeicheldrüse, die Leber und die Nieren. Organe und Gewebe, zwischen denen gerade bei der Werbung neuer potentieller Organspender häufig nicht hinreichend differenziert wird, werfen völlig unterschiedliche philosophische Probleme auf und sind daher auch völlig unabhängig voneinander zu betrachten. Diese umfassende Untersuchung kann hier leider nicht geleistet werden. Zu erwähnen ist jedoch, dass Gewebe auch Spendern entnommen bzw. aus und von deren Körpern entfernt kann, wenn sie nicht hirntot sind, sondern bereits seit einiger Zeit herz-kreislauf-tot. Damit würden sich die im fünften Abschnitt diskutierten Probleme für Gewebespenden nicht ergeben. [7]

Zweitens beschränke ich mich nicht nur auf Organe sondern sogar nur auf die Organe, die *lebensnotwendig* sind, deren Transplantation also das Leben eines Menschen retten kann. Meine Argumente sollen also nicht für die zweite funktionierende Niere gelten, die die Lebensqualität eines Menschen zwar deutlich erhöht, nicht aber für ein menschenwürdiges Überleben im engeren Sinne erforderlich ist (wie die Praxis der Lebendnierenspende illustriert). Die hier vorgelegten Argumente, die sich aus didaktischen Gründen also nur mit dem ‚einfachsten' Fall befassen, also den für den Leidtragenden am medizinisch schlimmsten, lassen sich für nicht lebensnotwendige aber die Lebensqualität verbessernde Organ- und Gewebetransplantationen ebenfalls vorbringen, müssen aber erheblich modifiziert werden.[8]

---

[7] Einige der Probleme, die sich wiederum nur für Gewebespenden ergeben, sind zusammenfassend in Schmidt-Petri und Himpsl (2012) diskutiert.

[8] Aber selbst bei den lebensnotwendigen Organen sind eigentlich weitere Differenzierungen vonnöten. Schließlich stirbt jeder Mensch früher oder später, selbst wenn (bis zum Tod) alle lebensnotwendigen Organe funktioniert haben; man kann also auch mit einer Reihe von Austauschorganen den Tod nicht unbeschränkt heraus zögern. Insofern mag es irreführend erscheinen, eine Leber als lebensrettend oder

Drittens werde ich Fragen der gerechten *Verteilung* der Organe nicht thematisieren, sondern mich nur auf die Frage der gerechten *Beschaffung* konzentrieren. Durch die hier vorgeschlagene Konfiszierungslösung dürfte sich das Problem des lautstark beklagten ‚Organmangels' (z.B. Breyer et. al. 2006) zwar nicht mehr in dem Ausmaß stellen wie dies bisher der Fall ist. Trotzdem kann selbst mit einer Konfiszierungslösung vielleicht nicht jeder, der sich gerne eine bestimmte Behandlung wünscht, diese auch erhalten. Denn vielleicht sind nicht alle Organe gleichermaßen geeignet, vielleicht die Behandlung nicht aller Patienten gleichermaßen medizinisch indiziert oder gleichermaßen durch Gerechtigkeitsüberlegungen gefordert. Verteilungsfragen stellen sich also fast mit Sicherheit selbst bei einem starken Anstieg der verfügbaren Organe. Es muss immer noch geklärt werden, wer warum welche Organe erhalten soll. Hierzu leistet der folgende Text keinen Beitrag.

Viertens stelle ich erklärtermaßen nicht den Anspruch, einen konkreten Vorschlag zur politischen Gestaltung Deutschlands zu machen. Eine Konfiszierungslösung hat keinerlei realistische Chance auf Verwirklichung und auch ich möchte mich eigentlich nicht für ihre Einführung aussprechen. Ich möchte nur aufzeigen, dass eine sorgfältige Betrachtung zeigt, dass bessere Argumente benötigt werden um sie dauerhaft ablehnen zu können.

## 2. Das Direkte Argument

Cécile Fabre hat unlängst ein m.E. fast vollständig gelungenes Argument für die postmortale Organkonfiszierung präsentiert (Fabre 2006, Kap. 4.). Da das Argument ‚direkt' ist, also nicht wie meins, mit einem Analogieschluss arbeitet, kommt es natürlich nicht ohne explizite Prämissen aus. Diese gliedern sich in gerechtigkeitstheoretische und empirische. Fabres Argumentation ist, wie darüberhinaus anzumerken ist, vollständig idealtheoretisch aufgebaut (S. 8). Dies bedeutet, dass sie sich nicht mit dem ‚real life' der gesellschaftlichen Wirklichkeit beschäftigt, in denen Menschen gerechte Regeln ungerechterweise zu ihrem eigenen Vorteil interpretieren oder sogar legitime Gesetze brechen, sondern mit der ‚ideal world', die der unseren nur ähnelt. Die ideale Welt unterscheidet sich von der Wirklichkeit dadurch, dass in ihr die Forderungen der Gerechtigkeit allesamt erfüllt sind, alle Personen sich an der Aufrechterhaltung einer gerechten Gesellschaft beteiligen und niemandem etwas ihm Zustehendes vorenthalten wird.

Ausbuchstabiert bedeutet dies in ihrer Gerechtigkeitstheorie unter anderem, dass alle Menschen ein Dach über dem Kopf haben, genug zu essen, Zugang zu Gesundheitsvorsorge und überhaupt über all das verfügen können, was für ein *minimally flourishing life* erforderlich ist, also ein Leben, dass in der bestimmten Gesellschaft, in der es geführt wird, vernünftigerweise als minimal erfülltes oder minimal gutes Leben angesehen wird. Das minimally flourishing life (MFL), das in einem Suffizienzprinzip ausformuliert wird (S. 33 und passim), ist also der Lackmustest einer jeden gesellschaftlichen Praxis.

> **Suffizienzprinzip**
>
> Alle Menschen haben Anspruch auf die Dinge, die sie für ein minimal erfülltes Leben benötigen.[9]

---

lebensnotwendig zu bezeichnen. Für eine erste Annäherung an das Thema möchte ich dies aber vernachlässigen.

[9] Dies ist das für meinen Kontext entscheidende Prinzip. Ihre Gerechtigkeitstheorie beinhaltet jedoch noch ein zweites Prinzip, das Autonomieprinzip. Dies erklärt, dass alle Menschen mit ihrem Eigentum tun und lassen dürfen was sie wollen - insofern das Suffizienzprinzip bereits erfüllt ist, das dem Autonomieprinzip somit lexikalisch vorgeordnet ist (S. 34 und passim).

Wie unschwer zu erkennen ist, führt das in ihrer Theorie u.a. dazu, dass (sorgfältig regulierter und staatlich abgewickelter) Organhandel zulässig ist, insofern Organe zur Verfügung stehen, die nicht zur Erreichung eines MFL konfisziert werden müssen und es Menschen gibt, die ihre Organe verkaufen

Diese Anspruchsrechte, die moralischer Natur sind und dann in positives Recht überführt werden müssen, sind durch Rückgriff auf menschliche Interessen begründet, wobei das gerechtigkeitstheoretisch wichtigste und fundamentale Interesse das an *Selbstrespekt* ist.[10] Ein MFL ist also, zusammenfassend, die einzige notwendige Bedingung dafür, dass man Selbstrespekt haben kann.

Was benötigt man nun ganz konkret für ein MFL? Diese entscheidende Frage wird sich hinsichtlich vieler Einzelfälle nicht ohne weiteres beantworten lassen.[11] So könnte man sich fragen, ob man als erwachsener Mensch ein Auto oder eine Spülmaschine benötigt, um ein MFL haben zu können oder ob man normalerweise auch mit einem Fahrrad und Handwäsche zu Rande kommen sollte. Die Antwort dürfte von Gesellschaft zu Gesellschaft unterschiedlich ausfallen und sicherlich müsste man auch die berufliche und familiäre Situation berücksichtigen. Für viele andere Fälle liegt die Antwort aber auf der Hand. So ist *Gesundheit* für ein MFL sicherlich generell erforderlich, oder zumindest das, was ich „Grundgesundheit" nennen möchte. Vereinzelter Karies oder ein Schnupfen, wenngleich Krankheiten, tangieren die Grundgesundheit kaum (sicher nicht, wenn sie rechtzeitig behandelt werden und sich nicht zu anderen gesundheitlichen Problemen entwickeln können), wohingegen schwere Herzrhythmusstörungen dies zweifellos tun.

In der Suffizienztheorie von Cécile Fabre haben also Patienten mit Herzrhythmusstörungen Anspruch auf die Implantation eines Herzschrittmachers und die erforderliche Behandlung und Medikamente, da ein einigermaßen normal funktionierendes Herz für ein MFL erforderlich ist. Natürlich gilt das nur in Gesellschaften, in denen Herzschrittmacher überhaupt gerechterweise zur Verfügung stehen können. Aber für Länder wie Deutschland ist das eindeutig der Fall. Die anderen Mitglieder der Gesellschaft sind somit in der Bringschuld, Herzschrittmacher und die erforderliche Behandlung zu finanzieren. Sie wären hingegen nicht dazu verpflichtet, Behandlungen zu finanzieren, die nicht dazu beitragen nur ein MFL zu erreichen, sondern ausschließlich oberhalb dieser Schwelle die Gesundheit optimieren, wie zum Beispiel preisgünstige Tabletten, die leichte temporäre Kopfschmerzen lindern.

Kann man nun nicht nur auf ‚normale' materielle Güter wie Herzschrittmacher (oder künstliche Herzen) sondern auch auf Körperteile anderer Menschen einen Anspruch haben? Fabre bejaht dies. Wenn man andere Menschen dazu verpflichten kann, einen Teil ihres Einkommens für die Bereitstellung von medizinischer Behandlung abzugeben, dann kann man sie auch dazu verpflichten, einen Teil ihres Körpers für die Bereitstellung medizinischer Behandlung abzugeben.[12] Selbstverständlich ist dabei das Suffizienzprinzip auch in Bezug auf den Körperteilegeber zu respektieren, nicht nur in Bezug auf den Empfänger. Es wäre daher nicht zulässig, einem Menschen ein Herz zu explantieren um es einem anderen Menschen zu implantieren, auch wenn dieser es tatsächlich benötigt. Denn der Spender benötigt es auch und würde durch die Explantation das MFL verlieren, das der andere – vielleicht – gewönne.[13] Wenn der Spender aber tot ist und das ist der für die hier relevanten Organe

---

möchten (und einige weitere Bedingungen erfüllt sind). Durch das Suffizienzprinzip ist bereits gewährleistet, dass dies nicht aus Not geschehen kann, reine Armut des Verkäufers ist als Grund für den Verkauf also ausgeschlossen. Ich gehe auf dieses kontroverse Thema hier nicht ein, da es für das vorliegende Argument nicht relevant ist und umfassender diskutiert werden müsste.

[10] Wobei dies natürlich sogenannter „recognition respect" ist, vgl. Darwall (1977).

[11] Fabre bezieht sich hier auf den „capability approach" in der Variante von Martha Nussbaum (2000), dessen allgemeine Gültigkeit situationsabhängige Spezifizierungen vorsieht.

[12] Zur Klarstellung: dies ist nicht als reductio im Stile Nozicks gedacht (vgl. Nozick 1974, 169ff). Fabre spricht sich auch explizit für Arbeitsdienste aus, insofern diese zur Erreichung eines MFL erforderlich sein sollten (2006, Kap. 2).

[13] Hier entstehen einige diffizile Probleme, die ich leider zurückstellen muss. Man könnte sich beispielsweise fragen, ob der Gesunde denn nur durch seinen (hier in erster Näherung als zufällig angesehenen) Gesundheitsstatus einen Vorteil erlangen soll, nämlich gute Gesundheit, um deren gerechte Verteilung es doch gerade geht. Pragmatisch ist dies leicht zu lösen, denn jede Transplantation

wichtige Fall, hat er ohnehin kein MFL mehr, das beachtet werden muss. Insofern dann Organe transplantiert werden können, sollte dies also getan werden. Um genau zu sein: der Bedürftige hat das Anrecht darauf (das er aber nicht zwangsläufig wahrnehmen muss), das Organ transplantiert zu bekommen, und die anderen Mitglieder der Gesellschaft sind verpflichtet, ihm das Organ zu verschaffen.[14]

Dies bedeutet natürlich, dass weder dem Organgeber – von einer Spende möchte man wohl nicht mehr sprechen – noch seinen Angehörigen ein Mitspracherecht eingeräumt werden kann. Es handelt sich also um ein System der Organkonfiszierung.

## 3. Analogie zur Erbschaftssteuer

### *3.1. Die Analogie*

Die Analogie zur Erbschaftssteuer wird ebenfalls von Fabre angeführt, obwohl dies für den Erfolg ihres Arguments eigentlich nicht erforderlich ist. Aus ihrer Theorie ergibt sich direkt, dass Organkonfiszierungen geboten sind, völlig unabhängig davon, ob es Erbschaftssteuern gibt oder nicht. Darüberhinaus ergibt sich in ihrer Theorie keineswegs ohne Weiteres die Legitimität von Erbschaftssteuern. Denn es muss als strittig angesehen werden, ob Erbschaftssteuern erforderlich sind, um allen Bürgern ein MFL zu garantieren. Selbst wenn sich in einer Gesellschaft ohne Erbschaftssteuern enorme Vermögensunterschiede ergäben, die durch eine Erbschaftssteuer verhindert würden, wäre dies allein noch kein Grund, eine Erbschaftssteuer einzuführen. Enorme Vermögensunterschiede dürften kaum an sich dazu führen, dass das MFL der weniger Begüterten gefährdet ist. Aber auch dies ließe sich bestreiten.

Das Argument, das ich hier vorstellen möchte soll eine andere Struktur haben. Ich möchte ohne Rückgriff auf eine explizite Gerechtigkeitstheorie aus der – somit nur angenommenen – Legitimität von Erbschaftssteuern und der – ebenfalls nur angenommenen – Legitimität von Zwangsobduktionen auf die – daher nur bedingt aufgezeigte – Legitimität von Organkonfiszierungen schließen.

Dieses Argument soll also niemanden überzeugen können, der Erbschaftssteuern oder Zwangsobduktionen ablehnt. Argumente für die Legitimität von Erbschaftssteuern oder Zwangsobduktionen werde ich auch nicht vorbringen. Meine Argumente wären sogar gänzlich uninteressant, würden Erbschaftssteuern und Zwangsobduktionen nicht tatsächlich von vielen Menschen als legitim angesehen. Ich stelle also nur die Behauptung auf, dass jemand, der, aus welchen Gründen auch immer, sowohl Erbschaftssteuern wie auch Zwangsobduktionen als legitim ansieht (in ungefähr der Art, wie beide in Deutschland praktiziert werden), ebenfalls Organkonfiszierungen befürworten muss. Ich glaube, dass dies bereits ein interessantes Ergebnis ist.

Inwieweit sind Organkonfiszierungen und Erbschaftssteuern nun analog? Die Ähnlichkeiten liegen auf der Hand. Entscheidend ist der *erzwungene Ressourcentransfer nach dem Tod*. Organgeber und Erblasser sind beide tot. Beiden wird etwas genommen, das sie aufgrund

---

ist für den Empfänger zwangsläufig riskanter als das Weiterleben eines Gesunden für ihn ist. Der Endzustand wäre also, wenn man von der Identität der Personen absieht, nicht besser als der Ausgangszustand. Aber es ist zu bezweifeln, ob dies allein bereits philosophisch befriedigend sein kann.

[14] Diese Argumentation ist nicht utilitaristisch. Ziel ist nicht, das gesellschaftliche Wohlergehen zu maximieren, sondern Individuen ein MFL zu garantieren. Die Überlegung, ob der Spender mehr verliert als der Empfänger gewinnt ist irrelevant, insofern nicht bei einem von beidem das MFL tangiert wird. Auch dann geht es nicht darum, Wohlergehensverluste oder –gewinne miteinander aufzurechnen. Ganz im Gegenteil ist damit zu rechnen, dass zur Gewährleistung eines MFL relativ hohe Opfer gebracht werden müssten, so dass eine gerechte Welt à la Fabre kaum wohlergehensmaximierend sein dürfte. Hier sind die Parallelen zur Gerechtigkeitstheorie von Rawls (1971) augenfällig.

ihres Todes nicht mehr benutzen können und anderen Personen übereignet, die es weiterhin benutzen können.

In dieser kurzen Form kann das Argument sicher nicht überzeugen. In Sektion 5. werde ich erörtern, ob wirklich beide Personen tot sind. Dass der Transfer erzwungen ist, gilt für Erbschaftssteuer und Organkonfiszierung gleichermaßen. Zu klären ist damit vor allem, ob der Körper als ‚Ressource' betrachtet werden kann. Diese Schwierigkeit und einige andere potentielle Einwände werde ich in den folgenden Seiten bearbeiten.

### *3.2. Einwand 1: Eine Steuer ist keine Konfiszierung; es gibt auch Freibeträge*

Der erste Einwand zielt auf das Ausmaß des Transfers: Eine Steuer ist ja nur eine Steuer, so lautet der Einwand, und keine Konfiszierung. Ein Großteil des Erbes wird doch weitergegeben, da die Steuer nur einen vergleichsweise geringen Anteil des Erbes ausmacht (der Steuersatz beträgt z.B. 20%). Darüberhinaus sind ‚normale' Erbschaften ja häufig völlig steuerfrei. Wenn Eltern ihren Kindern oder Enkeln Vermögen vermachen, wie beim typischen Erbfall, wird erst ab einem relativ hohen Vermögen überhaupt Erbschaftssteuer fällig (z.B. wenn der Wert des Nachlasses 400.000 Euro übertrifft) und dann auch nur auf den Teil, der diesen Freibetrag überschreitet. Dies kann man doch nicht mit einer Konfiszierung gleichsetzen.

Dieser Einwand kann nicht überzeugen. Auch bei der Organkonfiszierung würde nicht der ganze Leichnam konfisziert, sondern nur die Organe entnommen, die als Transplantate in Frage kommen. Auch bei der Organkonfiszierung gilt, dass sie nur sehr selten zum Tragen käme. Denn weniger als 0,5% aller Todesfälle in Deutschland (nämlich ungefähr 4.000 Menschen pro Jahr[15]) kämen als hirntote Organgeber überhaupt in Frage und auch dann müssen weitere medizinische Kriterien erfüllt sein. Tatsächlich wäre die Wahrscheinlichkeit zur Organkonfiszierung herangezogen zu werden ein Bruchteil von der, Erbschaftssteuer bezahlen zu müssen, denn in immerhin fast 13% aller Erbfälle wird Erbschaftssteuer fällig.[16]

### *3.3. Einwand 2: Wie soll der Wert eines Organs veranschlagt werden?*

Der zweite Einwand beschäftigt sich mit der Berechnung der Abgabenlast: Wenn Organe nach dem Tode besteuert werden sollen wie alle anderen Vermögensarten, müsste man den Wert eines Organs ja genau beziffern können. Denn da, wie gesagt, nur ein Bruchteil des Vermögens als Steuer eingezogen wird, ist zu klären, wie der Rest der Erbmasse durch die Organkonfiszierung beeinflusst wird. Denn u.U. ist ja noch die konventionelle Erbschaftssteuer fällig. Transplantierbare Organe sind zwar rar, haben aber keinen pekuniären Wert, da es keinen Markt für Organe gibt. Damit kann die Steuerlast nicht berechnet werden, der Vorschlag ist also impraktikabel und die Analogie hinfällig.

Dieser Einwand beruht auf einem Missverständnis. Es ist zwar richtig, dass es keinen Marktpreis für Organe gibt und ein solcher wäre vielleicht tatsächlich nicht erstrebenswert. Die Analogie setzt aber auch keinen voraus. Das Ziel ist nicht, die Erbschaftssteuer auch auf Organe anzuwenden, sondern die Organe den Organbedürftigen zur Verfügung stellen zu können. Es handelt sich somit nicht um eine Erweiterung der Erbmasse, sondern um ein völlig separates Verfahren.

---

[15] Denn in „deutschen Krankenhäusern sterben jährlich rund 400.000 Menschen. Lediglich bei ungefähr einem Prozent der Verstorbenen tritt der Hirntod vor dem Herzstillstand ein" Siehe http://www.organspende-info.de/information/spende-und-transplantation/organspender der BZgA (letzter Zugriff 31.1.2013).

[16] Laut Statistischem Bundesamt (2012a, S. 11) gab es 2011 genau 110.595 steuerpflichtige Erwerbe von Todes wegen bei 852.328 Todesfällen (2012b, S. 1), der genaue Prozentsatz beträgt also 12,98%.

### *3.4. Einwand 3: Erbmasse wird langfristig akkumuliert, Hirntod ist Zufall*

Der dritte Einwand zweifelt indirekt die Gerechtigkeit einer solchen Organkonfiszierung an. Wieso sollten denn nur die Verstorbenen aufgeschnitten werden, die einen Hirntod gestorben sind? Hier handelt es sich doch um eine eklatante Ungleichbehandlung. Schließlich ist es meist Zufall (im umgangssprachlichen Sinne), wenn man durch einen Unfall einen Hirntod erleidet. Das zu vererbende Vermögen wird aber über die gesamte Lebenszeit angespart. Darüberhinaus gibt es diverse Möglichkeiten die Erbschaftssteuer völlig legal zu vermeiden, z.B. durch Schenkungen zu Lebzeiten. Für die Organkonfiszierung gilt auch das nicht. Die beiden Szenarien sind also nicht analog.

Es ist richtig, dass die Organkonfiszierung wohl kaum umgangen werden kann. Dies erscheint aber genauso wenig als guter Einwand *gegen* die Organkonfiszierung gelten zu können, wie die Tatsache, dass die Erbschaftssteuer häufig legal vermieden werden kann, *für* die Erbschaftssteuer spricht (darüber hinaus kann man bei hinreichend höhen Vermögen die Erbschaftssteuer natürlich nicht wirklich völlig umgehen). Wichtiger ist den Einwand der Ungleichbehandlung auszuräumen. Gerade weil der Hirntod ‚zufällig' auftritt, also alle Menschen ungefähr die gleiche Wahrscheinlichkeit haben ihn zu erleiden,[17] scheinen Fairnessüberlegungen kaum gegen die Organkonfiszierung sprechen zu können.

### *3.5. Einwand 4: Organmangel ist gar kein Gerechtigkeitsproblem*

Der vierte Einwand versucht die Sphäre der distributiven Gerechtigkeit noch stärker von der gesundheitlichen Sphäre abzukoppeln. Einige Menschen haben Pech, so der Einwand, und leiden an schlechter Gesundheit, sehr wenige benötigen sogar Transplantate. Aber entweder haben wir im Bereich der individuellen Gesundheit gar keine Gerechtigkeitspflichten oder wenn doch, dann reichen diese nicht so weit, dass wir Verstorbenen Organe entnehmen müssen, um Organbedürftige zu versorgen.

Dieser Einwand hat viel Potential. Da aber klar erscheint, dass wir anderen Menschen auch medizinische Hilfe schulden (insofern sie gerechterweise zur Verfügung stehen kann), erscheint eine *strikte* ‚Sphärentrennung' nicht plausibel. Eine *Begrenzung* der Gerechtigkeitspflichten gerade im Bereich der Gesundheit ist jedoch seit Längerem ein kontrovers diskutiertes Thema; diese Debatte kann hier nicht in ausreichender Länge wiedergegeben werden. Es erscheint aber sehr fraglich, dass bei einer Beschränkung dann gerade die Art von Behandlung entfallen sollte, die lebensrettend ist.[18] Bestenfalls könnte argumentiert werden, dass die Kosten oder Lasten auf Seiten der Organgeber unzumutbar wären. Hier setzt der nächste Einwand an.

### *3.6. Einwand 5: Der menschliche Körper ist keine Sache*

Der fünfte und letzte Einwand ist der im ersten Augenblick überzeugendste. Er setzt an, wo der vierte Einwand abbrach und bestreitet die Gleichartigkeit von materiellen Gütern und dem menschlichen Körper. Wir haben zwar die Pflicht, uns gegenseitig alle medizinisch notwendigen Behandlungen und Medikamente zukommen zu lassen, aber der Körper eines Menschen ist doch keine Sache! ‚Unser' Körper gehört uns auch nicht in dem Sinn, in dem uns z.B. unser Fahrrad gehört. Dies ist doch eindeutig. Es ist ja geradezu konzeptionell unmöglich, seinen eigenen Körper im buchstäblichen Sinne zu verkaufen. Wir sind leibliche Wesen, unser gesamter Erfahrungsschatz wurde durch und mittels unseres Körpers

[17] Dies stimmt nicht ganz, denn der typische Hirntote war anscheinend lange Zeit der jugendliche Motorradfahrer.

[18] Dies erscheint mir unabhängig davon zu gelten, ob wir Fabres Theorie akzeptieren oder nicht.

gewonnen, sodass unser Körper Teil, und zwar ein entscheidender Teil nicht nur unseres Selbstbilds, sondern auch unserer personalen Identität darstellt.

Diesem Einwand, der hier ganz bewusst etwas diffus dargestellt wurde und der für eine sorgfältige Behandlung deutlich klarer zu fassen wäre, ist etwas schwieriger zu begegnen. Die ihm zugrundelegende Intuition ist letztlich nicht zu widerlegen. Wer den menschlichen Körper als separat von der restlichen materiellen Welt begreift, muss sich von dieser Position nicht abbringen lassen. Ich möchte dennoch einige Punkte vorbringen, warum ich diese Sicht für verfehlt halte.

Zuerst möchte ich gewissermaßen die materielle Welt aufwerten. Auch die Erbschaftssteuer macht nämlich vor den Dingen, die dem Verstorbenen am wichtigsten oder heiligsten waren, nicht Halt. Stellen wir uns einen Musiker vor, dem eine Stradivari im Wert von zwei Millionen Euro gehört. *Diese* Stradivari ist *sein* Instrument und nur mit diesem Instrument musizierend kann er wirklich er selbst sein. Die Geige ist, wie man unter Musikern sagen würde, ein Teil von ihm. Sein ganzes Leben hat er der Musik gewidmet, sonstiges nennenswertes Vermögen hat er nicht.

Bei seinem Tod wird die Stradivari dennoch verkauft werden müssen, da jegliche Freibeträge überschritten sind. Man könnte meinen, dass der Geiger sicher froh darüber sein wird, dass andere Musiker nun mit ihr spielen können; vielleicht wird er sich aber auch darüber ärgern, dass seine Stradivari nun von anderen Musikern, die er für wohlhabend aber untalentiert hält, rücksichtslos ruiniert wird. Dem Gesetz ist dies egal. Die Erbschaftssteuer muss bezahlt werden, ob der Steuerzahler dies befürwortet oder eben nicht. Die Parallelen zur Organkonfiszierung sind eindeutig und müssen nicht weiter ausgeführt werden.

Zweitens möchte ich gewissermaßen den menschlichen Körper abwerten. Hier ist nur zu konstatieren, dass viele Menschen durchaus ein distanziertes Verhältnis zu ihrem eigenen Körper haben und ihn nicht als Inbegriff ihrer Persönlichkeit sehen. Er ist ja doch ein Produkt der genetischen Ausstattung, auf die man keinen Einfluss hatte. So sind, selbst wenn wir anerkennen, dass der eigene Körper Persönlichkeit und Weltwahrnehmung entscheidend beeinflusst, eben Persönlichkeit und Weltwahrnehmung ebenfalls Ausfluss von genetischen Faktoren, auf die man keinen Einfluss hatte. Zumindest zeigen gerade die an sich hohen Zustimmungsquoten zur Organspende, dass viele Menschen mit der Idee, dass ihrem Körper nach dem Tod Organe entnommen werden, anscheinend gut leben können.

Drittens müssen wir uns ins Bewusstsein rufen, dass diese Zustimmung auch praktische Auswirkungen hat, Organtransplantationen nämlich bereits gängige Praxis sind. Wir müssen also doch davon ausgehen, dass es möglich ist, Organe von einem Menschen zum anderen zu transferieren, ohne dass dies die ‚Identität' des Verstorbenen in diesem weiteren Sinne zerstört. Es stimmt natürlich, dass bisher die Zustimmung des Verstorbenen (bzw. seiner Angehörigen) vorliegen muss, aber dieser fünfte Einwand geht doch so tief, dass er eigentlich auch die Organspende insgesamt ablehnen muss. Es ist ganz im Gegenteil unkontrovers, dass menschliche Körperteile von einem Menschen zum anderen transplantiert also transferiert werden können. Mehr ist für das vorliegende Argument nicht erforderlich.

Viertens darf auch nicht vergessen werden, dass Organspender tot sind. Wie wichtig einem der eigene Körper zu Lebzeiten auch erscheinen mag und wie bedeutsam er als Mittler zwischen den Menschen fungieren mag, so hinfällig erscheinen diese Überlegungen doch nach dem Tod.

Ich möchte gar nicht vorgeben, dass ich diese skizzenhaften Antworten für zwingend überzeugend halte; sie scheinen den Einwand aber doch zu schwächen. Die beste Replik auf diesen fünften Einwand (und viele seiner hier unerwähnten Variationen, die sich mit der Frage beschäftigen, wie ein Leichnam behandelt werden darf) ist aber eine völlig andere: Es gibt bereits eine gesellschaftlich anerkannte Praxis, auf die er ebenso anwendbar wäre. Diese Praxis ist aber völlig unkontrovers. Es handelt sich um gerichtlich angeordnete Obduktionen,

die, wenn der fünfte Einwand stichhaltig wäre, ebenfalls nicht zulässig wären – was sie aber offensichtlich sind. Es ist also inkonsistent, Organkonfiszierungen abzulehnen, Erbschaftssteuern und Zwangsobduktionen jedoch nicht.

## 4. Analogie zur Obduktion

### *4.1. Bestandsaufnahme*

Die innere Leichenschau hat eine viel längere Tradition als Organtransplantationen, die erst seit der Mitte des 20. Jahrhunderts erfolgreich durchgeführt werden. Obduktionen sind in Deutschland tägliche Routine.[19] Sie werden an knapp 5% aller Verstorbenen vorgenommen, was über 40.000 Fällen pro Jahr entspricht. Die Wahrscheinlichkeit obduziert zu werden, ist also zehnmal so hoch, wie die Wahrscheinlichkeit einen Hirntod zu erleiden. Die Obduktionsquote, die in der BRD in den 1980er Jahren noch doppelt so hoch war, ist im internationalen Vergleich sogar niedrig. In Finnland werden fast ein Drittel aller Verstorbenen obduziert, in Großbritannien immerhin noch 15%. Die Bundesärztekammer (BÄK) spricht sich dafür aus, deutlich mehr Obduktionen durchzuführen (u.a. an 30% aller in Krankenhäusern Verstorbenen, vgl. BÄK 2005, S. 44).

Obduktionen können aus vielerlei Gründen durchgeführt werden. Die sogenannte „klinische Sektion“ wird von der Bundesärztekammer als „letzte ärztliche Handlung im Rahmen der medizinischen Behandlung der Patienten“ verstanden (*op. cit.,* S. 5) und dient primär der medizinischen Qualitätssicherung und Überprüfung der Todesursachenstatistik. Die sogenannte „gerichtliche Sektion“ ist für mein Argument wichtiger. Sie betrifft deutschlandweit knapp 2% aller Todesfälle (in Berlin 6%, in Rheinland-Pfalz 1,2%). Sie wird angeordnet, wenn ein nicht natürlicher Tod nicht völlig auszuschließen ist und die Todesursache geklärt werden soll oder wenn die Sicherung von Beweisen erforderlich scheint (§§ 87 StPO). Auch erwähnenswert ist die Möglichkeit, Obduktionen bei meldepflichtigen Krankheiten durch das Infektionsschutzgesetz anzuordnen (IfSG §§ 1, 25, 26). [20]

### *4.2. Die Analogie*

Inwieweit sind Organkonfiszierungen und Zwangsobduktionen nun analog? Die Ähnlichkeiten liegen auch hier auf der Hand.[21] Entscheidend ist in diesem Fall die *erzwungene Körperteilentnahme nach dem Tod*. Organgeber und zu Obduzierender sind beide tot. In beiden Fällen wird der Körper geöffnet und werden Organe entnommen. Der einzige Unterschied scheint zu sein, dass bei der Transplantation die Organe transplantiert werden, wohingegen bei der Obduktion die Organe nach der Untersuchung entweder entsorgt werden (häufig werden sie durch die Obduktion zerstört) oder in den Körper des Verstorbenen zurück gelegt werden.

Wie gesagt werde ich in Sektion 5. erörtern, ob wirklich beide Personen tot sind. Unkontrovers ist sicher, dass man bei angeordneten Obduktionen genauso wenig widersprechen kann wie man es bei einer Organkonfiszierung könnte. Ebenfalls irrelevant dürften die etwaigen Unterschiede bei der Behandlung der Leichname sein. Welche Einwände gegen die Analogie gibt es noch?

---

[19] Die Zahlen in den folgenden Abschnitten sind zur besseren Illustration sehr großzügig gerundet und stammen aus dem Jahr 1999 (Brinkmann et al. 2002).

[20] Diese Liste soll nur die hier relevanten Szenarien berücksichtigen und enthält daher nicht alle möglichen Begründungen.

[21] Umso überraschender, dass sie erst vor kurzem in der Literatur Erwähnung fand, nämlich in Hershenov und Delaney 2009. Die hier diskutierten Einwände finden sich ähnlich dort.

### *4.3. Einwand 1: Die Verbrechensaufklärung ist wichtiger als ein neues Organ*

Der erste Einwand zweifelt daran, dass die Begründungen der beiden Praktiken vergleichbar sind. Wenn eine Obduktion angeordnet wird, geht es immerhin darum, ein mögliches Verbrechen aufzuklären. Das Vermeiden von Mord und Totschlag ist aber für das Gemeinwesen bedeutender als Organtransplantationen. Zwangsobduktionen sind daher gerechtfertigt, Organkonfiszierungen aber nicht.

Tatsächlich wird gerade die hohe Dunkelziffer bei Tötungsdelikten als ein wichtiger Grund für die Erhöhung der Obduktionsrate genannt (BÄK 2005, S. 7). Wir wissen naturgemäß nicht genau, wie viele Verbrechen nicht aufgeklärt werden und durch zusätzliche Obduktionen aufgeklärt werden könnten. Derzeit wird davon ausgegangen, dass ca. 1.000 Tötungsdelikte pro Jahr in Deutschland verübt werden (und weitere 1200-2400 unentdeckt bleiben, *loc. cit.*) Dies entspricht aber genau der Anzahl der Menschen, die, wie es häufig ausgedrückt wird, aufgrund von Organmangel ‚auf der Warteliste sterben'. Es ist nun anzunehmen, dass nur ein Bruchteil der Tötungsdelikte, von denen derzeit über 95% aufgeklärt werden, nur durch eine Obduktion aufgeklärt werden konnten. Dementsprechend ist es ebenfalls unmöglich zu sagen, wie viele Tötungsdelikte durch die aufgeklärten Tötungsdelikte, die nur durch eine Obduktion aufgeklärt werden konnten, durch Abschreckung verhindert wurden – die Bedeutung der über 16.000 gerichtlich angeordneten Obduktionen für die Verbrechensbekämpfung und -vermeidung ist also unklar. Andererseits können jedem Organspender durchschnittlich mehr als drei Organe entnommen werden. Wir können also leicht vereinfachend davon ausgehen, dass *eine* Organkonfiszierung ungefähr *drei* Menschenleben rettet. Selbst bei gutwilliger Betrachtung der gerichtsmedizinisch angeordneten Obduktionen scheint diese Quote bei Weitem nicht erreichbar, denn dann müssten bei 16.000 Zwangsobduktionen ja ungefähr 48.000 Tötungsdelikte abgeschreckt werden.[22]

### *4.4. Einwand 2: Töten ist schlimmer als Sterbenlassen*

Der zweite Einwand beschäftigt sich mit einem ähnlichen Problem. Es wird gemeinhin angenommen, dass Töten schlimmer als Sterbenlassen ist. Jemand der tötet, tut etwas viel schlimmeres als jemand, der ‚nur' nicht rettet. Dies scheint intuitiv auch angebracht, denn sonst wäre jemand, der nur wenige Menschen vor dem Verhungern rettet (z.B. durch Geldspenden nach Afrika) aber trotzdem viele andere sterben lässt, obwohl er sie retten könnte, einem Massenmörder gleichzusetzen, was abstrus wäre. Obduktionen können nun, wie gerade ausgeführt, Tötungen verhindern. Transplantationen können dies aber nicht. Sie können nur Sterbenlassen verhindern.[23] Der potentielle Organempfänger wird ja, trotz aller Rhetorik, klarerweise nicht getötet, sondern stirbt an Organversagen, wenn er auch gerettet werden könnte. Auch dieser Einwand schlussfolgert also, dass Obduktionen wichtiger sind als Organkonfiszierungen, da Obduktionen etwas Schlimmeres verhindern.

Die moralische Differenzierung zwischen Töten und Sterbenlassen ist philosophisch höchst umstritten.[24] Diese Debatte muss hier jedoch gar nicht interessieren. Hier geht es darum, was wir, die Lebenden, mit den jeweiligen Leichnamen tun dürfen. Die noch lebenden Menschen, die wir vor dem Tod retten könnten (durch Verbrechensaufklärung bzw. Organtransplantationen), töten auch wir nicht, sondern schlimmstenfalls lassen wir sie sterben. Dies gilt aber für beide Fälle. Denn in beiden Fällen (dem unnatürlichen Tod durch

[22] Nur am Rande braucht hier erwähnt werden, dass der Schutz vor meldepflichtigen Krankheiten ebenfalls als legitimer Grund für Obduktionen gelten kann. Hier handelt es sich also nicht um die Bekämpfung von Verbrechen, sondern um die Gefahreneindämmung.

[23] Ich ignoriere hier einige philosophisch denkbare Fallkonstellationen.

[24] Siehe beispielsweise die in Steinbock und Norcross (1994) gesammelten klassischen Beiträge.

Tötung und dem natürlichen Tod durch Organversagen) sind wir in der Lage zu retten. Mit anderen Worten müssen wir uns also entscheiden, ob wir beide sterben lassen wollen, beide retten wollen, oder nur die retten wollen, die (von einer dritten Person) getötet würden. Hier erscheint es völlig arbiträr, zwischen diesen Szenarien zu differenzieren. Da wir grundsätzlich zweifellos auch Menschen vor dem Tod retten wollen, für deren Tod wir nicht kausal verantwortlich sind – was u.a. dadurch deutlich wird, dass wir Zwangsobduktionen im Dienste der Verbrechensabschreckung akzeptieren – sollten wir nicht ohne weiteren Grund zwischen den Todesursachen differenzieren, an denen wir ohnehin nicht als Verursacher beteiligt sind. Dies bedeutet nur, dass alle Gelegenheiten, Leben zu retten als gleich dringend angesehen werden müssen, nicht, dass Töten und Sterbenlassen moralisch äquivalent sind. Die Analogie kann so nicht angegriffen werden.

### *4.5. Einwand 3: Zustimmung kann angenommen werden*

Der dritte Einwand stellt eine empirische Hypothese auf. Er bestreitet, dass die Tatsache, dass Obduktionen angeordnet werden *können* auch bedeutet, dass sie angeordnet werden *müssten*.[25] Die zu Obduzierenden hätten der eigenen Obduktion sicher zugestimmt, wenn sie nur in Betracht gezogen hätten, dass sie – vermeintliches – Opfer von Verbrechen werden würden. Natürlich kann man nach dem Tod nicht mehr die Zustimmung des Verstorbenen einholen, so dass es auch möglich sein muss, Obduktionen anzuordnen. Man kann aber nicht davon ausgehen, dass dies nicht dem Willen des Verstorbenen entspricht, der sicher zumindest vor seinem Tod hohes Interesse an Verbrechensaufklärung hatte. Der Zwang hat also hier nur verfahrenstechnische Gründe. Die eigentliche Rechtfertigung ist, was Juristen als die „mutmaßliche" Zustimmung des Verstorbenen bezeichnen.

Diesem Einwand kann auf dreierlei Art begegnet werden. Erstens kann schlicht bestritten werden, dass alle zu Obduzierende der Obduktion zugestimmt hätten. Da die Datenlage hier naturgemäß keine Entscheidung bringen kann, beruht der Einwand auf Spekulation.

Wir können uns dann zweitens fragen, wenn wir diese Spekulation dennoch anstellen wollen, ob es eher vernünftig ist, Zwangsobduktionen zur Verbrechensaufklärung zu befürworten oder Organkonfiszierungen. Unter Berücksichtigung der Repliken auf Einwände in 4.3. und 4.4. scheint es mir klar, dass Zustimmung zu Organkonfiszierungen mindestens genauso vernünftig ist wie die Zustimmung zur Zwangsobduktionen. Die Wahrscheinlichkeit, sein eigenes Leben zu retten, erhöht man durch die Zustimmung zu Organkonfiszierungen noch stärker als durch die Zustimmung zu Zwangsobduktionen, und zu verlieren hat man in beiden Fällen das Gleiche. Wenn wir also Zwangsobduktionen durch die mutmaßliche Zustimmung des Verstorbenen zu Lebzeiten legitimieren können, können wir sicher auch die Zustimmung zu Organkonfiszierungen durch die mutmaßliche Zustimmung zu Lebzeiten legitimieren. Die Analogie kann also so nicht bestritten werden.

Drittens ist festzuhalten, dass Tötungsdelikte Straftaten sind. Straftaten werden, anders als privatrechtliche Streitigkeiten, von Staats wegen verfolgt, unabhängig davon, ob Täter oder Opfer selbst die rechtlichen Schritte einleiten. Selbst wenn das Opfer also an der Aufklärung der Straftat kein Interesse hat, steht es ihm nicht frei darüber zu entscheiden, ob Aufklärung stattfinden soll. Die mutmaßliche Zustimmung des Opfers zur Zwangsobduktion im Dienste der Aufklärung der Straftat sollte also ohnehin keine entscheidende Rolle spielen.

Mir scheint also, dass Erbschaftssteuern und Zwangsobduktionen jeweils wichtige Bausteine für die Rechtfertigung von Organkonfiszierungen liefern. Der erzwungene Ressourcentransfer nach dem Tod und die erzwungene Körperteilentnahme nach dem Tod sollten zusammen auch den erzwungen Körperteiletransfer nach dem Tod legitimieren.

---

[25] Dieser Einwand ließe sich auch hinsichtlich der Erbschaftssteuer vorbringen, wo er jedoch deutlich weniger plausibel erscheint.

## 5. Problem Hirntod

### *5.1. Problemaufriss*

Die in den vorangegangenen Abschnitten entwickelte Analogie zwischen der Erbschaftssteuer und angeordneten Obduktionen einerseits und der postmortalen Konfiszierung von lebensnotwendigen Organen andererseits hat bisher ein in der Debatte um die Transplantationsmedizin häufig und leidenschaftlich diskutiertes Problem vernachlässigt: Den Hirntod. Viele Menschen glauben, dass Menschen die hirntot sind noch nicht ‚richtig' tot sind,[26] u.a. weil das Herz von Hirntoten, die künstlich beatmet werden und die daher nicht den andernfalls unwiderruflich dem Hirntod folgenden Herz-Kreislauf-Tod erleiden, weiterhin schlägt. Darüberhinaus sind – bis auf das Gehirn – alle Organe im Wesentlichen in dem Zustand, in dem sie auch vor dem Hirntod gewesen sind, wenngleich dieser Zustand künstlich aufrechterhalten werden muss. Der Hirntod ist also in den Augen der sogenannten „Hirntodkritiker" nicht der Tod des Menschen, sondern nur der Tod *des Hirns* des Menschen.

Die Details dieser Debatte sind für mein Argument nicht von Belang. Wahrscheinlich treffen hier fundamentale Intuitionen über die menschliche Existenz aufeinander, die sich durch Argumente nur begrenzt beeinflussen lassen. Entscheidend ist hier, dass es die Diskussion gibt und dass sie im Allgemeinen als relevant für die Frage angesehen wird, ob Organexplantationen moralisch zulässig sein können.

Für meine bisherige Argumentationskette stellt sich also das folgende Problem: Hirntodkritiker können die zugrundeliegende Analogie ohne weiteres ablehnen. Erblasser und zu Obduzierende sind nämlich immer richtig tot,[27] Organspender jedoch nur hirntot.[28] Das ist ein wichtiger Unterschied.

Es bieten sich für meine Argumentationskette somit drei mögliche Positionen, von denen die erste die derzeit gültige Rechtslage akzeptiert, die anderen beiden das Hirntodkriterium ablehnen, aus dieser Ablehnung aber unterschiedliche Schlüsse ziehen:

(1) Hirntote sind richtig tot. Also sind die vorangegangenen Argumente ohne Einschränkung zu akzeptieren, postmortale Organkonfiszierungen zulässig.

(2) Hirntote sind nicht richtig tot. Also werden sie durch die Organentnahme getötet. Da nur Tote für eine moralisch legitime Organentnahme zur Verfügung stehen dürfen (die sogenannte „Dead Donor Rule") sind alle Organentnahmen moralisch unzulässig.

(3) Hirntote sind nicht richtig tot. Also werden Sie durch die Organentnahme getötet. Dies ist aber moralisch zulässig, denn die Dead Donor Rule ist falsch und die Organentnahme unter gewissen Umständen auch bei Menschen legitim, die nicht richtig tot sind.

---

[26] Ich werde die Anführungszeichen im Folgenden weglassen ohne die durch sie ausgedrückte Skepsis hinsichtlich der Idee des ‚richtigen' Todes aufgeben zu wollen.

[27] Andere Fälle dürften in der Praxis kaum auftreten, wenn sie auch rein rechtlich vielleicht möglich wären; schließlich ist der Herz-Kreislauf-Tod die absehbare Folge des Hirntods, insofern der Hirntote nicht künstlich beatmet wird. Und selbst wenn die künstliche Beatmung eines Hirntoten abgeschaltet wird, ohne dass er zum Organspender wird, stellt sich die Frage einer etwaigen Obduktion bzw. der Testamentseröffnung wohl bereits aus praktischen Gründen erst nach dieser Entscheidung. Das künstliche Beatmen eines Hirntoten – der ja de jure tot ist – hätte wohl ohne Aussicht auf Besserung oder Organentnahme zu Transplantationszwecken keinerlei Sinn und dürfte daher nicht praktiziert werden.

[28] Letzteres stimmt auch nicht mehr, da Spender, deren Herz nicht mehr schlägt, die aber noch nicht hirntot sind (sogenannte „non-heartbeating donors") vereinzelt auch zur Organspende herangezogen werden, wobei dies in Deutschland als illegal angesehen wird und nicht praktiziert wird (vgl. Stoecker 2010, S. xli). Nicht nur Hirntodkritiker dürften dieses Verfahren ablehnen.

### *5.2. Die erste Position*

Die erste Position entspricht der momentanen Gesetzeslage und dem bisherigen Argumentationsverlauf und bedarf momentan keiner weiteren Diskussion. Ich werde weiter unten auf sie zurückkommen.

### *5.3. Die zweite Position*

Die zweite Position bringt das Problem auf den Punkt. Es ist klar, wie diese Ansicht die gängige Praxis der Organtransplantationen in Frage stellt: Kaum eine moralische Regel ist so weit anerkannt, so unkontrovers und so eingängig wie das Tötungsverbot. Von einigen – moralisch keineswegs unproblematischen – Ausnahmen wie Notwehr, gerechten Verteidigungskriegen oder vielleicht noch der Todesstrafe abgesehen, ist es zweifellos normalerweise moralisch falsch, Menschen zu töten. Wenn Hirntote aber am Leben sind, wie die Anhänger dieser Position betonen, dann werden sie bei der Organentnahme zwangsläufig getötet. Aber die durch die Tötung erreichten Ziele, wie wünschenswert sie abstrakt gesprochen auch sein mögen, sind Notwehr und gerechten Kriegen nicht hinreichend ähnlich (und die Todesstrafe als Strafmaßnahme ja ohnehin irrelevant), als dass sie die Tötung des Organspenders rechtfertigen können. Es ist daher eindeutig unzulässig, an hirntoten Menschen Organexplantationen durchzuführen. Die zweite Position lehnt Organtransplantationen als ungerechtfertigte Tötung ab.

Wie wäre somit die Praxis der Organspende in Deutschland zu beurteilen? Werden Organspender also systematisch getötet? Dass diese Sicht auf die Dinge nicht noch mehr Aufruhr verursacht als sie es tatsächlich tut, liegt sicher auch daran, dass die Organspender (bzw. ihre Angehörigen) derzeit der Organentnahme zustimmen. Niemand wird in Deutschland gegen seinen Willen (bzw. den Willen seiner Angehörigen) Organspender. Daher hat nur selten ein direkt Betroffener berechtigten Anlass, sich über eine erfolgte Organspende zu entrüsten.[29]

Das momentane Verfahren der Zustimmungslösung bedeutet also, so muss es die zweite Position sehen, dass in Deutschland die *Tötung auf Verlangen* gängige und akzeptierte Praxis ist. Die Tötung auf Verlangen ist in Deutschland zwar ebenfalls illegal, aber es ist doch auch klar – wie sich bereits heute an den im Strafgesetzbuch angedrohten Strafen ablesen lässt[30] -, dass eine Tötung auf Verlangen sich sehr erheblich von einem Totschlag oder Mord unterscheidet. Insofern müsste also, wenn trotz Hirntodkritik die Praxis der Organspende beibehalten werden soll, zwar die Dead Donor Rule aufgegeben werden, das allgemeine Tötungsverbot jedoch könnte beibehalten werden. [31] Bei Vorliegen einer gültigen Zustimmung wäre es vielleicht nicht sonderlich tangiert, denn seine Geltung erklärt sich wohl primär aus der klarerweise viel problematischeren ‚unverlangten' also vermutlich unfreiwilligen Tötung, der Tötung *gegen* den Willen des Getöteten.[32] Davon kann bei einer Zustimmungslösung aber nicht die Rede sein.

---

[29] Selbstverständlich gilt dies nur, insofern die Zustimmung gültig ist, also die notwendige Aufklärung geleistet wurde, was in Einzelfällen vielleicht bezweifelt werden kann.

[30] § 216 StGB Töten auf Verlangen: Freiheitsstrafe zwischen sechs Monaten und fünf Jahren, § 211 StGB Mord und § 212 StGB Totschlag: Freiheitsstrafe zwischen fünf Jahren und lebenslänglich, § 213 Totschlag in einem minder schweren Fall: zwischen einem und zehn Jahren.

[31] Darüber hinaus könnte argumentiert werden, dass Tötung auf Verlangen *an einem Hirntoten* ein vielleicht separat zu betrachtender Fall ist.

[32] Natürlich ist es möglich, die Organspende abzulehnen, gerade weil sie eine – eben illegale – Tötung auf Verlangen darstellt. Wie das mit der gesellschaftlichen Wirklichkeit (in der die Organspende staatlicherseits befürwortet wird) in Einklang zu bringen sein soll ist dann jedoch kaum ersichtlich.

### *5.4. Die dritte Position*

Eigentlich wäre dies bereits die dritte Position, die zwar hirntodkritisch ist, die Organspende aber trotzdem nicht ablehnt. Wäre die Idee der Tötung auf Verlangen geeignet, um als Grundlage der derzeitigen Praxis der Organspende zu dienen? Dies scheint aus mindestens vier Gründen mehr als fraglich, die ich hier aus Platzgründen nur erwähnen kann.

Erstens müsste den potentiellen Organspendern vor Augen geführt werden, dass sie, wenn sie denn als Organspender in Frage kommen sollten, noch gar nicht tot sind. Es muss also eine Zustimmung zur Tötung eingeholt werden. Auch wenn sich vermitteln ließe, dass ein Hirntoter keinerlei Bewusstsein hat, nie wieder gesund werden kann und ohnehin bald sterben würde, wenn er nicht intensivmedizinisch behandelt wird, dürfte dies für die gesellschaftliche Akzeptanz der Organspende eine unüberwindbare Hürde darstellen.

Zweitens wäre eine solche Tötung nicht einfach analog zur passiven Sterbehilfe zu sehen, da durch die Organexplantation aktiv in den Körper des Hirntoten eingegriffen wird. Darüber hinaus kommt dieser Eingriff nicht dem Sterbenden, sondern einer anderen Person zugute.

Drittens konfligiert eine so verstandene Praxis sehr deutlich mit dem typischen Selbstverständnis der Ärzte, die sich an Tötungen nicht direkt beteiligen dürfen.

Viertens lässt sich die den Angehörigen derzeit eingeräumte Rolle kaum nachvollziehen. Natürlich kann niemand berechtigterweise über die Tötung einer anderen Person verfügen, auch nicht als sein Angehöriger. Dieser Ansatz könnte also bestenfalls eine enge Zustimmungslösung rechtfertigen. [33]

Es ergibt sich also, dass die Position, die die Tötung auf Verlangen als Grundlage der Organspende ansieht, insgesamt nicht besonders attraktiv erscheint – obwohl sie vielleicht den Vorteil hätte, dass sie die unangenehme Konfiszierungslösung nicht mittragen würde, bei der ja auch ohne Zustimmung explantiert werden könnte. Aber wenn die Transplantationsmedizin nicht rundweg abgelehnt werden soll, bietet sich die Tötung auf Verlangen als naheliegendste alternative Begründung einer legitimen Organentnahme an.[34]

### *5.5. Der Hirntod und das Analogieargument*

Fassen wir die Problematik des fünften Abschnitts zusammen. Hirntodkritiker dürften die von mir vorgeschlagenen Analogien ablehnen. Damit lehnen sie aber entweder die Praxis der Organspende überhaupt ab – wenn sie auf der Dead Donor Rule bestehen – oder sie müssen eine neuartige Rechtfertigung der Praxis der Organspende finden. Sie könnten dann die Dead Donor Rule ablehnen und explizite Zustimmung zur Organspende als Tötung auf Verlangen verstehen, womit das allgemeine Tötungsverbot, das primär auf die Szenarien von Mord und Totschlag abzielt, nicht unbedingt hinfällig würde. Aber Tötung auf Verlangen ist in Deutschland illegal und als explizite Rechtfertigung würde sie selbst nach ihrer Legalisierung neue Probleme aufwerfen, deren Lösbarkeit zumindest auf den ersten Blick noch weniger absehbar scheint als eine allgemein akzeptable Auflösung der Hirntodproblematik. Eine allseits zufriedenstellende Lösung ist nicht zu erwarten, aber es scheint sich doch anzubieten – und zwar gerade unter der Voraussetzung, dass wir die Praxis der Organspende generell befürworten – weiterhin den Hirntod als den Tod des Menschen zu akzeptieren.[35] Dies

---

[33] Bei einer Widerspruchsregelung von Tötung auf Verlangen zu sprechen erscheint absurd, hier wäre also eine solche Legitimierung der Organspende von vornherein unplausibel.

[34] Auch dies ist also kein zwingendes Argument, da es weitere Begründungen geben mag die ich hier nicht diskutiere. Es bietet sich als hypothetische Alternative zur Konfiszierungslösung noch eine hypothetische Rechtfertigung der Organspende, die i) hirntodkritisch ist, ii) Organspende *nicht* als Tötung auf Verlangen begreift und iii) dennoch der Konfiszierungslösung widerspricht.

[35] Dies soll natürlich keine pragmatische Rechtfertigung des Hirntodkriteriums darstellen, das ich als medizinisch begründet ansehe.

bedeutet also, dass meine Analogieargumente weiterhin Bestand haben, insofern die derzeitige Praxis der Organspende überhaupt zu rechtfertigen ist.

## 6. Schlussfolgerungen

Ich habe in diesem Text folgendermaßen argumentiert: Wenn wir Erbschaftssteuern als legitim ansehen und auch Zwangsobduktionen als legitim ansehen, dann sollten wir auch eine Konfiszierungslösung für Organe von Hirntoten befürworten. Dies gilt jedoch nur, wenn wir das Hirntodkriterium als Todeskriterium akzeptieren – was aber, ebenso wie Erbschaftssteuern und Zwangsobduktionen der derzeitigen deutschen Gesetzeslage entspricht – oder das Hirntodkriterium ablehnen, jedoch Organexplantationen an Hirntoten als dennoch zulässig ansehen. Eine philosophische Begründung dafür, dass Erbschaftssteuern, Zwangsobduktionen oder das Hirntodkriterium als legitim anzusehen sind, habe ich aber nicht geliefert.

**Christoph Schmidt-Petri**
Institut für Philosophie
Universität Regensburg
christoph.schmidt-petri@psk.uni-regensburg.de

## Literatur

Bundesärztekammer 2005: „Stellungnahme zur ‚Autopsie' – Langfassung", Berlin, Bundesärztekammer.

Breyer F., W. Van den Daele, W. Engelhard, G. Gubernatis, H. Kliemt, C. Kopetzki, H. J. Schlitt, und J. Taupitz 2006: *Organmangel: Ist der Tod auf der Warteliste unvermeidbar?* Berlin, Springer.

Brinkmann, A., A. Du Chesne, und B. Vennemann 2002: „Aktuelle Daten zur Obduktionsfrequenz in Deutschland", *Deutsche Medizinische Wochenschrift* 127, 791-795.

Fabre, C. 2006: *Whose Body is it Anyway? Justice and the Integrity of the Person*. Oxford: Oxford University Press.

Deutsche Stiftung Organtransplantation 2012: *Organspende und Transplantation in Deutschland. Jahresbericht 2011,* Frankfurt/Main, DSO.

Darwall, S. 1977: „Two Kinds of Respect", *Ethics* 88, 36-49.

Hershenov, D. und J. Delaney 2009: „Mandatory Autopsies and Organ Conscriptions", *Kennedy Institute of Ethics Journal* 19, 367-91.

Johnson E., und D. Goldstein 2003: „Do Defaults Save Lives?" *Science* 302, 1338-39.

Nozick, R. 1974: *Anarchy, State and Utopia.* New York, Basic Books.

Nussbaum, M. 2000: *Women and Human Development*, Cambridge, Cambridge University Press.

Rawls, J. 1971: *A Theory of Justice*, Cambridge MA, Harvard University Press.

Truog, R. 1997. „Is it Time to Abandon Brain Death?", *Hastings Center Report* 27, 29-37.

Schmidt-Petri, C. und F. Himpsl 2012: „Zellfrei, gefriergetrocknet - Knochenmehl, Haut, Achillessehnen: Was geschieht eigentlich mit Gewebespenden?", *Süddeutsche Zeitung*, Feuilleton, 31.5.2012, S. 19.

Schmidt-Petri, C. 2012: „Der mutmaßliche Wille im deutschen Transplantationsgesetz“, in *Ethics-Society-Politics*, Martin G. Weiss und Hajo Greif (Hg.), Kirchberg/Wechsel: ALWS, 300ff.

Schmidt-Petri, C. 2013: „Mandatory Autopsies, Organ Confiscations and the Definition of Death“, Manuskript, Universität Regensburg.

Steinbock, B. und Norcross A. (Hrg.) 1994: *Killing and Letting Die*, Cambridge, Cambridge University Press

Statistisches Bundesamt 2012a: *Finanzen und Steuern. Erbschaft- und Schenkungssteuer*, Wiesbaden, Statistisches Bundesamt.

Statistisches Bundesamt 2012b: *Gesundheit. Todesursachen in Deutschland*, Wiesbaden, Statistisches Bundesamt.

Stoecker, R. 2010: *Der Hirntod. Ein medizinethisches Problem und seine moralphilosophische Transformation. Studienausgabe*. Freiburg/München: Alber.

# Two Problems with the Socio-Relational Critique of Distributive Egalitarianism[1]

Christian Seidel

Distributive egalitarians believe that distributive justice is to be explained by the idea of distributive equality (DE) and that DE is of intrinsic value. The socio-relational critique argues that distributive egalitarianism does not account for the "true" value of equality, which rather lies in the idea of "equality as a substantive social value" (ESV). This paper examines the socio-relational critique and argues that it fails because – contrary to what the critique presupposes –, first, ESV is not conceptually distinct from DE, and second, the idea of ESV cannot serve as a "foundation" or "root" of distributive egalitarianism.

Distributive egalitarianism is a view about distributive justice. Proponents of this view believe in at least two claims: First, that a conception of distributive justice is (at least partially) to be explained by the idea of distributive equality (DE); and second, that DE is of *intrinsic* (and not purely instrumental) value. This view is subject to many objections. Some critics – most prominently, perhaps, Elizabeth Anderson (1999) and Samuel Scheffler (2010a,b) – have pointed out that distributive egalitarianism fails to account for the "true" value of equality. The true point of equality, they claim, is the idea of "equality as a substantive social value" (ESV) which essentially manifests itself in certain kinds of egalitarian social relationships. The present paper examines this "socio-relational critique", starting with an explanation of the main claims of distributive egalitarianism (section 1) and a brief summary of the socio-relational critique (section 2), it argues that this critique faces two related problems (section 3): The first arises once we ask what the relation between DE and ESV is supposed to be and I will argue that – contrary to the critics' claim – ESV is *not* independent of DE, but actually *presupposes* some form of DE such that DE is both conceptually and explanatory prior to ESV. The second problem concerns the way in which one may argue for DE; this is a question about the justificatory or argumentative structure of egalitarian conceptions of distributive justice and I will argue that – again, contrary to the critics' claim – the idea of ESV cannot serve as a "foundation" or "root" of DE: There is just no way to both provide a vindication of DE based on ESV *and* to hold a genuinely *egalitarian* view about distributive justice. These problems are not insurmountable, but – if the reasoning is sound – they suggest that the socio-relational critique fails to be a decisive refutation of distributive egalitarianism; they also suggest that the underlying conception of "equality as a substantive social value" is seriously underdeveloped and needs further refinement.

## 1. A Framework for Distributive Egalitarianism

As with virtually any "-ism"-terms, there is considerable disagreement about what the term "egalitarianism" really stands for. In order not to beg the question against the distinction

[1] This paper has benefited from questions, comments and suggestions by Gerhard Ernst, Iwao Hirose, Sabine Hohl, Christian Kietzmann, Weyma Lübbe, Sebastian Muders, Juri Viehoff and Gabriel Wollner. I would like to thank them, and especially Sam Scheffler, who took the time to thoroughly discuss the issues raised here and to respond to them.

between DE and ESV, I will confine myself to the term "distributive egalitarianism". To be a distributive egalitarian is, first, to hold a conception of distributive justice – a conception in which distributive justice is (at least partially) *explained* by the notion of distributive equality (DE). Hence a view fails to qualify as distributive egalitarian if either it is not at all concerned with distributive justice (but with, say, political liberty or moral rightness in general) or if it does not explain distributive justice in terms of distributive equality. Second, a distributive egalitarian values DE *intrinsically* and not just as a means to some other value. As is well known, competing accounts of distributive justice like prioritarianism, sufficientarianism or utilitarianism (understood as a view about just distributions) may also concede that distributive equality is of instrumental value: equality in the distribution of some good may, for contingent reasons, be the best means to give priority to the worst off, to make sure that everyone has enough or that total utility is maximised. Nevertheless, none of these conceptions values equality for its own sake. So the second claim is crucial in that it distinguishes distributive egalitarianism from rival conceptions of distributive justice.

So, as a first approximation, distributive egalitarianism is committed to at least two claims:

(T1) Distributive justice is to be explained in terms of DE.

(T2) DE is of intrinsic value.

I say "at least" because there may be more to distributive egalitarianism than these two claims, e.g. distributive egalitarians *may* also be committed to ESV; but any conception of distributive justice which does not subscribe to both claims is not distributively egalitarian. Moreover, note that the first claim does not say that DE *fully* explains distributive justices; this leaves open the "pluralist option", i.e. a conception of distributive justice in terms of several values (only one of which is DE); to the extent that a pluralist partially explains distributive justice in terms of DE and also believes in the intrinsic value of DE, he may be called a pluralist distributive egalitarian.

As they stand, (T1) and (T2) are not very specific; the first claim probably means that a distributive egalitarian defines "Distribution *D* is just" by a sentence which, *inter alia*, contains the condition "in *D*, everyone has equal amount of *C*". Here *C* is the currency (or set of currencies) of justice, i.e. the good which matters for the distributive egalitarian; *C* may be a resource, a moral right, well-being, social recognition, advantages, opportunities, capabilities or whatever (as long as it is such that two people have equal amount of it). The second claim is more tricky because it is notoriously difficult to make sense of intrinsic values. Here is a proposal:

(T2*) The fact that a distribution *D* is unequal in currency *C* is, *in itself*, a reason (of type *T* and weight *W*) that counts against distribution *D*.

I am not going to defend the claim that (T2*) is a proper explication of "valuing DE intrinsically". I think it is (although I happily concede that (T2*) may not be the whole story about valuing DE intrinsically; at least, it is part of that story). Instead of defending (T2*), let me try to at least explain it: First, it says that a distribution's inequality is *in itself* a reason against that distribution; the contrast is with a conception which holds that a distribution's inequality is a reason against that distribution *insofar as* or *to the extent that* or *given that* or *under the supposition that* or *by virtue of the fact that* this undermines some other value. In the latter claim, DE would be of derivative (or instrumental) value only and some other value would do the normative work. Second, (T2*) also qualifies the nature (or type) and the weight of the reason. This is mainly to account for the fact that distributive egalitarianism may come in different strengths:[2] Some may hold that the reason constituted by a distribution's inequality is overriding (this may be dubbed "strong distributive egalitarianism"), others may

[2] Cp. Cohen's (1989: 908) distinction between strong and weak *equalisandum* claims.

allow for trade-offs with other reasons (so again, (T2*) allows the pluralist to enter the stage). The equality-grounded reason may be of different types in that it may have various structural features (e.g. it may be a "silencing" or an "exclusionary" reason), although it will probably be a moral (rather than an aesthetic or prudential) reason. For the sake of the present argument, we do not have to bother with these details. Third, (T2*) can be specified with a whole lot of currencies of justice (see the remark above). What currency we will plug in for *C* will of course affect the plausibility of the resulting position; but it is not constitutive of being a distributive egalitarian that one is committed so some specific currency; all that is needed is that there is *some* currency with regard to which unequal distributions count as reasons against the distribution.

For the present purposes, (T1) and (T2*) provide a suitable explication of distributive egalitarianism. Indeed I think that this explication is quite uncontroversial; in particular, it should be acceptable to all critics of distributive egalitarianism.

## 2. The Socio-Relational Critique

Distributive egalitarianism, is subject to many objections. Perhaps the most widely discussed problem is the so-called "levelling-down" objection; it purports to show that (T2*) is false, i.e. that distributive equality is not intrinsically valuable. I am not going to deal with this objection.

A second prominent problem for distributive egalitarians concerns (T1); on a strong reading of (T1) – a view that may be called "pure" distributive egalitarianism –, DE is all that is needed to account for distributive justice; i.e. distributive justice is to be explained in terms of DE *only*. The quite common problem for pure distributive egalitarianism is to account for the proper place of individual responsibility in our intuitions about distributive justice. It seems that a simple "everyone gets the same, equal amount of the currency of justice" does not adequately reflect the idea that people sometimes deserve more or less of whatever is up for distribution simply because of good (clever) or bad (foolish) choices they made. In response, distributive egalitarians tried to "incorporate" individual responsibility into their accounts, thereby creating various forms of "luck-egalitarianism". Luck-egalitarianism is the main target of the socio-relational critique since it is – rightly or wrongly – often regarded as the most plausible form of distributive egalitarianism. *That* it is, indeed, a form of distributive egalitarianism can be explained within the proposed framework: The luck-egalitarian distinguishes between two kinds of inequalities – acceptable and unacceptable inequalities. She does so by some further distinction which she takes to be normatively relevant: the distinction between features for which a person can be held responsible (like autonomous choices) and features for which a person cannot be held responsible (like being born with a handicap). The idea is that inequalities in something (resources, advantages, well-being, ...) which are due to features for which a person can be held responsible are acceptable while inequalities which are due to features for which a person cannot be held responsible are not acceptable. Where exactly the line should be drawn (i.e. which features we can be held responsible for) is a matter of dispute within the luck-egalitarian camp.[3] Disregarding this dispute, what makes luck-egalitarianism a version of distributive egalitarianism in the sense defined above is that first, it tries to explicate distributive justice in terms of *certain (in)equalities* – namely (in)equalities in the distribution of resources, advantages, well-being, ...which are due to features for which someone cannot be held responsible; and second that luck-egalitarians, too, think that equality with regard to something (namely with regard to resources, advantages, well-being, etc. which derive from features for which one cannot be

[3] For instance, Dworkin (2002) and Cohen (1989) disagree about whether we can be held responsible for some of our expensive tastes or character traits.

held responsible) is intrinsically valuable – and conversely that there is something intrinsically bad about some forms of inequalities. This means that the luck-egalitarian will subscribe to the following thesis:

($T2_L$*) The fact that a distribution *D* is unequal in those *C*s (resources, advantages, well-being,...) which derive from features for which persons cannot be held responsible is, *in itself*, a reason (of type *T* and weight *W*) that counts against *D*.

Hence in the proposed framework, luck-egalitarianism turns out to be a version of distributive egalitarianism with a special currency.

Suppose that some version of ($T2_L$*) indeed solves the problem of accounting for the proper place of individual responsibility in our intuitions about distributive justice. Now proponents of the socio-relational critique argue that luck-egalitarianism is still bound to fail.[4] The critique's strategy is, first, to direct attention to the idea of equality as a social value (ESV) as something utterly different from distributive equality (of either the luck-egalitarian or some other type), and, second, to argue that *this* idea of ESV, rather than DE, is the "true" point of equality. As Jonathan Wolff puts it:

> [T]he basic idea is [that] an equal society is one that has the right quality of relations between individuals, rather than one which distributes the 'currency' of justice, the right way. (Wolff 2007: 135)

If this critique was successful, it would be an attack on (T1), since it implies that what – ultimately – accounts for distributive justice is ESV rather than DE.[5] In this respect (in attacking (T1)), the socio-relational critique is an extension of or a follow-up to the aforementioned criticism that distributive egalitarianism fails to account for the proper role of responsibility in our judgements about distributive justice.

A very forceful and instructive instance of the socio-relational critique has been put forward by Samuel Scheffler in his pair of articles "What Is Egalitarianism?" and "Choice, Circumstance, and the Value of Equality". If I understand this critique correctly, it consists of three parts:

(C1) There is some ideal of equality as a social value (ESV) which is distinct from the ideal of distributive equality (DE).[6]

(C2) Any vindication of DE has to be based on ESV.[7]

(C3) Luck-egalitarianism cannot be rooted in ESV. [8]

It is important to see that the the socio-relational critique needs the first premise (C1) to get off the ground; indeed, the very idea of the critique is to argue for a re-orientation in the distributive egalitarian's project: Any vindication of DE should begin from and do justice to a quite *different* idea, namely ESV. But if ESV was not a distinct idea but just a variation of DE,

---

[4] In "What Is Egalitarianism?", Scheffler entangles this critique with the claim that luck-egalitarianism cannot be regarded as an extension of Rawls's arguments in *A Theory of Justice* (Scheffler 2010b).

[5] So although the primary target of the socio-relational critique is luck-egalitarianism, the critique easily generalises to other types of distributive egalitarianism.

[6] Cp. "Equality, as it is more commonly understood, is not, in the first instance, a distributive ideal" (Scheffler 2010b: 191); "Equality as a social and political value expresses an ideal of how human relationships should be conducted. That ideal has distributive implications, and the task for an egalitarian conception of distributive justice is to draw out those implications" (Scheffler 2010a: 232).

[7] Cp. "any form of distributive egalitarianism, if it is to be persuasive, must be rooted in a more general conception of equality as a moral value or normative ideal" (Scheffler 2010b: 178).

[8] Cp. "the luck-egalitarian conception of equality diverges from a more familiar way of understanding that value [i.e. ESV]" (Scheffler 2010b: 191); "the most serious reason for declining to ground egalitarianism in the principle of responsibility [as luck-egalitarians do] is that to do so is to lose touch with the value of equality itself [i.e. ESV]" (Scheffler 2010a: 225).

then the "critique" would amount to saying that luck-egalitarians have the wrong conception of *distributive* equality. That is, without (C1), the socio-relational critique would be entirely *internal* in the sense that it would amount to a dispute between different conceptions of DE; instead, I take this critique to be *external* in the sense that it suggests that a successful vindication of DE has to "look outside" and to resort to some further argumentative resource not already given by the idea of DE.

As is already obvious from the fact that ESV figures in all three premises, the strength of the argument heavily depends on what, precisely, ESV consist in. The set-up so far already suggests that there are three structural constraints on any substantive characterisation of the ideal of "equality as a substantive social value": First, ESV must be distinct from DE, as (C1) holds. Second, ESV must somehow figure in a proper account of the idea of distributive justice (and not in an account of some other moral value). This follows from the fact that luck-egalitarianism is the target of the ESV-based socio-relational critique: Luck-egalitarianism *is* a conception of distributive justice and if the critique is supposed to show that this conception is misguided because it does not properly account for ESV, then a convincing conception of distributive justice must account for ESV. Third, ESV must not only be distinct from DE, but also from what one might call "moral equality" – the idea of equality in basic moral status which is taken to ground human rights or taken to be constitutive of membership in the moral community.[9] This is simply because *any* conception of distributive justice – be it sufficientarian, prioritarian, egalitarian etc. – will allow for moral equality; hence, without further specification, the ideal of "moral equality" does not *per se* differentiate and, therefore, not settle the dispute between competing conceptions of distributive justice. So if ESV was equivalent to "moral equality", it could not have the implications that the socio-relational critiques takes it to have – namely to refute a specific account of distributive justice.

These remarks suggest that a substantive characterisation of ESV must look for a middle ground: On the one hand, "equality as a substantive social value" must be "more" (specific) than basic moral equality, but on the other hand, it must be "less" (specific) than distributive equality. So how do proponents of the socio-relational critique characterise ESV? Unfortunately, this is a place where the critics' argument becomes a bit evasive; what follows is a collection of quotations from Scheffler's papers which are supposed to shed some light on the idea of ESV:

> [Equality] is, instead, a moral ideal governing the relations in which people stand to one another. (Scheffler 2010b: 191)
>
> [ESV] claims that human relations must be conducted on the basis of an assumption that everyone's life is equally important, and that all members of a society have equal standing. (Scheffler 2010b: 191)
>
> As a social ideal, it holds that a human society must be conceived of as a cooperative arrangement among equals, each of whom enjoys the same social standing. (Scheffler 2010b: 191)
>
> Insofar as equality is understood as a substantive social value [...], the basic reason it matters to us is because we believe that there is something valuable about human relationships that are, in certain crucial respects at least, unstructured by difference of rank, power or status. (Scheffler 2010a: 225)
>
> [I]negalitarian societies compromise human flourishing; they limit personal freedom, corrupt human relationships, undermine self-respect, and inhibit truthful living. By contrast, a society of equals supports the mutual respect and the self-respect of its members, encourages freedom of interpersonal exchange, and places no special

[9] Cp. Scanlon's (2002: 41) contrast of "substantive" and "formal" equality.

> obstacles in the way of self-understanding or truthful relations among people. (Scheffler 2010a: 227)
>
> Equality as a social and political value expresses an ideal of how human relationships should be conducted. (Scheffler 2010a: 232)

So the main idea seems to be that ESV is concerned with relationships between humans and that these relationships are governed by some generic conception of encountering others "as equals". With regard to *this* idea, the critics claim that (1) it is distinct from the idea of DE, (2) that it is the basis of any successful vindication of DE, and (3) that luck-egalitarians cannot use it as a basis to support their position.

I think there are two serious problems with this route of critique; but before I address them in the following section, let me explain the normative implication of the critique: If successful, the socio-relational critique would be able to support a restriction of the *scope* of claims of distributive justice to contexts with certain social relationships – namely to those contexts in which it is at least possible to "encounter each other as equals". It is not entirely clear, for instance, whether (and how) the idea of "encountering each other as equals" applies to injustices that arise from global economic institutions. Is it possible to describe the normative problem of global inequalities in terms of categories like "truthful living", "respect and disrespect", "self-understanding" which are supposed to explicate the content of ESV and which are most clearly rooted in the context of direct interpersonal relations (e.g. between spouses, relatives, friends, neighbours, colleagues etc.)? If not then it seems that proponents of the socio-relational critique will not (be able to) provide a rationale for global redistribution in the face of global inequality; rather they could point out that from the point of distributive justice, there are no claims for redistribution as long as the relevant human relationships do not exist.[10] The situation is very different for distributive egalitarians: They may easily hold that global inequality (in the relevant currency) provides, in itself. a reason for global redistribution – after all, they believe in the intrinsic value of equality and the intrinsic disvalue of inequality, and this commitment is not necessarily qualified in scope.

So the success of the socio-relational critique may have implications for the scope of claims of distributive justice. This is why we should be sure about whether or not the socio-relational critique succeeds. In the following section, I will argue that it does not.

## 3. Assessing the Socio-Relational Critique

The socio-relational critique does not succeed for two reasons. The first is that I can see no way of making sense of equality as a social value *without* resorting to the idea of distributive equality. In other words: Claim (C1) is false because DE and ESV are not completely distinct ideas. The second problem is that – even if ESV was a distinct idea – the project of anchoring distributive egalitarianism in ESV is bound to fail. In other words: Claim (C2) is false, too.

### *3.1 DE is Conceptually and Explanatory Prior to ESV*

Let me elaborate on the first problem. While reading the collection of quotations you may have already been wondering whether the given characterisations of ESV are neutral with respect to the idea of *distributive* equality. For instance, the claims that "all members of a society have equal standing" (second quotation), that "each [...] enjoys the same social standing" (third quotation) or that "human relationships [...] are, in certain crucial respects at least, unstructured by difference of rank, power or status" (fourth quotation) seem to be a

[10] Of course, they will be able to account for some duties related to global inequalities, e.g. a duty to assist the starving poor; but these duties would not be duties of distributive justice, but rather duties grounded in some other value (like the value of human life).

characterisation of ESV in terms of distributive equality: The first and second say that the distribution of social standing is equal; and the converse of the third is that human relationships are (or rather ought to be), in at least some respects, governed by equality of rank, power and status. So here ESV is obviously characterised in terms of DE, although the "currency" is unusual in that it is neither opportunities nor capabilities nor resources nor well-being but social standing or some form of rank, power or status. My point is not that ESV is said to have some distributive egalitarian *implications*; the point is that the *very idea* of ESV employs the idea of DE: The concept of ESV is explained with reference to the concept of DE. And this means that, contrary to the critics' claim, ESV is *not* distinct from DE – instead, DE seems explanatory prior to ESV.

The critic may reply that this merely shows that the given explications of ESV are infelicitous in so far as – as a matter of fact – they are not neutral with respect to DE; but, she continues, this does not show that it is *impossible* to give an explication of ESV which is neutral with respect to DE. But this reply will not work. First of all, the burden of proof is on the critic's side – *she* has to show that it is possible to give a characterisation of ESV that is neutral with respect to DE. But more importantly, the following argument shows that this burden cannot be met: To give a characterisation of ESV that is neutral with respect to DE is to give an explication of ESV that is not conceptually tied to the idea of distributive equality. That is, in principle, it would have to be conceivable that we can apply the concept of equality as a social value in a situation in which there is *no respect whatsoever* with regard to which a distribution between two or more people can be said to be equal. This, I submit, is impossible: You cannot use the concept of difference unless two things are different *in at least some respect*; similarly, you cannot talk of equality without some people being equal in at least some respect, that is: without some people being equal *with regard to something*. But to say that some people are equal with regard to something is to say that this something is distributed equally between these people. It follows that the concept of ESV – if it concerns equality at all – *presupposes* the concept of DE. And if DE is indeed conceptually prior to ESV, then ESV is not distinct from DE but can, at least partially, be accounted for in terms of DE.

The critic may concede the point and reply that it is just a dispute about words: "The label 'equality as a social value' is a misnomer – it is not about equality at all. Let's call it 'the ideal of how human relations should be conducted' (cp. the sixth quotation from above) or 'the ideal of freedom'. Then the objection is blocked". Well, yes and no. First note that in conceding that ESV really concerns something quite distinct from equality, the critic simply changes the subject. If, for instance, ESV is just a name for the ideal of "how human relations should be conducted", then the requirements supported by this ideal – e.g. "treat each other with mutual respect" – would be requirements of morality in general rather than requirements of distributive justice. So ESV would no longer serve the purpose of accounting for distributive justice – a violation of the second structural constraint on any substantive characterisation of ESV noted above – and in changing the subject, the socio-relational critique looses much of its appeal: It is undeniable that there are pressing moral questions which are *essentially distributive* in character, i.e. moral questions which are about the fair distribution of burdens and benefits: How should organs be allocated among the needy? What is a fair distribution of wealth within a state? How should we share the burdens of combating climate change? Distributive egalitarians try to tackle these questions. Whether their answer is satisfactory, is of course up for dispute. But this dispute is a dispute about distributive justice. If the critics change the subject and take ESV to be a general moral ideal not specifically concerned with questions of distributive justice, these questions will be left unanswered. Instead of establishing an attractive alternative account of distributive justice,

the socio-relational critique would simply sidestep essentially distributive questions.[11] Moreover (and consequently), if ESV is really about something else (like freedom or valuable human relationships), then not undermining this value would be a condition of adequacy for *any* conception of distributive justice and not just for distributive egalitarianism. Sufficientarians and prioritarians, too, may also accept that we should treat each other with equal respect; but their agreement on *this* understanding of ESV would not settle their substantive disagreements about what constitutes a fair distribution. In other words: By re-stating ESV in this way, ESV ceases to have any bearing on distributive questions – it would be "distributively inert". Again, the socio-relational critique looses much of its appeal: If taking this ideal into account is a general criterion for any plausible conception of justice, why should the (luck-)egalitarian be especially worried about it?

A last reply to my line of reasoning may be called the "stipulative currency objection": "Of course," the critic may say, "we *may* talk about distributing *something* (e.g. rank, power, status, social standing) equally; but it would be completely stipulative to do this, a purely formal and artificial way to plug in something as a currency just for the sake of having a currency." I think that the stipulative currency objection is either self-refuting or overgeneralises. Regarding the first horn of this dilemma, note that "artificial" cannot mean "unimportant", "unwarranted" or "not being supported by further reasons". The currencies under consideration (e.g. rank, power, status, social standing) clearly *are* important from the critic's point of view, because these are the very concepts employed to characterise the content of ESV; and while it is true that no reasons are given why differences in the currencies under consideration (differences in e.g. rank, power, status, social standing) are supposed to be relevant, this charge would, *mutatis mutandis*, apply to ESV as well because no further reasons are given why certain human relation ought not to be governed by differences in rank, power, status, social standing etc. Let us turn to the second horn: The stipulative currency objection may also be read as complaining not so much about *what* is regarded as a currency or *distribuendum* but rather as objecting to the fact *that* something is regarded as a *distribuendum* in the first place. The objection will be then be that it is misleading to think of rank, status, power etc. as something like apples, newspaper or spam mails which may be literally "distributed" by a distributor who, at some point in time, has all of the *distribuendum* at her disposal and then hands it out to the recipients. But this objection is too strong; if valid, it would apply to all *distribuenda* which are discussed in the debate on distributive justice (e.g. welfare, resources, opportunities): Of course, no single entity, person or institution possesses (or ever possessed) all of the welfare, resources or opportunities and then, after careful thinking, hands it out to the people. The existence of such a distributor or "social planner" is not necessary for questions of distributive justice to arise since we can also be said to "distribute something" if we have some sort of causal influence on the distribution of something by shaping our social institutions. For instance, as long as we are able to influence the distribution of income through taxes, expropriation, education or whatever, we may call any action that so affects the distribution of income an act of (re)distributing income. And given that, through institutional arrangements, we surely *do* have such a causal impact on differences in rank, status, power, social standing etc., there is nothing artificial or stipulative in talking about the *distribution* of rank, status, power or social standing. I thus conclude that the stipulative currency objection fails and that the first problem with the socio-relational critique is still unsolved: ESV is not distinct from DE but rather presupposes some from of DE.

[11] This answer to the critic naturally invites the following question: What distinguishes "a question of distributive justice" from a "question of morality in general"? While I am not able to give a fully-developed answer, a plausible first approximation is that *paradigmatic* questions of distributive justice arise where (1) a group of persons benefits from (2) some important good which (3) is provided by a (possibly different) group of persons and (4) whose distribution can intentionally be affected (either directly or through social institutions).

### *3.2 It Is Impossible to Anchor Distributive Egalitarianism in ESV*

But maybe I am totally wrong and there is some way to make sense of the idea of equality as a social value without resorting to the idea of distributive equality. So suppose that ESV really were an idea distinct from DE. Then there is a second problem, which concerns the alleged justificatory relation between ESV and DE (claim (C2)). As Scheffler puts it,

> unless distributive egalitarianism is anchored in some version of that ideal, or in some other comparably general understanding of equality as a moral value or normative ideal, it will be arbitrary, pointless, fetishistic. (Scheffler 2010b: 192)

My problem with this claim has two parts. The minor part is that I am less optimistic that ESV is of the right kind to "anchor" DE. Once we treat ESV as an idea really distinct from DE, it becomes less clear that the distributive implications ESV is said to have are of any use to the distributive egalitarian. If, for instance, ESV really is about the absence of oppressive, corrupt, humiliating, freedom-undermining or self-respect-undermining relationships (cp. the fifth quotation in the collection mentioned above), then it seems that ESV is fully compatible with a range of conceptions of distributive justice; again, prioritarians or sufficientarians may equally well subscribe to ESV in this sense. So once we throw out anything in the concept of ESV which has to do with DE, it is no longer clear why it should be of any use for vindicating distributive egalitarianism. And to the extent that we allow DE to be part of the concept of ESV, this problems of course vanishes, but then ESV ceases to be a distinct idea.

The major part of my second problem with the proposal quoted above is less obvious, but probably more severe: Suppose that ESV really were an idea distinct from DE and suppose also that I am wrong with my last criticism; that is, let us assume that ESV really has distributive implications for a distributive egalitarian. Then, as I see it, the requirement to "root" or "anchor" DE in ESV – claim (C2) – and the thesis (T2*) are incompatible; that is, there is no way to provide a vindication of DE based on ESV *and* to stay a distributive egalitarian. Or, alternatively: By rooting DE in ESV, one ceases to be a distributive egalitarian.

Here is the argument: To anchor distributive egalitarianism in the idea of ESV is to provide an argument that supports distributive egalitarianism and employs ESV and its distributive implications as a premise. Supporting distributive egalitarianism by an argument is, *inter alia*, to provide an argument which has (T2*) as its conclusion (of course, one also need an argument with (T1) as its conclusion but I will leave this aside). So the task is to construct an argument which has (T2*) as its conclusion and which makes use of ESV and its distributive implications. To make use of ESV means to start from the fact that ESV is of intrinsic value, that is, to use something like the following premise:

(P1) If *D* undermines ESV, this fact is *in itself* a reason that counts against *D*.

And to make use of ESV's distributive implications for distributive egalitarianism means to employ a conditional which connects ESV and the equal distribution of something, i.e.:

(P2') If a distribution *D* satisfies ESV, then it is equal in currency *C*.

or, by taking its converse,

(P2) If a distribution *D* is unequal in currency *C*, then it undermines ESV.

So the argument needed to root or anchor DE in ESV takes the general form:

(P1) If *D* undermines ESV, this fact is *in itself* a reason that counts against *D*. (= ESV as an intrinsic value)

(P2) If a distribution *D* is unequal in currency *C*, then it undermines ESV. (= distributive implications of ESV)

(C) If a distribution *D* is unequal in currency *C*, the fact that *D* undermines ESV is, *in itself*, a reason that counts against *D*.

Hence all that follows from ESV and its distributive implications is (C), which says that the reason which counts against the distribution is constituted by its *undermining ESV*. In particular, it is *not* the distribution's inequality which provides the reason. That is, (C) is incompatible with

(T2*) The fact that a distribution *D* is unequal in currency *C* is, *in itself*, a reason (of type *T* and weight *W*) that counts against distribution *D*.

because if (C) is true, then it is not the distribution's inequality *in itself* which constitutes the reason; rather, the distribution's inequality is a reason only *insofar as* or *by virtue of the fact that* or *to the extent that* this undermines ESV. To put it a bit less precise but more suggestive: Once we engage in the socio-relational project of anchoring DE in ESV, DE will be of derivative (or instrumental) value only, but it will no longer be intrinsically valuable. But what is distinctive of the distributive egalitarian is that she believes in the *intrinsic* value of distributive equality. So one has to choose: Either engage in the socio-relational project or stay a distributive egalitarian. But one cannot have both on pain of inconsistency.[12]

Of course, showing that you cannot both believe in the intrinsic value of distributive equality and vindicate distributive equality by the idea of equality as a social value leaves open the separate question which of the two incompatible options one should choose. Socio-relational accounts seem to go for the latter. But this will only work if we can make sense of the ideal of ESV. My first worry in section 3.1 was that this may not be the case.

## 4. Conclusion

So what are we left with? I argued that the first and second part of the three-step socio-relational critique of (luck-)egalitarianism fail. That is, neither is there some ideal of ESV which is distinct from the ideal of DE; nor has any vindication of distributive egalitarianism to be rooted in this ideal of ESV. Along the way, I suggested that "equality" is essentially a distributive ideal: Since you cannot talk about equality without some people being equal with regard to something, equality always consists in the equal distribution of something.[13] While this claim certainly needs a more thorough defence, it poses a challenge to the proponents of the socio-relational critique: They have to further elaborate on the idea of equality as a substantive social value and have to give a substantial characterisation which, on the one hand, avoids any reference to the idea of distributive equality and which, on the other hands, is recognizable as an ideal of *equality* (rather than an ideal of something else). Whether or not this challenge can be met remains to be seen.[14]

---

[12] This point has already been noted by Thomas Nagel: "[T]he defense of economic equality [i.e. distributive equality] on the ground that it is needed to protect political, legal, and social equality may not be a defense of equality *per se* – equality in the possession of benefits in general. Yet the latter is a further moral idea of great importance. Its validity would provide an independent reason to favor economic equality as a good in its own right" (Nagel 2002: 61).

[13] This conceptual claim is, of course, fully compatible with the evaluative claim that equality is not intrinsically valuable or the critique that it is a "harmful" ideal (cp. e.g. Frankfurt 1988).

[14] In a recent paper, Samuel Scheffler has responded to the objections presented here by pointing to the "practice" of equality (Scheffler 2012): Egalitarian relations often manifest themselves in the way persons practically deliberate. While I think that this a promising route, the challenge is to characterise "egalitarian practical deliberation" in a way that cannot be rephrased in distributive terms. For instance, if egalitarian practical deliberation consisted in giving "equal" weight to "equally important" interests (whether yours or mine), DE would re-enter through the back-door.

**Christian Seidel**
Friedrich-Alexander-Universität Erlangen-Nürnberg
christian.seidel@fau.de

## References

Anderson, E. (1999): 'What Is the Point of Equality?', *Ethics* 109, 287–337.

Cohen, G. A. (1989): 'On the Currency of Egalitarian Justice', *Ethics* 99, 906–44.

Dworkin, R. (2002): *Sovereign Virtue. The Theory and Practice of Equality*. Cambridge, MA/London: Harvard University Press.

Frankfurt, H. G. (1988): 'Equality as a moral ideal', in *The importance of what we care about. Philosophical essays*, Cambridge/New York/Melbourne: Cambridge University Press, 134–58.

Nagel, T. (2002): 'Equality', in M. Clayton, and A. Williams (eds.): *The Ideal of Equality*, Houndmills: Palgrave Macmillan, 60–80.

Scanlon, T. M. (2002): 'The Diversity of Objections to Inequality', in M. Clayton, and A. Williams (eds.): *The Ideal of Equality*, Houndmills: Palgrave Macmillan, 41–59.

Scheffler, S. (2010a): 'Choice, Circumstance, and the Value of Equality', in: *Equality and Tradition. Questions of Value in Moral and Political Theory*, Oxford/New York: Oxford University Press, 208–35.

— (2010b): 'What Is Egalitarianism?', in: *Equality and Tradition. Questions of Value in Moral and Political Theory*, Oxford/New York: Oxford University Press, 175–207.

— (2012): 'The Practice of Equality', unpublished manuscript presented at the workshop "On the Scope of Distributive Justice", Central European University, 5 July 2012.

Wolff, J. (2007): 'Equality: The Recent History of an Idea', Journal of Moral Philosophy 4, 125–36.

# The Role of Economic Analysis in Combating Climate Change

Joachim Wündisch

This paper is concerned with the appropriate role of economic analysis in the context of the climate change debate. It seeks to offer a careful evaluation of the economists' tools and the underlying philosophy of economics in order to build a limited defense of economic analysis against common criticisms voiced by philosophers.

## 1. Introduction

Anthropogenic climate change is one of humanity's most important and urgent challenges. If we fail to intervene with conviction, anthropogenic climate change will radically reshape our biosphere wreaking havoc on human and animal welfare. While lagged effects make reversing or even halting climate change unattainable for now, there are feasible strategies to slow climate change and to avert catastrophe. The most salient but also one of the most costly of these strategies is a significant reduction of global $CO_2$ emissions.

Neither the reality of climate change nor its effects are disputed by credible sources. However, the question of how to evaluate the effects of climate change as well as the question of how and to what extent to reduce $CO_2$ emissions is the subject of serious debate. Due to the success of economists in influencing policy debates, the tools of their trade are ever present in discussions of public policy. Their critics, though, claim that the contributions economists can make to public policy debates are vastly overrated. Specifically, some claim that underlying philosophical perspectives need to be more carefully considered when making public policy decisions in general and decisions about combating climate change in particular.

This paper is concerned with the appropriate role of economic analysis in the context of the climate change debate. It seeks to offer a careful evaluation of the economists' tools and the underlying philosophy of economics in order to build a limited defense of economic analysis against common criticisms voiced by philosophers. Toward this end section two introduces a particularly pointed statement of these criticisms and summarizes them. Sections three, four, and five defend economic analyses against these criticisms before section six concludes.

## 2. Management Approaches to Climate Change

In his article "Ethics, Public Policy, and Global Warming" the influential environmental ethicist Dale Jamieson delivers a scathing critique of what he calls management approaches. These management approaches are said to be based on

> [m]anagement techniques mainly [...] drawn from neoclassical economic theory and [...] directed toward manipulating behavior by controlling economic incentives through taxes, regulations, and subsidies. (Jamieson 1992: 142)

Examples of these management approaches within the climate change debate are taxes on $CO_2$ emissions, outright bans on particularly wasteful technologies such as the incandescent

light bulb, subsidies for electric cars, and, of course, cap-and-trade systems for greenhouse gas emissions.

According to Jamieson these management approaches are rooted in the unjustified assumption that "the perception of self-interest is the only motivator for human beings" which in turn suggests that "[i]f you want people to do something give them a carrot; if you want them to desist, give them a stick" (Jamieson 1992: 143).[1] These components of management approaches are geared toward reaching a previously identified goal of public policy by means of structuring individual incentives in such a way that individuals have reason to align their behavior with that public policy goal. Jamieson criticizes that course of action because he believes that "exploiting people's perceptions of self-interest may not be the only way to move them" (Jamieson 1992: 144). Jamieson's first critique of management approaches assumes that they are based on an incorrect view of human psychology. It amounts to the claim that management approaches therefore champion deficient means of public policy implementation.

However, Jamieson's misgivings about the role of economic analysis within the context of the climate change debate go deeper. The management approaches fuelled by economic analyses are not only purported to be deficient in how they implement public policy but also misguided regarding which public policy they support and for what reasons they do so. To illustrate this point Jamieson correctly recounts the basic idea of cost-benefit analysis within economics:

> [w]hen faced with a policy decision, what we need to do is assess the benefits and costs o[f] various alternatives. The alternative that maximizes the benefits less the costs is the one we should prefer. This alternative is "efficient" and choosing it is "rational." (Jamieson 1992: 143)

However, according to Jamieson the application of cost-benefit analysis is misguided in the context of public policy because it is narrowly focused on economic efficiency. This he argues to be problematic because "economic efficiency is only one value, and it may not be the most important one" (Jamieson 1992: 143). Rather than focusing on economic efficiency public policy makers should take a broader view and consider other values – such as equity – which Jamieson believes to be often more important. Jamieson's second critique of management approaches and economic analyses assumes that they put far too great an emphasis on economic efficiency within public policy and therefore support the wrong ends.

Lastly, Jamieson argues that because even the overall impacts of climate change – such as the increase in the global mean surface temperature or the changes in precipitation – are so uncertain, there is virtually no way to forecast their economic consequences. This supposed special exposure of economic forecasts to uncertainty is the essence of Jamieson's third critique and according to him, "a further reason why economic considerations should take a back seat in our thinking about global climate change [...]" (Jamieson 1992: 144).

If justified, these criticisms amount to a strong argument against giving special weight to economic analyses in the context of climate change. Supposedly, management approaches attempt to reach the wrong ends by deficient means and are in the process particularly vulnerable to forecasting uncertainties. On a superficial level these claims appear to be plausible. However, as detailed assessment will reveal, these criticisms of economic analysis are in large part misguided. Sections three, four, and five provide that assessment.

[1] To support his claim Jamieson also refers to Meyers 1983.

## 3. Means for Achieving Pre-Specified Policy Goals

The claim that economic management approaches rely on an incomplete understanding of human psychology has some plausibility given that neoclassical economic models indeed rely upon the assumption that humans are exclusively self-interested. On the one hand this claim implies a theoretical problem for neoclassical economics depending on how significantly and under what kind of circumstances humans are differently motivated. However, as purely theoretical questions are less relevant for public policy this problem does not need to be addressed here. On the other hand this claim could well imply a problem for the theory's applicability. However, in the context of management approaches to climate change – and the question of whether the means these management approaches employ are well chosen to guide individual actions toward particular pre-specified ends – the potential problem of applicability does not materialize.

Recall that Jamieson's first critique implies that management approaches fail to be effective because it is false that "the perception of self-interest is the *only* motivator for human beings" [emphasis mine] (Jamieson 1992: 143). Like many authors before and after him, Jamieson makes plausible that in fact humans are also motivated by factors other than the perception of their self-interest such as "concern for family and friends, and religious, moral, and political ideals" (Jamieson 1992: 144).[2] Although this is true, it evidently does not spell trouble for economic management approaches for all they rely upon is that in the particular choice situation that is relevant to a particular approach – be it taxation, regulation, or subsidies – *one* strong motivator of human action is the perception of self-interest.

Take, for example, the standard case of a tax on $CO_2$ emissions. There might be many ways to motivate humans to reduce their $CO_2$ emissions. However, such alternatives neither imply that an appropriate tax on $CO_2$ emissions would be ineffective nor that it would be inefficient. As long as demand for $CO_2$ emissions – or rather the products and their production processes that cause them – is price sensitive, such a tax is effective. One look at consumer decisions at the fuel pump or the thermostat vindicates economic management approaches with respect to the means they employ to achieve pre-specified ends.

That management approaches which target a narrow perception of self-interest are effective in a context where that self-interest provides a strong motivator for specific consumption decisions does not, however, mean that these approaches always work. If other motivators take center stage or the perception of self-interest cannot easily be addressed, management approaches need to be adapted. As economists are not wedded to the neoclassical paradigm, they can integrate these different motivations into management approaches.

Consider the example of the anti-littering campaign "Don't Mess with Texas" which was initiated by the State Department of Highways and Public Transportation in 1985 and reduced roadside littering by more than 70% in the following five years alone (Stafford and Hartman 2012). Strict anti-littering regulations in combination with fines had limited success due to prohibitive enforcement costs. It was and is simply not feasible to patrol highways so extensively that the probability of conviction for littering in combination with a justifiable fine for individual violations produces a sufficient financial incentive to desist from littering.[3] Therefore, the State Department of Highways and Public Transportation amended their management approach in order to draw on state-specific conceptions of honor and pride to increase social pressure and exact change. The widely communicated slogan *Don't Mess with Texas* "cast litterers as 'outsiders' or 'imposters' insulting the honor of Texas" (Stafford and Hartman 2012). This effort to reframe the meaning of littering within Texas does not directly

[2] For a particularly influential example see Hume 1751. For an overview of the current debate about altruism within moral psychology see Stich, Doris, and Roedder 2010.

[3] For the general concept that underlies this particular assessment see Becker 1968.

appeal to self-interest but nevertheless follows the same strategy of aligning individual incentives with a pre-specified public policy goal. Many similar management approaches are championed by economists and public policy makers alike.[4]

In summary, the claim that management approaches employ deficient means of implementing public policy is misguided. First, standard management approaches do not rely on the assumption that self-interest is the *only* consideration that motivates humans. Rather, for them to be effective it must merely be true that in the particular choice situation *one* strong motivator for action is the agents' perception of their self-interest. Especially in the context of climate change where it is the primary goal to reduce $CO_2$ emissions this condition appears to be met. Almost all significant sources of $CO_2$ emissions are directly tied to monetized exchanges of goods and services in markets where demand is usually price sensitive – at least over the long run. Second, management approaches can easily be adapted to reflect different assumptions about human psychology. Should other concerns such as ideals of honor be seen as action guiding, management approaches can reach pre-specified public policy goals by integrating these assumptions and employing adapted techniques to sway actors.

## 4. Ends and Methodology

The more important question regarding the role of economic analysis and management approaches does, of course, not concern their means but their ends. Recall that, according to Jamieson, management approaches are too narrowly focused on economic efficiency within public policy and therefore support the wrong ends. Note that most economists differentiate between positive and normative economics. Within this distinction positive economics describes what *is* while normative economics prescribes what *should be*. Most economists view their work as positive economics which is silent regarding the ends of state action but merely analyses the effects of policy changes and recommends means to reach desired ends. Thus, to fruitfully discuss the ends of economic management approaches we must focus on normative economics also known as welfare economics.

To determine whether the ends of normative economic analysis and management approaches within public policy are well chosen, we need to first consider suitable goals of public policy and how the goals of welfare economics measure up to them. This comparison will in turn raise important questions of methodology. The answers to these questions will shed light on the "economic" aspect of the term "economic efficiency". In a second step – to avoid misinterpretations – we should consider the meaning and importance of "efficiency". This will allow for a conclusive evaluation of "economic efficiency" as a goal of public policy.

Which goals of public policy we endorse depends, of course, on our ethical theory. While utilitarianism demands that agents maximize overall welfare, virtue theory prescribes that we uphold particular virtues, and egalitarianism requires that we work toward ensuring equality. From the outset, a variety of goals such as the promotion of welfare, justice, and equality appear suitable for public policy. However, any plausible ethical theory must accept the promotion of welfare as one central concern and – although I cannot argue that point here – there are good reasons to accept the promotion of welfare as the central goal of public policy.[5]

---

[4] For a different but related subject see the debate on "nudging" within behavioural economics. The basic idea is that under certain circumstances biases cause people to make decisions that are not in their best interest and that policy makers can counteract those biases without restricting people's liberty – thereby promoting better choices. See Thaler and Sunstein 2008, Hausman and Welch 2010, as well as Sunstein 2011. Note that under the circumstances where nudging is helpful, expressed preferences have no sufficient evidential connection to welfare and that the technique must presuppose a theory of welfare.

[5] See Goodin 1995.

For the sake of the following argument, however, it suffices to presume that the promotion of welfare is *one* important aim of public policy.

Welfare economics – as the term indicates – is exclusively concerned with welfare and therefore guides public policy toward its promotion. Thus, the ultimate end of welfare economics is also at least one important aim of public policy. This is a strong prima facie reason to give considerable weight to economic analysis when making public policy decisions.

What we need to consider now is how economists make welfare measurable – we need to address questions of methodology. While philosophers vigorously debate the advantages and shortcomings of different theories of welfare – such as hedonism, desire-fulfillment theories, and objective list accounts[6] – economists assume that welfare is preference satisfaction. As any other choice of a theory of welfare would do, the focus on preference satisfaction provokes disputes as well as theoretical difficulties. To mention only a few:

> Preferences may be based on mistaken beliefs. People may prefer to sacrifice their own well-being for some purpose they value more highly. Preferences may reflect past manipulation or distorting psychological influences [...].[7] (Hausman 2008: 28)

However, rather than viewing the economists' use of a specific theory of welfare as a deep-seated theoretical endorsement, it should rather be understood as a practical decision based on the theory's virtues of applicability. For example, it would be particularly difficult to measure the mental states necessary to implement the promotion of welfare based on a theory of hedonism. In contrast, economists can measure agents' preferences via their willingness-to-pay as observed in market interactions or as derived from surveys.[8] These preferences are a good guide to agents' welfare as long as there is an "evidential connection between preferences and well-being" (Hausman and McPherson 2009: 16). This evidential connection can be utilized to predict the agents' welfare "in circumstances in which people are concerned with their own interests and reasonably good judges of what will serve their interests" (Hausman and McPherson 2009: 1).[9]

Do these conditions obtain in the context of consumption decisions that are relevant to managing climate change? When considering this question, note that it is here not relevant whether agents will react as predicted to a tax on $CO_2$ emissions which has been implemented given a pre-specified goal of reductions in $CO_2$ emissions – as was discussed in section three. Rather, the question is whether the preferences agents have with respect to $CO_2$-abatement strategies and climate change are a sufficiently good guide to their wellbeing. If their preferences are a sufficiently good guide then public policy can be designed based on these preferences in order to specify and then implement a goal of reducing $CO_2$ emissions that promotes the agents' welfare.

Two main challenges arise. First, given that climate change will have impacts on individuals around the globe, agents are likely not to be entirely self-interested with respect to climate change.[10] Questions regarding the fate of inhabitants of other continents are likely to affect

---

[6] See Crisp 2008.

[7] Also see Elster 1983.

[8] Unfortunately, given that willingness-to-pay positively correlates with income this technique has the problematic implication that it gives greater weight to those with larger incomes (see Baker 1975 as well as Hausman 2008). Although it is possible to offset these distorting effects (Harberger 1978) this is usually not done (Hausman 2008). In the context of climate change where the welfare of people with radically different levels of income and wealth living in all parts of the world needs to be taken into account it is obviously of paramount importance to make these adjustments.

[9] Also see Scanlon 1998: 116-8.

[10] A related problem is that individuals are likely to have limited concern for future generations leading them to implicitly support public climate change policies that are far less demanding than those prescribed by any plausible ethical theory. The public policy question that arises in this case is how to take into consideration the welfare of people other than those living at present and how many

agents' willingness-to-pay to slow climate change. What may appear as a veritable problem of economic cost-benefit analysis is actually an advantage when it comes to guiding public policy regarding climate change.[11] Preferences for the welfare of others – even those in different nation states – might not be the worst factors to guide public policy. These preferences may in fact be one justification for international aid. Especially in the context of climate change – where what is at stake is in large part harm inflicted upon others – those altruistic preferences should be welcomed and counted rather than rejected.

Second, given the complexities of climate change, we have to assess whether individual agents are able to be sufficiently good judges of what will serve their own welfare in the context of policy choices pertaining to climate change. Individual actors are likely to be ill informed regarding the effects of climate change and their dependence on particular policies such as those allowing for increased or decreased levels of $CO_2$ emissions. To a certain extent this ignorance is unavoidable as there simply is no conclusive information available and even experts face considerable uncertainties.

However, to the extent that experts can guide agents whose preferences are to be measured, economists can and do rely on such experts to make the preferences of agents well informed. For example, when constructing a cost-benefit analysis with the aim of evaluating policies regarding air pollution, survey participants in a study conducted by the U.S. Environmental Protection Agency have been confronted with pictures displaying different levels of possible pollution in order to allow participants to make informed judgments about the visual impact of that pollution.[12] Similar study designs need to be employed in order to improve the ill informed preferences of agents in the case of climate change.[13] Beyond supporting agents' preference formation with intricate survey designs and expert information, psychologists and economists may be able to correct ill-informed preferences for mistakes and cognitive distortions.[14]

If we assume that the information gap between experts and agents can be bridged, we are back to the more general question of whether agents' expressed preferences are likely to be a good guide to their respective welfare as long as agents are well informed. Those who are doubtful should remember that economists offer and policy makers need a practical solution to a difficult problem. Even if expressed preferences are unlikely to be an *excellent* guide to welfare that does not imply that they are not the *best available* guide. While one may not be confident that individuals are good judges of their own welfare, there is even less reason to suppose that legislators know more about what is good for the individuals affected by public

---

generations of those people living in the future we should bear in mind (Pearce, Markandya, and Barbier 1989: 172). Of course, this problem is not limited to the design of economic management approaches or the economic rationale for public policy more generally.

[11] Preferences that are not self-interested may be more problematic in the context of cost-benefit analysis when they concern beings whose welfare is not supposed to guide public policy. Assume – and I do not mean to suggest this to be true – that public policy should exclusively be concerned with human welfare. Under these circumstances a proper cost-benefit analysis that has the aim of guiding public policy should discount preferences of humans for the wellbeing of non-human animals. See also Adler and Posner 2006: 126-7 and Hausman 2009: 21. In any case it needs to be clear which specific preferences are taken into consideration.

[12] See Adler and Posner 2006: 127-8 as well as Hausman and McPherson 2009: 22.

[13] Whenever we are concerned with the valuation of environmental goods – which are usually not traded in markets – we face particular methodological obstacles because preferences for such goods cannot easily be observed. To overcome these obstacles the field of environmental economics has branched out and developed intricate techniques of benefit measurement. For an introduction see Pearce, Markandya, and Barbier 1989. A discussion of these techniques is beyond the scope of this paper not only for reasons of space but also and more importantly because the common criticisms philosophers voice against economic analysis are directed toward standard welfare economics.

[14] Hausman and McPherson 2009: 23.

policy regarding climate change.[15] Therefore, it is reasonable to assume that the well-informed and self-interested preferences of agents are – for lack of a better alternative rather than their infallibility – the best available guide to the agents' welfare.[16]

This completes the first step of our evaluation of the ends and methodology of normative economic analysis. Where does it leave us? If we face normal cognitive limitations – e.g. if we are not clairvoyant – and if our theory of ethics is purely concerned with welfare as is the case, for example, if we are utilitarians, then economic analysis should be the central tool to guide our public policy. The same result emerges if we believe that a democratic state should do nothing but further the preference satisfaction of its inhabitants.

If, on the other hand, we prescribe to a different ethical theory or our views on the responsibility of the state are less limited, the role of welfare economics in guiding public policy must be more modest. Under those circumstances welfare economics is seen as operating within legal and moral side-constraints prescribed by the law and our relevant ethical theory.[17] Though, given that welfare must play an integral role in any plausible ethical theory, welfare economics will still have an important function. Therefore, although the analysis offered by welfare economics is unlikely to be suitable as the sole criterion for good public policy, it is of central importance.

Why is the role of welfare economics within public policy nevertheless often vehemently criticized for its narrow focus? Amongst others there are two reasons. First, critics do not appear to appreciate that welfare economics is focused on promoting welfare which is indeed a rather broad goal. How else can we interpret claims like these: "Although economic efficiency may be a value, there are other values as well, and in many areas of life, values other than economic efficiency should take precedence" (Jamieson 1992: 144). It is plausible that values other than welfare should be considered in public policy; however, it is implausible that other considerations should take precedence over considerations of welfare in 'many areas of life'. At times it appears that the critics of economic analysis misinterpret economic considerations as purely financial.

Second, some appear to assume that every single policy decision must weigh all reasons that should figure in the overall public policy of the state. For example, if welfare, equity, and justice are identified as guiding principles to public policy of equal importance one may think that any specific policy decision – such as the introduction of a new tax – should be informed by all these principles to an equal degree. However, that is only true for the overall policy position of the state. Consider the example of the introduction of a tax on $CO_2$ emissions. Such taxes are often criticized because of their disproportionate negative effects on earners of low wages and the resulting distortion of equity. Here, economic analysis is criticized and it is claimed that values such as equity should be given more weight in order to exempt low wage earners from a carbon tax. However, this would be a bad policy decision for it would fail to expose low wage earners to the resulting incentive to amend their $CO_2$ "consumption". A better policy would rely on the separate instrument of the welfare state to ensure that the

[15] For the related discussion of how to evaluate adaptive preferences see Barnes 2009. Above the question of whether to take into consideration preferences that are not self-interested has been discussed and answered affirmatively for certain circumstances. That raises the question of whether the preferences of agents for the welfare of others are a good guide to the welfare of others. Again one may be doubtful. However, as these preferences are only counted in addition to the individual preferences of those others the problem is not as pronounced as when policy makers replace observed preferences with their own judgements.

[16] Those who are still sceptical of this result should consider another disadvantage of the judgements of policy makers, economists, and philosophers with respect to the welfare of others: if they go wrong they go wrong systematically with a strong impact on policy. In contrast the errors individuals make are most likely unsystematic and therefore run a much lower risk of negatively impacting public policy (Hausman and McPherson 2009: 16).

[17] Hausman 2008: 30.

value of equity is duly considered. To put it another way: if side constraints put in place by considerations of justice and equity are observed, a focus on the promotion of welfare may well be legitimate.

Having shed light on the "economic" aspect of the term "economic efficiency" we can take the second step of our evaluation of the ends and methodology of normative economic analysis. This brings us to the meaning and importance of the term "efficiency".

Jamieson claims that "[e]conomics [...] cannot tell us what our goals should be or even whether we should be concerned to reach them efficiently" (Jamieson 1992: 147). The first part of that claim is misleading because it seems to suggest that economics wants to tell us what our goals should be. Here we need to differentiate between the level of individual agents and the level of policy makers. Normative economics assumes that individuals choose their own goals and that policy makers should respect those goals by observing – and basing policy decisions on – agents' preferences. Further, normative economics tells us that policy makers should be concerned with the promotion of welfare and that economics has tools at its disposal to help them achieve that goal. In doing so, economics is not even committed to a particular theory of welfare, but relies merely on an evidential connection between preferences and welfare to offer a meaningful analysis.

However, the second part of the statement which implies that it might be up for debate whether we should be concerned to reach our goals efficiently, or whether economics should tell us one way or another, is likely a reflection of a misunderstanding regarding the concept of efficiency. To identify that misunderstanding we should consider two meanings of the term efficiency. First, efficiency as it is used in daily conversations and second, efficiency in the sense of "Pareto efficiency" as it is often used within economics. I suggest that in the above statement Jamieson seeks to imply the more common version of the term. However, I argue that independently of which version one refers to, the importance of efficiency is difficult to dispute.

According to the common interpretation of the term efficiency, a process becomes more efficient as we use fewer inputs to achieve the same output, incur fewer costs to achieve the same benefits, must sacrifice less to reach the same ends, or accept less disutility to reach the same level of utility. Similarly a process is more efficient than another if it uses fewer resources per unit of output independently of the specific level of output of these two processes. Therefore, efficiency is intimately tied to whatever our ends or values are and to whatever we see as inputs, costs, disutility, disadvantage or bads. To question whether we should be concerned to reach our goals efficiently is almost as nonsensical as questioning whether we should be concerned to be effective.

Similar misunderstandings about the concept of efficiency are prevalent in public discourse. Some argue that efficiency is of paramount importance and recommend adapting production processes such as to produce the same output with a smaller workforce or cheaper materials. Others respond that efficiency is not obviously good as seen in the resulting higher workload per employee or the lower quality of the output. This is not a basic argument about the concept of efficiency but a disagreement about whether to evaluate efficiency from a private or a public perspective. Alternatively, the disagreement can be understood as pertaining to the question of what to count as goods and what to count as bads.

These conflicting claims reflect a fundamental difference in the interpretation of inputs and outputs. While the producer disregards the longer working hours she does not recompense (additional input) and the lower quality of her products (lower output) – as long as these changes do not in turn reduce her profits – her employees and customers do not share her private perspective. Therefore, they are correct to point out that the restructuring has not increased efficiency from a public perspective. However, this claim does not pertain to the question of whether efficiency as such is good or bad. It might be due to a global and socially

irresponsible push toward lean manufacturing that efficiency has an undeserved bad reputation.

This skepticism aside, Jamieson does believe that efficiency and even economic efficiency is a value albeit "it may not be the most important one" (Jamieson 1992: 143). I am unclear on whether to even refer to efficiency as a value given its entirely derivative status and I would find the claim of economic efficiency being the most important value blatantly absurd for the very same reason.[18] Nevertheless, inefficiencies are not of marginal importance but reflect sacrifices in domains that are important to us. On one level they imply that we waste resources but on another level they imply that we get less of what we want and value. Therefore, the importance of efficiency is independent of any particular goals we may or may not have and it cannot be disputed based on the common interpretation of the term.

Let us now consider efficiency in the sense of "Pareto efficiency" as it is often used within economics.[19] In contrast to the standard notion of efficiency Pareto efficiency is concerned with evaluating outcomes based on their capacity to satisfy preferences. An outcome is Pareto efficient if it is not possible to achieve a higher level of preference satisfaction for anybody without accepting a lower level of preference satisfaction for somebody else. Conversely, an outcome is Pareto inefficient if it is possible to achieve a higher level of preference satisfaction for somebody without accepting a lower level of preference satisfaction for somebody else.[20] Thus, if an outcome is Pareto inefficient, policy makers are needlessly foregoing an opportunity to satisfy someone's preferences better. If they make use of that opportunity the new outcome is Pareto superior to the old outcome, therefore it is a Pareto improvement. Thus, all other things being equal, policy makers should aim at Pareto efficient outcomes. Just as in the case of the common reading of efficiency it is clear that it is not in anybody's interest to accept Pareto inefficient outcomes.

This concept can fruitfully be applied to the tool of cost-benefit analysis even beyond cases where direct Pareto improvements are possible. Suppose policy makers wish to compare the current outcome A to the alternative outcome B neither of which is Pareto superior to the other. Some citizens prefer A to B while others prefer B to A. Should policy makers favor the status quo or a shift to outcome B? One way to frame this question is to ask whether B is a *potential* Pareto improvement over A. Whether that is so depends on a comparison between the willingness-to-pay to bring about B of those who prefer B and the compensation necessary to convince those who prefer A to accept B. B is a potential Pareto improvement over A if those who favor B are willing to pay more for the policy change to B than what would be needed to compensate those who favor A for that same policy change. If that is so there is a net-benefit associated with the policy change to B.

Should policy makers view this increase in net-benefit and the resulting increase in economic efficiency as decisive in comparing A to B? No, they should not. But if one goal of their policy is to promote welfare then a potential increase in economic efficiency should have considerable weight in their policy decision. As Daniel M. Hausman correctly states:

> According to cost-benefit analysis, among eligible policies (which satisfy legal and moral constraints), one should, other things being equal, employ the one with the largest net benefit. (Hausman 2008: 30)

---

[18] Jamieson appears to interpret the term "economic efficiency" in reference to a very narrow interpretation of "economic grounds" (Jamieson 1992: 144). Economists would argue that this interpretation is based on a serious misunderstanding of the proper inputs of economic analysis.

[19] For the basics of Pareto efficiency refer to any introductory economics textbook such as Mankiw 2012 but also the excellent discussion in Hausmann 2008 upon which this presentation of the concept is based.

[20] See, for example, Mankiw 2012.

Therefore, even according to the concept of cost-benefit analysis itself and economic analysis more generally the criterion of economic efficiency is not of exclusive importance.

Having completed the second step of our analysis of the term "economic efficiency" we can draw on that analysis and our discussion of the ends of welfare economics in order to evaluate Jamieson's second critique. The primary results are these: First, the promotion of welfare should be at least one important goal of public policy. Therefore, the ends of welfare economics and the ends of management approaches that draw on it are well-chosen albeit not all encompassing. Thus, economic analysis should play an important role in guiding public policy. Second, on any plausible reading of the term efficiency we should be concerned to reach our goals efficiently. Therefore, we should seriously consider the results of cost-benefit analyses. How central they should be to the decisions of policy makers again depends on their views regarding the importance of welfare.

## 5. Uncertainty

Lastly, Jamieson claims that "economic considerations should take a back seat in our thinking about global climate change" because of their special exposure to risks and uncertainties (Jamieson 1992: 144). Here the basic argument is that because we cannot reliably predict the specific climatic effects of climate change (disastrous storms, floods, temperature volatility and precipitation changes, etc.) we are even less able to predict the economic effects of climate change. This is due to uncertainties about regional effects, the coping strategies and reactions of humans, the workings of local and global economies, as well as complex economic interaction effects.[21]

In part, this claim again appears to reflect a misinterpretation of the term "economic effects". If we take economic effects to relate exclusively to the workings and the development of economies Jamieson's analysis appears to be well constructed. Indeed it finds anecdotal evidence in the inability of central bankers and other experts in economics to forecast the developments of their economies during the upcoming months – entirely independently of massive exogenous shocks in the further future such as climate change.[22] However, as argued in section four, economic analysis has a much broader concern and takes into consideration a policy's effect on welfare even if that effect is independent of any interaction with the economy (e.g. welfare reductions through the negative aesthetic effects of air pollution). Importantly, these effects of climate change on welfare need to be taken into account by any plausible theory and therefore cannot be used to criticize economic analysis in particular. At present there simply is no way around the uncertainty inherent in that task.

Further, even if we narrow our perspective and exclusively consider the workings and the development of economies, it is unclear how we can avoid attempting to measure and consider these effects, as long as we have reason to believe that they will have an impact on the promotion of welfare or other policy goals we may want to pursue. However, Jamieson believes that we should refrain from doing so because "conventional economic analysis is practically useless" in the context of climate change and that other forms of analysis or other approaches to mitigating the problem are inherently better suited to deal with uncertainty (Jamieson 1992: 144). Specifically, Jamieson claims that:

> [a] bad analysis can be so wrong that it can lead us to do bad things, outrageous things – things that are much worse than what we would have done had we not tried to assess the costs and benefits at all [...]. (Jamieson 1992: 146)

[21] Jamieson 1992: 144-6.

[22] For examples consider the *Monthly Bulletin* of the European Central Bank.

This assessment carelessly neglects the analysts' awareness of the weaknesses of their predictions as well as their ability to take into account the risk aversion of those the analysis is supposed to guide. Just as climatologists offer confidence intervals for their predictions, any thoughtful cost-benefit analysis will comment on the contingencies of its predictions and address uncertainties in any recommendation. To deny this implies that economists are strangers to analyzing decisions under uncertainty – little is further from the truth. If we use an imperfect analysis in the knowledge that it is imperfect it can very well help us.

In summary, the basic observation that it is challenging to forecast the development of economies especially in the face of exogenous shocks is justified and well supported. However, there are three reasons to be skeptical of the general claim that economic analysis is practically useless when confronted with the uncertainties brought about by climate change. First, economic analysis is not only concerned with the development of economies alone but seeks to take into account all effects a policy change has on welfare. No plausible analysis can avoid considering these effects and facing the relevant uncertainties. Second, even if economic analysis was focused on the workings and the development of economies alone, it is unclear how an ethical theory could avoid taking them into account as long as they are assumed to impact welfare or any other important policy goal. Third, economic analysis is well adapted to analyzing decisions under uncertainty and therefore produces normative recommendations that take into account its exposure to that uncertainty.

## 6. Conclusion

Many of the common criticisms voiced against economic analysis as a tool within public policy are either based on misinterpretations about the aims and capabilities of normative economics or formulated without taking into view the specific practical challenges that arise in any attempt to make a theory of welfare operational. However, critics are correct to point out that the criterion of economic efficiency is not of exclusive importance. Given the dominance of economic analysis within the public policy discourse this may be an important reminder for some – even if economics as a science says nothing different and most economists are well aware of that.

One important effect of the critiques raised against economics by moral philosophers is that they inspire collaboration and bring about theoretical advancements. The work of that kind best known among philosophers is probably the capabilities approach to well-being developed by Martha Nussbaum and Amartya Sen as well as Sen's application of that approach to questions of egalitarianism.[23] As long as criticisms like those of Jamieson encourage further investigation into the important field of the philosophy of economics they are valuable.

**Joachim Wündisch**

Heinrich-Heine-Universität Düsseldorf
Joachim.Wuendisch@uni-duesseldorf.de

## References

Adler, M. D. and E. A. Posner 2006: *New Foundations of Cost-Benefit Analysis*. Cambridge, (MA): Harvard University Press.

[23] Sen 1992, Nussbaum and Sen 1993, Sen 1999, as well as Hausman 2008: 31.

Baker, C. E. 1975: 'The Ideology of the Economic Analysis of Law', *Philosophy & Public Affairs* 5, 3–48.

Barnes, E. 2009: 'Disability and Adaptive Preference', *Philosophical Perspectives* 23, 1–22.

Becker, G. S. 1968: 'Crime and Punishment: An Economic Approach', *The Journal of Political Economy* 76, 169–217.

Crisp, R. 2008: 'Well-Being', in: E. N. Zalta, (ed.) 2008.

Elster, J. 1983: *Sour Grapes: Studies in the Subversion of Rationality*. Cambridge: Cambridge University Press.

Executive Board of the ECB 2012: '*Monthly Bulletin*', European Central Bank. Available online: http://www.ecb.int/pub/mb/html/index.en.html.

Goodin, R. E. 1995: *Utilitarianism as a Public Philosophy*. Cambridge: Cambridge University Press.

Harberger, A. C. 1978: 'On the Use of Distributional Weights in Social Cost-Benefit Analysis', *Journal of Political Economy* 86, S87-S120.

Hausman, D. M. 2008: 'Philosophy of Economics', in: E. N. Zalta, (ed.) 2008.

Hausman, D. M. and M. S. McPherson 2009: 'Preference Satisfaction and Welfare Economics', *Economics and Philosophy* 25, 1–25.

Hausman, D. M. and B. Welch 2010: 'Debate: To Nudge or Not to Nudge', *Journal of Political Philosophy* 18, 123–136.

Hume, D. 1751: 'An Enquiry Concerning the Principles of Morals', in: L. A. Selby-Bigge (ed.): *Enquiries Concerning the Human Understanding and Concerning the Principles of Morals* (2nd ed.), Oxford: Clarendon Press, 1902, 176–323.

Jamieson, D. 1992: 'Ethics, Public Policy, and Global Warming', *Science, Technology & Human Values* 17, 139–153.

Mankiw, N. G. 2012: *Principles of Economics*. Mason, (OH): South-Western Cengage Learning.

Myers, M. L. 1983: *The Soul of Modern Economic Man: Ideas of Self-Interest, Thomas Hobbes to Adam Smith*. Chicago: University of Chicago Press.

Nussbaum, M. C. and A. Sen (eds.) 1993: *The Quality of Life*. Oxford: Clarendon Press.

Pearce, D. W., A. Markandya and E. Barbier (1989): *Blueprint for a Green Economy*. London: Earthscan.

Scanlon, T. 1998: *What We Owe to Each Other*. Cambridge (MA): Harvard University Press.

Sen, A. 1992: *Inequality Reexamined*. Cambridge (MA): Harvard University Press.

Sen, A. 1999: *Development as Freedom*. New York: Oxford University Press.

Stafford, E. R. and C. L. Hartman 2012: 'Making Green More Macho', *Solutions* 3, 25–29.

Stich, S., J. M. Doris and E. Roedder 2010: 'Altruism', in: J. M. Doris and F. Cushman (eds.): *The Moral Psychology Handbook*, Oxford: Oxford University Press, 147–205.

Sunstein, C. R. 2011: 'Empirically Informed Regulation', *University of Chicago Law Review* 78, 1349–1429.

Thaler, R. H. and C. R. Sunstein 2008: *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven: Yale University Press.

Zalta, E. N. (ed.) 2008: *The Stanford Encyclopedia of Philosophy*. Stanford (CA): The Metaphysics Research Lab.

# 8. Normative Ethik, Metaethik, Handlungs- und Entscheidungstheorie

# Defending Moral Intuitionism Against Debunking Arguments

Anne Burkard

One way of challenging someone's moral beliefs is to show that their sole basis is a religious doctrine which the person used to accept, but no longer does. This way of questioning a belief can be called a *debunking argument*. Recently, debunking arguments have been put forward as a challenge to robust versions of moral realism (and realism about normative discourse in general). Thus the target of these debunking arguments is not the justification of moral beliefs of just any description. Rather, the debunking is taken to challenge moral beliefs only when understood in a robustly realist way. Instead of pursuing the putative metaphysical implications of debunking arguments, this paper focuses on the relevance of such arguments for the epistemic justification of moral beliefs regardless of the metaphysical framework. More specifically, I propose ways to respond to debunking arguments within an intuitionist framework. I argue that a plausible form of intuitionism can successfully assimilate convincing debunking arguments. To do so, I first sketch a minimal form of moral intuitionism that can be spelt out both in realist and anti-realist versions. I then specify the nature of *local* and *global* debunking arguments, and show how intuitionists can and should respond to arguments of both sorts.

## 1. Introduction

There are many ways to challenge someone's moral beliefs. We could, for example, identify internal inconsistencies within a person's moral belief system, we could point out that many well-informed, sensitive and empathetic people hold moral beliefs that conflict with hers, or we could show that the sole basis of some moral conviction of hers is a religious doctrine which she used to accept, but no longer does.

This last way of challenging a moral belief can be called a *debunking argument*. Such arguments are currently receiving considerable philosophical attention.[1] A dominant theme in recent discussions is the support these arguments seem to lend to forms of anti-realism about morality (and normative discourse in general): Debunking arguments are put forward as a challenge to robust versions of moral realism, positions according to which there are mind-independent moral facts that serve as truth-makers for the relevant convictions (this, at least, is how a prominent, though disputed understanding of 'robust realism' would have it).[2] Thus the target of these debunking arguments is not the justification of moral beliefs of just any description. Rather, the debunking is taken to challenge moral beliefs only when they are understood in a robustly realist way (cf. e.g. Clarke-Doane 2012, Kitcher 2006 and Street 2006). The discussion often concerns not only moral beliefs, but also normative or evaluative

[1] Cf. Clarke-Doane 2012, Enoch 2010, Huemer 2008, Joyce 2006, Kahane 2011, Lillehammer 2003, Sinnott-Armstrong 2006, Skarsaune 2011, Street 2006 and Tersman 2008 for just a small selection of recent contributions to this debate. Mason 2010 discusses a number of prominent examples of debunking arguments taken from the history of religion and philosophy, including religious texts and the work of authors as diverse as Friedrich Nietzsche, William James and Sigmund Freud.

[2] Cf. e.g. Shafer-Landau 2003: 13-18 and Street 2006: 110-112 for roughly this understanding of realism and cf. Skarsaune 2011, section 5 for criticism thereof. The following discussion does not require taking a stance on this dispute.

beliefs in general. In this paper, I speak only of moral beliefs, but nothing in the presented arguments hinges on this difference.

Instead of pursuing the metaphysical implications that certain debunking arguments purportedly have, this paper focuses on the relevance of such arguments for the epistemic justification of moral beliefs regardless of the metaphysical framework. More specifically, I propose various ways to respond to debunking arguments from an intuitionist perspective on moral justification. I argue that a plausible form of intuitionism can successfully assimilate convincing debunking arguments. It can and should do so irrespective of whether the position is spelt out in a realist or an anti-realist manner.

I begin by briefly sketching a minimal intuitionist position and the sceptical consequences entailed in the rejection thereof (section 2). Secondly, I characterise the nature of debunking arguments in more detail and explain to what extent such arguments can challenge all non-sceptical positions about moral justification (section 3). Thirdly, I discuss what I call the *local* form of debunking arguments. These are arguments that question the putative justification of specific moral beliefs or subsets thereof. I show that one way of responding to these arguments is to use them constructively in an account of trustworthy intuitions (section 4). In the final section, I sketch an intuitionist response to *global* debunking arguments, whose aim is to debunk morality as a whole (section 5).

## 2. A Minimal Form of Moral Intuitionism and the Sceptical Alternative

The central tenet of the form of moral intuitionism that I want to defend against debunking arguments is this: When searching for justified answers to moral questions, we may, and indeed have to, provisionally rely on at least some moral beliefs for which we have no inferential justification (i.e., beliefs that are either *not (positively) justified* or justified *non-inferentially*). And, somewhat more precisely: Those beliefs are not acquired through sense perception, memory, introspection or testimony alone.[3] We can call these moral judgments 'moral intuitions'. Crucially, this minimal intuitionist position is compatible with many metaethical views, including forms of realism and anti-realism, naturalism and non-naturalism, as well as moderate variants of coherentist and foundationalist theories of justification.[4]

The intuitionist claim that we have to rely on moral intuitions when engaging in moral inquiry rests on two assumptions. First, it is based on the hardly controversial conviction that we need to start somewhere if we are to form any justified beliefs at all. Those starting points could be immediately justified beliefs, as in conceptions that take the contents of moral intuitions to be self-evident (cf. Audi 2004 and Ross 1930/2002 for prominent versions of this view). But the starting points can also simply be what we currently believe. So proponents of moderate forms of coherentism share this first assumption, insofar as they use notions like the 'permissive justification' or 'initial tenability' of beliefs (cf. Sayre-McCord 2007 and Elgin 2005 respectively for versions of coherentism using these terms).

---

[3] The following discussion of intuitionism and debunking arguments focuses on moral intuitions as a type of *belief*. However, the discussion could easily be reformulated so that it also applies to non-doxastic accounts of intuitions, according to which undefeated intuitions are *seemings*, i.e. non-doxastic mental states, which give us some justification to believe their propositional content (cf. Huemer 2005: 5.1 und Huemer 2008).

[4] Cf. DePaul 2006 and Nelson 1999 for similar positions on intuitions in moral inquiry. Intuitionism is often defined more narrowly though, namely as being committed to foundationalism about justification or knowledge (cf. e.g. Crisp 2006: 71-73 and Sinnott-Armstrong 2006: 185, 190). See Burkard 2012 for a more detailed presentation and defence of the minimal form of intuitionism briefly outlined here.

The second assumption motivating the intuitionist claim is that morality is autonomous in two senses. On the one hand, it is *logically autonomous*, which means that there are no valid, informative inferences from purely non-moral premises to moral conclusions. This thesis is a version of the famous 'is/ought gap' (cf. Schurz 1997 for a detailed discussion). And on the other hand, morality is *semantically autonomous*: we cannot give a convincing reductive analysis of moral propositions in purely non-moral terms.[5] The thesis of the logical autonomy of morality is of course widely accepted, but many also agree that morality is semantically autonomous. (Intuitionism in the minimal sense is *not* committed to a third version of the autonomy thesis, namely the claim that morality is *metaphysically* autonomous.) Anyone who subscribes to both theses, and who acknowledges that we 'need to start somewhere' in our inquiries, is committed to the intuitionist claim that we have to rely on some moral intuitions *if* we are to find justified answers to moral questions.

However, they could still deny that we are in fact justified in any of our moral beliefs. Walter Sinnott-Armstrong, for example, is an advocate of this view. He accepts that moral justification is only possible if we rely on moral intuitions at some point. But he defends a form of epistemic moral scepticism, because he believes that we can never trust our moral intuitions. And this mistrust of moral intuitions is partly based on his acceptance of certain debunking arguments (cf. Sinnott-Armstrong 2006: 207-210). Let us now take a closer look at the nature of such arguments.

## 3. How Debunking Arguments Can Challenge Moral Intuitionism

Debunking arguments aim at undermining the beliefs they target, in the sense of showing them to be unjustified or of removing their justification. They do so by providing an allegedly exhaustive explanation for why the relevant propositions are believed, an explanation that is (a) independent of the truth or correctness of the beliefs and (b) leaves no room for a justificatory account of them. The general form of debunking arguments can be put as follows:[6]

P1 The best explanation for *S*'s belief *that p* is origin or process *X*.

P2 *X* has nothing to do with the truth or correctness of *p*.

C *S*'s belief *that p* is not justified or trustworthy.

Debunking arguments can be formulated both from within internalist and externalist conceptions of justification. A third premise like "*S* is in a position to know that P1 and P2 are the case", i.e. a premise concerning the *access* to the undermining factors is needed in internalist views to warrant the conclusion. By adding a fourth premise such as the following, we could transform the argument into a deductive one: "When an origin or a process that best explains a belief has nothing to do with the belief's correctness or truth, then this belief is not justified or trustworthy." For the sake of simplicity, I ignore those modifications here.

Consider the following example as an illustration of the potential force of this type of argument:

---

[5] Some forms of semantic reductionism actually are compatible with intuitionism insofar as they are committed to the view that to find an adequate reduction of moral propositions in terms of non-moral propositions, we need to rely on moral intuitions (cf. Schroeder 2009: 199). Thomas Scanlon's contractualism, for example, which proposes that moral judgments be analysed in terms of reasons, is compatible with the semantic autonomy thesis. This is because the notion of 'reasonableness' that Scanlon uses here is, to use his phrase, "an idea with moral content" (Scanlon 1998: 194; cf. also Parfit 2011: vol., 1 sect. 53).

[6] This is a slightly adapted version of Guy Kahane's analysis of debunking arguments (cf. Kahane 2011: 106).

> Richard grew up a devout Catholic, accepting the sexual morals he was taught. However, he has long since become an atheist and has cast off many of the views of his youth. Only his belief in the special moral status of monogamous, heterosexual marriage has not changed. When he is challenged by his friend Susan to provide a justification for this view, Richard realises that he has no non-religious basis for it. As a consequence, he re-thinks his position on marriage.

Richard's reaction to Susan's challenge is clearly reasonable since he has been presented with an explanation for a belief of his which undermines any justification he might have had for it. But how exactly are we to understand the undermining force of such explanations? Although this is not the place to provide a comprehensive account of debunking arguments, a few clarifications are needed.

First, the fact that someone's justification for believing *that p* has been undermined by a debunking explanation does of course not imply the falsity of *p*. *P* could still turn out to be true, but the person currently has no justificatory basis for believing so. That is, successful debunking explanations function as so called *undercutting*, not as *rebutting defeaters* (cf. Pollock and Cruz 1987/1999: 196f. for this distinction).

Second, the explanations referred to in debunking arguments only work in this undermining fashion insofar as we understand them as leaving no room for alternative, non-undermining or *vindicating* explanations. In a variant of the scenario just described, Richard could have come to hold the same view on marriage but on a different basis. In that case, reference to his religious upbringing would not have provided the best (or a complete) explanation for the relevant belief, which would hence not have been undermined (cf. Kahane 2011: 106 for a similar point).

Third, it is a complex question what qualifies as a good explanation, and in many cases philosophers will disagree as to what best explains the acceptance of some moral proposition. For present purposes, just note that it is not presupposed here that the explanations referred to in debunking arguments must be causal, or that they cannot make reference to non-natural properties. [7] Also, I take it that there are at least some cases in which it is fairly easy to determine what best explains someone's acceptance of certain moral propositions (as I have stipulated in Richard's case above).

Fourth, the notions of 'correctness' and 'truth' used in debunking arguments should be understood broadly. Reference to moral truth is not limited to robust forms of moral realism (and the correspondence theory of truth with which it is associated). Rather, the term is also employed by defenders of moral constructivism, by proponents of a coherence theory of moral truth or by quasi-realists who subscribe to minimalism about truth (cf. e.g. Blackburn 1998, 75-80, Skorupski 1999 and Street 2010). Thus the unspecific formulation used in the second premise ("*X* has *nothing to do with* the truth or correctness of *p*") is deliberately chosen to make it compatible with different conceptions of truth.

And, finally, attempts to debunk beliefs by pointing out their supposedly dubious origins can of course misfire. Consider the following example:

> Irene learns about evolutionary hypotheses regarding the origins of our cognitive faculties, among them our ability to make simple calculations. She worries that the fact that these faculties were selected through evolutionary processes speaks for their usefulness for survival, rather than for our ability to recognise algebraic truths. But Joan points out to her that it is plausible to assume that the ability to calculate was useful for survival only because it enabled humans to calculate *correctly*.

[7] Cf. Mayes 2005 for a discussion of some central issues concerning the nature of explanations, including reference to anti-realist approaches to (scientific) explanations.

This example differs from the case of Richard in so far as it does not provide the right kind of material needed for the second premise of a debunking argument. This is because it is suggested in this case that the best explanation for our calculation ability makes reference to the correctness of the calculations. It can thus be taken to be a *vindicating*, instead of an undermining explanation.[8]

Given this characterisation of debunking arguments, it is easy to see that they can present a challenge to moral intuitions and to moral beliefs in general, just as any other type of belief is in principle vulnerable to them. If, for example, the best explanation for the widely held belief that sexual relations between siblings are always morally wrong is that an innate disgust mechanism leads us to make this judgment, then on most metaethical positions we have good reason to reconsider our views on incest. Similarly, any sensible view would epistemically require test persons of a psychological experiment to revise their judgments when they learn that they only judged certain behaviour as morally objectionable because they had been manipulated under hypnosis.[9] That is so because the explanations in both cases are plausibly seen as having nothing to do with the correctness or truth of the challenged beliefs. Clearly, this assessment is not only plausible on robust realist understandings of morality, but can also be accepted, e.g., by Scanlonian contractualists or by proponents of response-dependent views like John McDowell's.[10]

As indicated above, debunking arguments relevant to non-sceptical views of moral justification can be divided into two types: *local* and *global* debunking arguments. I will first discuss the former.

## 4. Responding to Local Debunking Arguments

Local debunking arguments are sometimes advanced in order to undermine a certain subgroup of moral beliefs. Peter Singer's attempt to undermine emotionally charged moral judgments is a prominent example for this strategy (cf. Singer 2005).

On the basis of psychological and neuroscientific studies in combination with evolutionary considerations, Singer argues that certain moral intuitions should not be given any epistemic weight. Specifically, he intends to debunk common intuitive reactions to certain instances of the famous *trolley cases*, reactions often invoked by defenders of deontological views. Consider the following pair of cases: If asked whether it is permissible to pull a switch in order to divert a train from a track where it would kill five and onto a track where only one person would be killed (when no other options are available), most people answer 'yes'. If asked, however, whether it is permissible to push a big man from a footbridge if this is the only way to stop a train that would otherwise kill five, most people answer 'no'.

The latter reaction can be explained, Singer claims, with reference to negative emotional dispositions toward 'up-close and personal' forms of violence which our ancestors developed in early stages of human development, where those dispositions were advantageous to humans

[8] Cf. also Joyce's suggestion: "So does the fact that we have such a genealogical explanation of our simple mathematical beliefs serve to demonstrate that we are unjustified in holding these beliefs? Surely not, for we have no grasp of how this belief might have been selected for [...] independent of its truth. False mathematical beliefs just aren't going to be very useful." (Joyce 2006: 182) We can leave it open here whether this account of simple mathematical beliefs is indeed convincing; cf. Clark-Doane 2012 for criticism of Joyce's view.

[9] Cf. Haidt 2001: 814, for a discussion of an empirical investigation of people's reactions to a scenario where siblings have consensual safe sex, as well as for a discussion of similar reactions to other scenarios; cf. Wheatley and Haidt 2006 for an experiment involving hypnotically induced disgust reactions that made moral judgments more severe.

[10] Cf. Southwood 2009 on contractualism and D'Arms and Jacobson 2006 for a discussion of response-dependent views.

living together in small groups. 'Impersonal' forms of violations such as throwing a switch to set a train in motion that will kill someone were not available at that stage and so no such emotional dispositions towards them were developed. It is not surprising, Singer contends, that pulling a switch to redirect the train in the first scenario is judged permissible by many. For in such impersonal cases, widely shared rational insights like "five deaths are worse than one" are not distorted by emotional reactions (cf. Singer 2005: 347-351; the direct quote is from 350). This assessment seems to gain further support from measurements of the reaction times of the subjects. The few who judged that it is permissible to push the big man from the footbridge apparently took significantly longer to come to this conclusion than those who judged it impermissible; this is taken to indicate that during the longer reaction time subjects 'overcame' their spontaneous emotional reaction (cf. ibid.).[11] Singer's position can be restated as follows:

P1 The best explanation for most people's intuitive reactions to up-close and personal cases is that they result from evolved emotional dispositions that used to be advantageous in earlier stages of human development.

P2 Evolved emotional dispositions that used to be advantageous in earlier stages of human development have nothing to do with the truth or correctness of people's intuitive reactions to up-close and personal cases.

C People's intuitive reactions to up-close and personal cases are not justified or trustworthy.[12]

How should we evaluate this debunking attempt? Both premises of the argument are controversial; we can first question whether the offered explanation for the intuitive reactions really is the best. The empirical studies purportedly showing this have received much critical attention that cast doubt on the claim. The second premise can be questioned as well, for it is far from clear that emotional dispositions need to be regarded as distorting influences in moral judgments.[13] However, we may also find that Singer's debunking argument is plausible. We could try to support his position by defending a strong version of cognitivism, for example, and thereby underpin the second premise.

Abstracting from the example, we can identify the following three ways of reacting to local debunking arguments. First: To defend an intuition (or a subset of intuitions) against debunking attempts, one may question the explanation given in the first premise of such arguments by offering an alternative, non-undermining or vindicating explanation for the judgment in question.

Second: Alternatively, one could accept the explanation but reject the epistemic standard used in the second premise, e.g. the assumption that evolved emotional dispositions necessarily distort moral judgments.

And third: We may also accept the debunking attempt because we find both premises convincing. If the debunking concerns a subset of moral intuitions, then an intuitionist can take that as a reason to exclude those intuitions or intuitions of that type from her positive account of moral justification. Although debunking arguments are usually formulated with a sceptical

[11] The distinction between 'up-close and personal' and 'impersonal' cases of violation was introduced by Joshua Greene and colleagues in one of the empirical studies Singer relies upon. The footbridge case belongs in the former category, the switch case in the latter (cf. Greene et al. 2001: 2106). Elsewhere Greene argues like Singer that the empirical findings support a consequentialist ethics (cf. Greene 2008).

[12] As noted above, on an internalist understanding of justification we would have to add a premise about people being in a position to know about P1 and P2.

[13] Cf. Berker 2009 and Tersman 2008 for criticism of both types. Cf. Kahane et al. 2012 for more recent empirical studies of intuitive responses to trolley cases and other, less extreme scenarios, which also question the empirical data Singer relies on.

impetus, they clearly have constructive potential: As intuitionism only claims that it is necessary to rely on *some* moral intuitions to avoid moral scepticism, and as it is highly plausible that not all intuitions are trustworthy, the intuitionist agenda involves distinguishing trustworthy from untrustworthy intuitions; we can say that it involves creating an 'intuition filter'.[14] Convincing local debunking arguments can and should be recruited for this task.

It is crucial to acknowledge that in all these ways of responding to local debunking arguments, metaethical, general epistemological and maybe even moral considerations are involved (on the latter, see the following section). Although we may start with results from empirical investigations, and indeed gain helpful insights from them, *assessing* those results involves philosophical, and among them normative, considerations. The outcome of this process will vary for different metaethical frameworks and can be more or less reversionary. But *in principle*, local debunking arguments can challenge moral judgments on any non-sceptical view of moral justification. Likewise, the three described options for responding to the challenge are plausible on versions of moral (as well as evaluative or normative) realism and anti-realism alike.

## 5. Responding to Global Debunking Arguments

The presented outline of defence strategies for intuitionism against debunking arguments is, however, incomplete. With regard to Singer's challenge it has recently been suggested that the attempt to keep the debunking of moral intuitions *local* is likely to fail. That is because analogous debunking arguments can be advanced to target the intuitions on which Singer himself relies. One could, for example, point out that the universalistic-egalitarian doctrine central to Singer's utilitarianism has its root in Christian ethics – a basis for morality Singer has repeatedly rejected as inadequate (cf. Tersman 2008: 401f.).[15] A global debunking argument of this sort can be formulated with regard to the influence of evolutionary forces on our moral outlook:

P1 The best explanation for the contents of our moral intuitions is that evolutionary forces had a tremendous influence on them.

P2 Evolutionary forces have nothing to do with the correctness or truth of our moral intuitions.

C Our moral intuitions are not justified or trustworthy.[16]

If this debunking argument is successful and given the intuitionist framework outlined above, we are left with moral scepticism. But *is* it successful? We can see immediately that the first two ways of responding to *local* debunking arguments are also applicable to this and other *global* debunking attempts; both the explanation adduced in the first premise and the

[14] Cf. Daniels 1996: 82 for the use of the filter metaphor within his reflective equilibrium approach.

[15] See also Kahane 2011 for a version of the worry that debunking arguments that were meant to target only a subgroup of moral beliefs end up undermining *all* moral justification (at least on a robust realist or, as Kahane calls it, an objectivist understanding of morality).

[16] This argument is an adaptation of the first horn of Street's much discussed Darwinian Dilemma for realist theories of value: "Evolutionary forces have played a tremendous role in shaping the content of human evaluative attitudes. The challenge for realist theories of value is to explain the relation between these evolutionary influences on our evaluative attitudes, on the one hand, and the independent evaluative truths that realism posits, on the other. Realism, I argue, can give no satisfactory account of this relation. [...] [T]he realist may claim that there is no relation between evolutionary influences on our evaluative attitudes and independent evaluative truths. But this claim leads to the implausible skeptical result that most of our evaluative judgements are off track due to the distorting pressure of Darwinian forces." (Street 2006: 109)

epistemic standard formulated in the second can clearly be challenged. There is no room to discuss these options in any detail here, but promising responses have been developed in the literature from both realist and anti-realist perspectives, responses which can be used in a defence of moral intuitionism. To just briefly state these two options with respect to the evolutionary debunking argument: We may doubt that the influence of evolutionary forces is indeed so considerable that all or most of our intuitions are (directly or indirectly) affected. Or we could argue that the influence of evolutionary forces is not necessarily detrimental to the trustworthiness of our moral intuitions. That could be so because we may regard our ability to form justified moral beliefs as a by-product of the development of those epistemic faculties that can reasonably be considered as providing us with justified or true beliefs *and* as beneficial for survival (or as 'fitness enhancing').[17] Remember, however, that a proponent of the minimalist intuitionist position is not necessarily committed to this kind of realism, or to moral realism at all. As Street argues, anti-realists have different resources to react to hypotheses about the influence of evolutionary forces on our moral (and other evaluative) views, and to avoid scepticism about the beliefs in question (cf. Street 2006: sect. 10).

Let me now comment on the third option described above, that of accepting a debunking argument but using it constructively in a positive account of moral justification. It may seem obvious that this response is not available in the case of global debunking arguments, for there would simply be no intuitions left to rely on if such an argument succeeded in showing that none of them are justified or trustworthy. However, this verdict is premature. We can see why when we adopt what may be called a holistic approach to debunking attempts.

According to such an approach, we may be rationally required to start our reflection process anew if a global debunking argument *seems* to force us to conclude that no moral belief is justified. This sort of response to debunking arguments is constructive in the following sense: Global debunking attempts can indicate that some moral convictions are so central to our belief system that in cases of conflict, it is more reasonable to give up some theoretical beliefs than those convictions. That is, global debunking arguments can function as a *reductio* of the empirical or theoretical thesis in question.

If it were the case, say, that the apparently best explanation for why we hold moral beliefs, in combination with a robust version of realism, leads us to conclude that no moral beliefs are justified, then this could be a reason to either reconsider the explanation in question or to reject this form of realism. In fact, also local debunking arguments can sometimes be treated as *reductiones* of this sort. For example, I might be impressed by Singer's debunking argument, but also be deeply convinced that it is permissible to throw the switch in the first trolley case but impermissible to push the big man in the second. In such a case, my moral conviction can give me a reason to treat the argument as a *recuctio* and search for a non-debunking account of my judgment and thus gain (new) justification for it (a search that might fail, of course).

This description of debunking arguments implies that (certain) moral beliefs can give us reasons to either reconsider an explanation of why we make moral judgments or to adopt some metaethical view rather than another. That is, although debunking arguments can put pressure on moral convictions as well as on theoretical views about morality, they do not necessarily have the final say. In this vein, Street and others take their conjecture that our moral beliefs have been shaped considerably by evolutionary forces as a reason to reject certain *metaethical* views rather than *moral* convictions.

The assessment that we may have reason to give up some theoretical assumptions or to rethink certain explanations so that we can coherently hold on to some moral convictions, dovetails well with the intuitionist outlook. For though we should certainly not claim that

[17] Cf. e.g. FitzPatrick 2008, Enoch 2010 and Skarsaune 2011, who all try to defend a robust version of moral realism against an evolutionary debunking argument.

moral judgments generally or even often take precedence over other types of judgments, we *should* accept that they can have independent epistemic weight which needs to be considered in a holistic evaluation of debunking arguments. This, in fact, is just another way of expressing the view that morality is autonomous in the sense described at the outset.[18]

**Anne Burkard**

Humboldt-Universität zu Berlin
anne.burkard@philosophie.hu-berlin.de

## References

Audi, R. 2004: *The Good in the Right*. Princeton: Princeton University Press.

Blackburn, S. 1998: *Ruling Passions*. Oxford: Oxford University Press.

Burkard, A. 2012: *Intuitionen in der Ethik*. Münster: Mentis.

Clarke-Doane, J. 2012: 'Morality and Mathematics: The Evolutionary Challenge', *Ethics* 122, 331–340.

Crisp, R. 2006: *Reasons and the Good*. Oxford: Oxford University Press.

D'Arms, J. and Jacobson, D. 2006: 'Sensibility Theory and Projectivism', in D. Copp (ed.): *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press, 186–218.

Daniels, N. 1996: *Justice and Justification: Reflective Equilibrium in Theory and Practice*. Cambridge: Cambridge University Press.

DePaul, M. R. 2006: 'Intuitions in Moral Inquiry', in D. Copp (ed.): *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press, 595–623.

Elgin, C. Z. 2005: 'Non-Foundationalist Epistemology: Holism, Tenability and Coherence', in M. Steup and E. Sosa (eds.): *Contemporary Debates in Epistemology*, Malden (MA): Blackwell, 156–167.

Enoch, D. 2010, 'The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It', *Philosophical Studies* 148, 413–438.

FitzPatrick, W. 2008: 'Morality and Evolutionary Biology', in E. N. Zalta (ed.): *The Stanford Encyclopedia of Philosophy*, URL: <http://plato.stanford.edu/archives/win2008/entries/morality-biology/>.

Greene, J. D. 2008: 'The Secret Joke of Kant's Soul', in W. Sinnott-Armstrong (ed.): *Moral Psychology, Vol. 3: The Neuroscience of Morality. Emotion, Brain Disorder, and Development*, Cambridge (MA): MIT Press, 35–79.

Greene, J. D., Sommerville, B. R., Nystrom, L. E., Darley, J. M., and Coheno, J. D. 2001: 'An fMRI Investigation of Emotional Engagement in Moral Judgment', *Science* 293, 2105–2108.

Haidt, J. 2001: 'The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment', *Psychological Review* 108, 814–834.

Huemer, M. 2005: *Ethical Intuitionism*. New York: Palgrave Macmillan.

— 2008: 'Revisionary Intuitionism', *Social Philosophy and Policy* 25, 368-392.

Joyce, R. 2006: *The Evolution of Morality*. Cambridge (MA): MIT Press.

Kahane, G. 2011: 'Evolutionary Debunking Arguments', *Noûs* 45, 103–125.

[18] Many thanks to Philipp Brüllmann, Benjamin Emerson, Simon Gaus, Benjamin Kiesewetter, Andreas Müller, Thomas Schmidt and to the audience of my talk given at GAP.8 for helpful comments on earlier versions of this paper.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., and Tracey, I. 2012: 'The Neural Basis of Intuitive and Counterintuitive Moral Judgement', *Social Cognitive and Affective Neuroscience* 7, 393–402.

Kitcher, P. 2006: 'Biology and Ethics', in D. Copp (ed.): *The Oxford Handbook of Ethical Theory*, Oxford: Oxford University Press, 163–185.

Lillehammer, H. 2003: 'Debunking Morality: Evolutionary Naturalism and Moral Error Theory', *Biology and Philosophy* 18, 567–581.

Mason, K. 2010: 'Debunking Arguments and the Genealogy of Morality and Religion', *Philosophy Compass* 5, 770–778.

Mayes, G. R. 2005: 'Theories of Explanation', in: *The Internet Encyclopedia of Philosophy*, URL: www.iep.utm.edu/explanat/.

Nelson, M. 1999: 'Morally Serious Critics of Moral Intuitions', *Ratio (New Series)* 12, 54–79.

Parfit, D. 2011: *On What Matters, Vol. 1 and 2*. Oxford: Oxford University Press.

Pollock, J. L. and Cruz, J. 1987/1999: *Contemporary Theories of Knowledge, 2nd ed*. Towota (NJ): Rowman and Littlefield Publishers.

Ross, D. 1930/2002: *The Right and the Good*, ed. by Ph. Stratton-Lake. Oxford: Oxford University Press.

Sayre-McCord, G. 2007: 'Coherentism and the Justification of Moral Belief', in R. Shafer-Landau (ed.): *Ethical Theory: An Anthology*, Malden (MA) et al.: Blackwell 123–139.

Scanlon, T. M. 1998: *What We Owe to Each Other*. Cambridge (MA): Harvard University Press.

Schroeder, M. 2009: 'Huemer's Clarkeanism', *Philosophy and Phenomenological Research* 78, 197–204.

Schurz, G. 1997: *The Is-Ought Problem: An Investigation in Philosophical Logic*. Dordrecht: Kluwer Academic Publishers.

Shafer-Landau, R. 2003: *Moral Realism: A Defence*. Oxford: Oxford University Press.

Singer, P. 2005: 'Ethics and Intuitions', *The Journal of Ethics* 9, 331–352.

Sinnott-Armstrong, W. 2006: *Moral Skepticism*. Oxford: Oxford University Press.

Skarsaune, K. O. 2011: 'Darwin and Moral Realism: Survival of the Iffiest', *Philosophical Studies* 152.2, 229–243.

Skorupski, J. 1999: 'Irrealist Cognitivism', *Ratio (New Series)* 12, 436–459.

Southwood, N. 2009: 'Moral Contractualism', *Philosophy Compass* 4, 926–937.

Street, S. 2006: 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies* 127, 109–166.

— 2010: 'What is Constructivism in Metaethics?', *Philosophy Compass* 5, 363–384.

Tersman, F. (2008), 'The Reliability of Moral Intuitions: A Challenge from Neuroscience', *Australasian Journal of Philosophy* 86, 389–405.

Wheatley, Th. and Haidt, J. 2005: 'Hypnotically Induced Disgust Makes Moral Judgments More Severe', *Psychological Science* 16, 780–784.

# Overdetermination in Inuitive Causal Decision Theory

Esteban Céspedes

Causal decision theory defines a rational action as the one that tends to cause the best outcomes. If we adopt counterfactual or probabilistic theories of causation, then we may face problems in overdetermination cases. Do such problems affect Causal decision theory? The aim of this work is to show that the concept of causation that has been fundamental in all versions of causal decision theory is not the most intuitive one. Since overdetermination poses problems for a counterfactual theory of causation, one can think that a version of causal decision theory based on counterfactual dependence may also be vulnerable to such scenarios. However, only when an intuitive, not analyzed notion of causation is presupposed as a ground for a more plausible version of causal decision theory, overdetermination turns problematic. The first interesting consequence of this is that there are more reasons to dismiss traditional theories of causation (and to accept others). The second is to confirm that traditional causal decision theory is not based on our intuitive concept of the causal relation.

## 1. Overdetermination and Two Theories of Causation

Overdetermination cases present problems to a diverse range of causal notions. The particularity of overdetermination is the fact that there are two or more causes that are sufficient to produce a single event. Some of the classical examples of symmetric and asymmetric overdetermination are the following (Lewis 1973, 2000; Hall 2004).

**Symmetric overdetermination**

Suzy and Billy are throwing rocks at a bottle. Two stones, one thrown by Suzy and the other by Billy, hit the bottle at the same time and it shatters. Anyhow, either of both throws would have been enough to shatter the bottle without the other.

The next is the case of asymmetric overdetermination, also called *preemption*, and its difference from symmetric cases lies, normally though not necessarily, on the temporal order of the two causes, i.e. Billy and Suzy throwing rocks, and on the interruption of one by the other.

**Asymmetric overdetermination**

Suzy and Billy are throwing rocks at a bottle. One of Suzy's rocks hits the bottle first, breaking it. Billy's stone would have hit the bottle later, if Suzy's throw had not broken the bottle.

The ways in which Billy's stone might have hit the bottle can change depending on the time at which his throw is prevented to hit the bottle. Thus, there are cases of early and late preemption.

**Early asymmetric overdetermination**

Suzy throws a stone at a bottle, breaking it. Billy was prepared to throw his stone, but he did not throw it. He would have thrown his stone, if Suzy had given him a signal, meaning that she would not throw. In such case, his throw would have broken the bottle instead.

**Late asymmetric overdetermination**

Suzy and Billy are throwing rocks at a bottle. One of Suzy's rocks hits the bottle first, breaking it into many glass pieces. Billy's stone flies a bit later exactly over the spot where the bottle was.

It has been argued that such a distinction between early and late asymmetric overdetermination is not necessarily a fundamental one (Hall 2004: 236), but I am not going to get into that debate. There is another, non-temporal kind of asymmetry implied in overdetermination by trumping (Lewis 2000, Stone 2009).

**Trumping asymmetric overdetermination**

The officer and the sergeant simultaneously shout at the soldiers to advance. The soldiers advance, but only following the officer's order, which seems to be the cause of the fact that the soldiers start marching.

Let us now see why these are taken to be problems. A counterfactual theory of causation, specially, suffers from this and the above described overdetermination scenarios.

**Counterfactual theory of causation**

One event is caused by another if and only if the former depends causally on the latter. An event e depends causally on a distinct event c, if and only if both events occur and if c had not occurred, e would not have occurred.

Now, symmetric overdetermination presents a problem to this theory, because although it seems plausible to affirm that Suzy's throw was a cause of the bottle's breaking, it is not true that the breaking would not have happened without Suzy's throw. The different types of asymmetric overdetermination are problematic for similar reasons.

In facing difficulties with overdetermination in general, probabilistic theories of causation are not an exception.

**Probabilistic theory of causation**

An event c causes a distinct event e, if and only if the probability that e occurs, given the fact that c occurs, is higher than the probability that e occurs, given c's absence. Expressed formally, $P(E|C) > P(E|\text{-}C)$.

Suppose a scenario of late asymmetric overdetermination, in which Suzy only throws if she sees that Billy is going to throw his rock as well. Suppose also that she always throws faster than him. Before the game starts, it would seem to be true that Billy's throw counts as a cause for the breaking of the bottle. For his throw raises the probability of the breaking. But it is actually Suzy's throw that breaks the bottle. It is thus a case of probabilistic influence without causation.

Symmetric overdetermination cases are also problematic. Suppose that both Suzy and Billy throw simultaneously and that, somehow, seeing other people throwing stones at the same time disconcentrates them. Such distraction lowers the probability of the shattering. Suzy's throws raise the probability of breaking the bottle when Billy is not round throwing rocks at the same time. Her throw is a cause of the breaking. But if Billy also throws, the probability that the bottle shatters, given her throw, will be very low. Again, overdetermination changes the causal scenario in a considerable way. While asymmetric overdetermination presents a case of probabity-raising without causation, symmetric ovedetermination illustrates causation without probability-raising.

Other accounts of causation, like regularity and dispositionalist theories, are also affected by overdetermination (see Lewis 1973, Maslen 2012 and Hitchcock 2013). Given that overdetermination affects somehow a general notion of causation, it might also affect theories in which the causal concept plays an important role. One particular theory that has such

characteristic is causal decision theory. In general, such theory recommends performing actions that tend to cause valuable outcomes. I will now explore a case that might be problematic to this theory.

## 2. Decisional Overdetermination

Before getting into some details of causal decision theory, I will describe a decision scenario in which overdetermination is present, called *decisional overdetermination*. Here are the different types:

**Decisional symmetric overdetermination**

An agent, Suzy, whose only interest is that the bottle in front of her breaks, has two options. She can either throw a stone at the bottle (T) or omit the throw (-T). She knows that if she throws the stone, Billy's stone and her stone will break the bottle simultaneously. Should she throw the stone at the bottle?

**Decisional asymmetric overdetermination**

The agent can either throw or abstain of throwing. Billy's stone will break the bottle in a second instance anyway, if she does not throw. Should she throw?

Two types of decisional asymmetric overdetermination can also be distinguished depending on whether the interruption occurs *early* or *late*. *Decisional trumping* scenarios may also be considered. Suppose that the officer wants the bottle to be shattered. He can either shout at the soldiers to break the bottle, knowing that the sergeant will also shout at the same time, or stay silent. The sergeant will shout the order anyway. Should the officer shout to the soldiers? I am not going to analyze decisional trumping cases in depth here. I assume that, for any account of expected utility, the recommendation in decisional trumping cases will be the same as in decisional symmetric overdetermination and decisional late preemption. The reason for this assumption is not necessarily the simultaneity of the potential causes, but, more fundamentally, the causal independence between both, which is implied by simultaneity. Thus, the relevant distinction will be between decisional symmetric overdetermination and decisional asymmetric overdetermination. However, it might also be helpful to distinguish the latter in decisional early and late preemption.

Here is the outcome's matrix for all cases of decisional overdetermination, which contrast the values of the bottle breaking (B) with the values of the bottle remaining unbroken (-B). As usual, the rows represent the agent's possible actions, while the columns represent the different outcomes:

| | **B** | **-B** |
|---|---|---|
| **T** | 1 | 0 |
| **-T** | 1 | 0 |

Before evaluating decisional overdetermination from the point of view of causal decision theory, let us see whether the classical version of evidential decision theory recommends the rational thing to do. I assume that every account of rational decision is based on the maximization of some notion of expected utility. Thus, when defining some theory of decision I might leave that unmentioned and only provide the definition of expected utility that characterizes it.

**Evidential Decision Theory**

The most rational option is to perform the action that has the higher expected utility. This is defined as the sum of the products of the outcome's conditional probabilities, given that the action is performed, and the values of those outcomes.

$EU(A) = \Sigma_j P(O_j|A)V(AO_j)$

We can easily see that in a situation of symmetric decisional overdetermination, the theory recommends indifference. On the one side, the expected utility of throwing is near to one, if the probability that the bottle breaks, given the agent's throw, is also extremely high, which seems natural.

$EU(T) = \Sigma\ P_j(B_j|T)V(TB_j) = 1$

$EU(-T) = \Sigma\ P_j(B_j|-T)V(-TB_j) = 1$

On the other side, the expected utility of not throwing is also near to one. For the probability that the bottle breaks, given that the agent does not throw, is also very high; Billy will surely break the bottle with his stone.

In the case of asymmetric decisional overdetermination the expected utilities are also equal. The bottle is going to be shattered by Billy anyway and evidential decision theory recommends indifference again. It must be noticed that this verdict is the same in cases of decisional *early* preemption as in cases of decisional *late* preemption. The conditional probabilities will be equally high in both cases.

Let us now consider causal decision theory, starting with one of its first and most general versions.

**Causal Decision Theory**

Gibbard and Harper (1978) developed a notion of expected utility based on a solution made by Stalnaker (1972), according to which conditional probabilities are no longer crucial. Instead, probabilities of conditionals should be considered. The particular sort of conditionals are counterfactuals ($\Box\!\!\rightarrow$). Thus, the expected utility of an act is defined as the sum of the products of the probability that the different outcomes would occur, if the action was performed, with the values of those outcomes:

$EU(A) = \Sigma_j P(A \mathbin{\Box\!\!\rightarrow} O_j)V(AO_j)$

Other accounts of causal decision theory were developed, prescribing the same results as Gibbard and Harper's version (Skyrms 1980, Lewis 1981). However, I am going to work with the one defined above. How does this theory confront decisional overdetermination? In the case of *symmetric decisional overdetermination*, causal decision theory recommends indifference as well. Both the expected utility of throwing and of not throwing approximate to one, because the bottle would break anyway, either by the agent's throw or her not throwing. Billy is throwing the stone at the same time the agent throws hers, which means that the bottle will surely break, even if the agent does not throw. *Asymmetric decisional overdetermination* does not make a difference in the expected values. The bottle would be broken by Billy's stone, if the agent decided not to throw. Thus, the counterfactual holds. Again, as well as in evidential decision theory, causal decision theory based on counterfactuals does not distinguish between early and late preemption. Indifference is recommended in both cases. Such result is already described by Lewis (1981) with the notion of *independence hypothesis*.

## 3. Intuitive Causal Decision Theory

It must be questioned whether the same results are obtained in a version of causal decision theory based on a simple, unanalyzed concept of causation. Instead of asking whether certain outcomes would take place, if a given action was performed, the agent should simply ask whether the action he is evaluating will *cause* the outcomes considered. Ignoring conditional probability, fundamental for evidential decision theory, or the probability of a conditional that could merely point out and perhaps inspire a plausible theory of causation, we have the option to take our most natural and general notion of the causal relation into account.

**Intuitive Causal Decision Theory**

Given an intuitive causal relation, the expected utility of an action is defined as the sum of the products of the probabilities that a certain outcome will be caused by that action, with these outcome's values. The following is the formal definition:

$EU(A) = \Sigma_j P(A \text{ causes } O_j)V(AO_j)$

In simple scenarios, this kind of expected utility delivers the same verdicts as causal decision theory and, even, as evidential decision theory. For instance, if I want to break a bottle (B) by smashing it with a hammer (H), the three kinds of probabilities about the breaking given my action will have, at a first moment, equal degrees, that is, $P(B|H) = P(H \mathrel{\square\!\!\rightarrow} B) = P(H \textit{ causes } B)$ and $P(B|\text{-}H) = P(\text{-}H \mathrel{\square\!\!\rightarrow} B) = P(\text{-}H \textit{ causes } B)$. It also tends to deliver the same as causal decision theory in Newcomb's problem, but I will not discuss this here.

In other situations of simple everyday decision making involving so-called *causation by omission*, intuitive causal decision theory might give the right advice for the wrong reason. Should someone water his plant? The expected utility of watering the plant will be high, given the fact that such action will cause the plant to be healthy. The expected utility of not watering the plant will be much lower. But why is it so? It depends on whether we support a theory that accounts for negative causation or not. If negative causation is real, then one might say that to omit watering the plant would probably cause the plant's death. But if we believe that negative events do not cause anything and prefer to describe the plant's possible death as a product of other positive events like, for example, the complex stress produced in the roots by a very dry soil, then we would not say that to omit watering the plant will cause its death. In this latter interpretation, it is not the omission itself but the conjunct of its *consequences* that causes the undesirable outcome. That is, to water the plant will not be extremely preferableto omitting the watering. Interestingly, the expected utility of letting the plant without water will be different in both cases, depending on our assumptions about negative causation. If we argue that intuitive decision theory recommends watering the plant, the argument would lie on the wrong reasons, namely that the omission of watering the plant can itself be the cause of its death. I will avoid the debate on negative causation, because I think that it is a point where the intuition goes too far (see Beebee 2004 or Dowe 2009). However, negative causation is going to be crucial for an elaborated version of intuitive causal decision theory and will be considered in our further discussion below.

My main interest now is to see whether intuitive causal decision theory gives the right advice in cases of decisional overdetermination, and negative causation might play an important role here. First, in asymmetric decisional overdetermination the probability that the agent's throw *causes* the breaking is near to certainty and thus, the expected utility of throwing will approach to the value of the fact that the bottle is broken. It must now be asked whether it is likely that the agent's *not throwing* can cause the bottle to shatter. Exactly on this point and in contrast to evidential and traditional causal decision theory, this version of expected utility tends to distinguish between early and late decisional preemption scenarios. In the case of early asymmetric overdetermination, the omission of throwing must lead to the signal that communicates to Billy that he can throw. Billy's throw would then cause the breaking. But

would we say that the absence of the agent's throw *causes* the bottle to break? Are we committed to say that absences and omissions can cause something? Again, this topic is not going to be discussed here. Anyway, the signal sending can help to grasp the decision scenario in such a way that one can avoid the debate. In a case of decisional *early* preemption, the agent's options are throwing a stone or sending a signal to Billy (and not throwing). We would not doubt in saying that the agent's signal causes Billy's throw and that Billy's throw causes the bottle to break. By the transitivity of causation, the breaking of the bottle would be caused in both cases by the agent's action. Therefore, the expected utility of throwing equals the expected utility of not throwing, making it indifferent to Suzy, whether she throws or not. This is the same verdict delivered by a traditional counterfactual-based causal decision theory.

In cases of decisional *late* preemption, the probabilities are different. For there is no signal that Suzy sends to Billy and his throw does not depend causally on what the agent does. But the fact that the bottle is not hit by the agent's stone could depend causally, according to our intuitive notion of causation, on the fact that she does not throw. Obviously, this is only plausible under the assumption that omissions and negative events can cause and be caused by others. On the one side, if omissions are considered to be possible causes, then it is possible to think that Suzy's omission of throwing causes the shattering by letting Billy's rock hit the bottle. Thus, intuitive causal decision theory recommends indifference in the scenario of decisional late preemption. On the other side, if omissions are not taken to have causal power or are not considered as possible causal relata, then the theory recommends to throw. This seems to be the wrong advice, since the interests of the agent are focused on the breaking of the bottle, no matter how or by whom it is broken.

The results at a symmetric decisional overdetermination scenario resemble the results of decisional late preemption. In this situation, the agent and Billy throw their stones simultaneously and with the same force. It is assumed that both stones are also going to hit the bottle at the same time. The probability that the agent's throw causes the bottle to break must be high, although it will not shatter the bottle alone. But what is the probability that her omission to throw her rock *causes* the bottle to break? Under a very general and natural notion of causation, it is a very low one, for we would not say that the agent's not throwing can *cause* the bottle to break. Her action is completely independent from Billy's throw, given that both events are simultaneous. Thus, the expected utility of not throwing approaches zero and, clearly, is much lower than the expected utility of throwing. It seems that causal decision theory recommends wrongly again. I am going to clarify later how this unexpected result can be rationalized. As was already mentioned, intuitive causal decision theory will recommend the same in decisional trumping as in decisional symmetric overdetermination.

There are cases where intuitive causal decision theory might give simply an irrational advice. Suppose that Suzy is deliberating about throwing the stone at a bottle or not in a scenario of decisional asymmetric overdetermination. This time she knows that she has a bad run at throwing rocks and it is unlikely that she will hit the bottle if she throws. But she also knows that it is very likely that Billy hits it if he throws. Suppose further that Billy will not throw if she does. Since it is less probable that the omission of the throw *causes* Billy's rock to break the bottle, the expected utility of omitting the throw will be lower than the expected utility of throwing.

EU(T) = P(T *causes* B)V(TB)

EU(-T) = P(-T *causes* B)V(-TB)

EU (T) > EU(-T)

However, if we take the definition of expected utility used by evidential decision theory, the inequality favors not throwing:

EU(T) = P(T|B)V(TB)

EU(-T) = P(-T|B)V(-TB)

EU (T) < EU(-T)

The probability that the bottle breaks given the agent's throw is lower than without her throw, because Billy, who is in a good run, will almost surely break it. This result should be the same in the framework of traditional causal decision theory. Thus, evidential decision theory and traditional causal decision theory recommend omitting the throw and letting Billy break the bottle, while intuitive causal decision theory recommends, maybe irrationally, to throw.

Let me go back and take an overview of our results. The following table shows the different results that the theories considered until now give in decisional overdetermination scenarios:

| | **Symmetric overdetermination** | **Asymmetric overdetermination** | |
|---|---|---|---|
| | | ***Early*** | ***Late*** |
| **EVT** | Indifference | Indifference | Indifference |
| **CDT** | Indifference | Indifference | Indifference |
| **ICDT** | Throw | Indifference | Throw (-NC) |

Recapitulating and describing the table, we can say that decisional overdetermination cases do not present differences between traditional accounts of evidential (EDT) and causal decision theory (CDT). Nevertheless, there is a clear difference between intuitive causal decision theory (ICDT) and the other two mentioned theories, especially in scenarios of decisional symmetric overdetermination. Another distinction occurs in cases of decisional asymmetric overdetermination. While both evidential and traditional causal decision theories do not recommend differently in cases of early and late preemption, intuitive causal decision theory distinguishes those variations. For, on the one hand, the two traditional theories prescribe indifference, no matter whether it is a scenario of early or late preemption. On the other hand, intuitive causal decision theory prescribes indifference only in decisional early preemption, recommending throwing in cases that involve late asymmetric overdetermination, if negative causation (NC) is not assumed literally.

## 4. Contrastive Causation

It is not hard to find other theories that might be compatible in some cases with the results given by traditional accounts. Take, for instance, a definition of expected utility based on a *contrastive notion of causation* (see Hitchcock 1996, Schaffer 2005, Northcott 2008), where EU(A) = $\Sigma_j$ P(A *rather than* C *causes* $O_j$) and C, a set of possible causes of the considered outcome different from the agent's action and its effects. For every action A, the consequences of its omission, i.e. of -A, are included in C. One can argue on whether the contrast class must be built around both the cause and the effect (Schaffer 2005) or just around the cause (Hitchcock 1996). For purposes of rational decision, only the side of the cause, i.e. of the action, is contrastive explicitly. The contrast class of the effect is already given by the outcomes' partition used in the definition of expected utility.

How do we interpret the contrast between different possible causes? The fact that, for instance, a hurricane rather than an earthquake caused the house to collapse simply means that the hurricane caused the collapse and the earthquake did not. Now, the fact that the earthquake did not cause the collapse either means that an earthquake occurred that was not strong enough (nor sufficient in any other sense) to cause the collapse or that there wasn't any

earthquake at all. The specificity of the contrast class may be determined by the agent's interests.

The recommendation under such account will be, at least at decisional early preemption scenarios, the same as in traditional causal decision theory. For it seems to be very likely that the agent's throw rather than Billy's throw causes the bottle shattering. On the other hand, it is highly probable that her omission of throwing and the sending of the signal rather than throwing cause the breaking. Indifference is, thus, recommended. In decisional late preemption, this definition of expected utility may recommend to throw, depending on our assumptions about the intuitive concept of the causal relation.

At decisional symmetric overdetermination, it seems implausible to think that the agent's throw *rather* than Billy's throw can break the bottle, assuming that Billy will throw anyway. It is also very unlikely that the agent's omission of the throw and its consequences rather than throwing can cause the breaking of the bottle. That is, the theory would recommend indifference. But this is so because both options have very low expected values, which is a very different reason than the one that takes traditional accounts to the same results. This is the same result of a decision theory based on possible causation, discussed below. I am not going to discuss further details about how an approach of *contrastive causal decision theory* could be elaborated, and I will assume that the recommendations such a theory would give depend on parameters that are either crucial for traditional accounts of rational decision (e.g. the definition of the contrast class) or for intuitive causal decision theory (e.g. the definition of the causal relation).

## 5. Does Intuitive Causal Decision Theory get the Right Notion of Causation?

As we have seen, a traditional version of causal decision theory is not actually based on causation, but on the counterfactual relation between options and possible outcomes. Hitchcock *(2003)* examines the vulnerability of traditional causal decision theory confronting overdetermination scenarios, concluding, as it was shown above, that such cases are far from being problematic.

Instead of simply saying that causal decision theory should not be considered as an account based on the notion of causation, Hitchcock's diagnosis is that the notion on which the theory is based is just a different one. A traditional account of causal decision theory is not based on a concept of *actual causation* supported by the counterfactual analysis. Instead, it uses the notion of the causal relation implied by the counterfactual conditional $A \mathrel{\Box\!\!\rightarrow} O$, called *causal dependence*. The main difference that Hitchcock presents between actual causation and causal dependence is the fact that the former is a retrospective notion, while the latter is a prospective one. That is, when one evaluates whether *c* causes *e* using actual causation, one refers to past events. Both events *c* and *e* already occurred and if *c* had not occurred, *e* would not have occurred. But when we evaluate whether *c* causes *e* using causal dependence, we think of future counterfactual events. Both events *c* and *e* have not occurred yet and if *c* occurred, then *e* would occur.

I think that Hitchcock's distinction makes the case more understandable. But understandability does neither imply accuracy nor relevance. After all, every agent should assume that the actions he is deliberating about have not occurred yet. Otherwise he would just be wasting his time giving very extreme probabilities to the different possible outcomes. Another reason to rethink the relevance of the distinction between retrospective and prospective causation is to ask whether temporal precedence is a necessary assumption for the causal analysis or whether the notion of time must be included at all as a condition in the definition of the causal relation. The distinction between actual causation and causal

dependence assumes temporal order, but I think that one of the virtues of the counterfactual analysis of causation is the way it handles with temporal asymmetry without presupposing it. If the conditional -C $\Box\rightarrow$ -E is true, the conditional -E $\Box\rightarrow$ -C might be false. The non-backtracking feature of these counterfactuals clarifies the precedence of the cause without postulating it. Why should we do this when we face decisional overdetermination scenarios, if these look harmless to our traditional account of causal decision theory?

Somehow, decisional overdetermination demands something from the theory. My diagnosis is, thus, double. Either traditional causal decision theory is not affected by decisional overdetermination, but we should strengthen our notion of causation to explain that, which shows that decisional overdetermination has some serious consequences after all, or traditional causal decision theory is indeed directly affected by such scenarios. In both cases, the causal notion involved in causal decision theory must be reconsidered.

## 6. Possible Causation in Expected Utility

As I have already shown, while decisional overdetermination seems not to be problematic for traditional causal decision theory, it is only problematic for a causal theory of decision based on some intuitive concept of causality. I have not said much about that concept, nor is it necessary to say much, since it must be a mere product of our most general intuitions. But is not the fact that the effect would not have happened without the cause a feature of the causal relation that fits our most general understanding of it? That is actually another virtue of the counterfactual account of causation. Let us consider a notion of expected utility construed with probabilities of causal facts between actions and outcomes, as well as in our intuitive version of causal decision theory, but using the counterfactual analysis of causation. In order to do that we cannot just take the probability of the conditional -A $\Box\rightarrow$ -O, because the counterfactual analysis of causation also presupposes that both events involved in the relation already occurred. Thus, we may have to face the following problematic definition, involving a would-cause counterfactual:

$$EU(A) = \Sigma_j P[A \,\&\, O_j \Box\rightarrow (\text{-}A \Box\rightarrow \text{-}O_j)]V(AO_j)$$

The agent must evaluate the probability of the fact that if he had performed some action and some outcome occurred, then that outcome would not have occurred without his action. The counterfactual considered turns out to be vacuously true, for, on the base of the importation principle (McGee 1985), the antecedent of the equivalent conditional (O & A & - A) $\Box\rightarrow$ -O is impossible. One should actually construct a definition of expected utility based on this latter conditional, but this is not going to be considered here, as the results will be the same. Since the validity of the importation principle can be put in doubt under specific assumptions (see Arló-Costa's work on epistemic conditionals [2001]), I am not going to take it here as a strong and uncontroversial assumption.

There are plenty of common cases where expressions of embedded conditionals of the form A $\Box\rightarrow$-A $\Box\rightarrow$-B do make sense. I think that one could make perfectly sense of that using traditional possible world semantics for counterfactuals (either Stalnaker's 1968 or Lewis's 1979). The relevant changes that such a semantic should confront in order to make sense of embedded counterfactuals are precisions regarding the notion of laws of nature and their violations. I am not going into these details now (see Dowe 2009). I will assume, however, that a counterfactual conditional is true, if and only if in the closest world (or in any of the closest worlds) where the antecedent is true, the consequent is also true.

For example, if I had a dry match, then, if it was not dry, it would not light. In the actual world, $w_1$, seeing a match box with a single wet match in it, I wonder what would happen if it was a dry match. Thinking about the constitution of the closest world (or worlds) where it is a

dry match, I wonder whether the match would light, if it was not dry. In order to do this, I have to speculate, this time from $w_2$, about the closest world in which the (counterpart of or the same) match is not dry. The hasty response would be to think that such a world is the actual world. But that is not necessarily so. For the physical changes that distinguish the actual world from $w_2$ might be simpler than the changes that would take us from $w_2$ back to the actual world. It might be that by drying the match, the surroundings were also dried and to consider a world where the match is not dry would imply to change the entire environment. In such a case, the closest non-A-world is not the actual world. Hence, the iterated counterfactual of the form A $\Box\rightarrow$ -A $\Box\rightarrow$ B cannot be converted to a conditional with a contradictory antecedent. The occurrence of A, the fact that the match is dry, is assumed from $w_1$, but its negation is not. The occurrence of the fact that the match is not dry, is assumed from the perspective of $w_2$. This means that A & -A are not assumed in $w_1$ either. The closest world to $w_2$ in which the match is not dry might be different from the actual and since it is true that the match won't light there, the expression with the embedded counterfactual makes perfect sense:

(The match in the box is dry) $\Box\rightarrow$ (The match is not dry) $\Box\rightarrow$ (The match does not light)

That is, if the match in the box was not dry, it would not light. Let us clarify the evaluation of the expressions involved from the perspective of the corresponding possible worlds, assuming, just to make the case clearer, that for every world there is only one closest possible world:

| | |
|---|---|
| $\mathbf{w_1}$: -A | The match is not dry. |
| $\mathbf{w_2}$: A | The match is dry. |
| $\mathbf{w_3}$: -A & B | The match is not dry and it does not light. |
| $\mathbf{w_1}$: A $\Box\rightarrow$ -A $\Box\rightarrow$ B | If the match was dry, then if it was not dry, it would not light. |

Again, the closest possible world to the actual world, in which A is true, is the possible world $w_2$. Nevertheless, in $w_2$, the closest possible world where A is false is not the actual world, but $w_3$. In that possible world, B is also the case. Thus, the iterated counterfactual conditional is true in the actual world.

Let us now consider the recommendations for decisional overdetermination by a causal decision theory based on *possible causation* with embedded counterfactuals supported by a would-cause semantics (see Dowe 2009). In the case of *decisional symmetric overdetermination*, the expected utility of throwing is as low as the expected utility of omitting the throw. For, on the one side, given that the agent throws and the bottle shatters, the bottle would still have shattered, if she had not thrown. On the other side, given that the agent did not throw and the bottle shattered, it would have shattered, if she had thrown. The theory recommends indifference, which fits to the traditional account of causal decision theory, but differs from the intuitive approach. In decisional asymmetric overdetermination scenarios, the theory prescribes indifference again.

The account based on possible causation gives the same recommendations as traditional causal decision theory. However, the reasons for this are, as in contrastive causal decision theory, extremely different for each theory. According to traditional causal decision theory, the bottle would break, no matter what the agent did, i.e. the probability that the bottle shatters is for both options very high. But for a decision theory based on possible causation, none of the possible actions would cause the breaking. The bottle might break, when Suzy throws, but the throw cannot count as a cause. The bottle would still have shattered, if the throw had not occurred.

We have a new reason to think that counterfactuals do not give an intuitive notion of causation for a theory of rational decision, neither used for definitions of causal dependence, nor for actual causation. As it has been mentioned, this also shows that a general theory of causation that pretends to grasp our intuitive concepts of that relation cannot be based only on counterfactual dependence.

## 7. Ranking Theory and Decisional Ovedetermination

There is a definition of expected utility that might, apparently, share some results with intuitive causal decision theory. I will consider such a definition with the ranking functional notion of causation at its basis (Spohn 2006, 2012). The ranking theoretic approach of causation proposes a clear solution to overdetermination cases that, instead of postulating fine-grained events (one of the most plausible solutions to that problem), suggests considering a *conceptual* fine-graining. Put in a very informal way the variable C is a *direct cause* of variable E, if and only if both C and E occur, C precedes E and C is a *reason* for E, given the obtaining circumstances (Spohn 2012b: 354). Taking a negative ranking function $\kappa$ that measures the degrees of disbelief for propositions, such that $\kappa(A) = 0$ expresses that A is disbelieved, the belief in C is a reason for E if and only if $\kappa(-E|C) > \kappa(-E|-C)$, i.e. if the absence of E has a higher degree of disbelief, given C's occurrence, than given its absence.

Using a positive ranking function $\beta$, such that $\beta(A) = \kappa(-A)$, and a two-sided function $\tau$, such that $\tau(A) = \beta(A) - \kappa(A)$, overdetermination cases can be described with clarity. In cases of symmetric overdetermination, the belief that the bottle breaks (B), given the fact that both Suzy ($T_1$) and Billy ($T_2$) threw their stones, is higher than the belief on the breaking, given only one throw. That is, for instance, $\tau(B|T_1 \& T_2) = 2 > \tau(B|T_1 \& -T_2) = 1 = \tau(B|-T_1 \& T_2) > \tau(B|-T_1 \& -T_2) = -1$. Clearly, it is highly disbelieved that the bottle breaks, given the fact that none of them throws a stone. Cases of early asymmetric overdetermination are clarified simply by appealing to the transitivity of the causal chain going from the actual cause to the effect, arguing that although the shattering does not depend counterfactually on Suzy's throw, it depends on some state of the stone before it hits the bottle and after the instant at which Suzy would have given the signal to communicate that she was not going to throw. Cases of late preemption are more problematic and find hardly an interpretation on this theory (Spohn 2012: 367).

To avoid getting into the debate about whether asymmetric overdetermination is really a problem for causation or not, let us go directly to our topic, namely decision theory, and see what this framework can do about the scenarios we are struggling with. I am going to consider a definition of expected utility based on the positive ranking function:

$$EU(A) = \Sigma_j \beta(O_j|A)V(AO_j)$$

A theory of decision that takes rational actions as the ones that maximize such notion of expected utility into account recommends the usual in decisional early preemption scenarios, namely indifference. For the fact that the agent's rock breaks the bottle does not have to be more believed than the fact that Billy's rock breaks it. Billy will not throw, if the agent does, so the degree of belief on the bottle shattering must not be too high. Let us say that $\beta(B|T) = \beta(B|-T) = 1$. I am not going to consider decisional late preemption for the already mentioned reason.

The scenario of decisional symmetric overdetermination is more interesting, since the account of expected utility based on ranking functions recommends the same as intuitive causal decision theory. For it is more believed that the bottle shatters (S), given the fact that both the agent and Billy throw their stones, than just given Billy's throw (B), i.e. $\beta(S|T \& B) > \beta(S|-T \& B)$. Therefore, the expected utility of throwing (T) is higher than the expected utility of omitting the throw. This case is also an example of one of the most important differences

between ranking theory and probabilistic causation, considering that P(S|T & B) = P(S|-T & B) and assuming that both the agent and Billy are well-trained stone-throwers. As expected, ranking causal decision theory also recommends throwing in decisional trumping. It also delivers the right verdict in the bad run scenario described above, in which intuitive causal decision theory fails.

In this way, the ranking account of causation does not only analyze cases of redundant causation correctly; it can also ground a notion of expected utility that adjusts to intuitive causal decision theory. The fact that a decision theory based on ranking functions is more similar to a decision theory based on the intuitive concept of causation than to other accounts that give different recommendations in particular scenarios means that the ranking theoretic framework of causation fits better our general intuitions about the causal relation.

Other accounts of causation, different from the ranking functional approach, might also fulfill the requirements for intuitive causal decision theory. Some of these are the causal modeling approach (Pearl 2000; Spirtes, Glymour & Sheines 2000), the theory of causal processes (Salmon 1984, Dowe 2000) and the dispositionalist theory of causation (Harré & Madden 1975, Mumford & Anjum 2011). There are not enough reasons to think that these accounts of causation exclude each other in a fundamental and relevant way. I will leave aside this time the discussion of such differences and on how these theories may support intuitive causal decision theory.

## 8. Is there Irrationality in the Intuition?

Now that it has been explained how traditional accounts of decision making do not fit our intuitive notion of causation and that intuitive causal decision theory (and some specific accounts supporting it) might serve as an alternative, it must be put in doubt whether the intuitive account is really a practical guide for rational decision. Is it really rational to throw a stone at the bottle in a scenario of decisional symmetric overdetermination? Apparently, the recommendations given by the traditional accounts are right in saying that it does not matter, because Billy will also throw and the bottle will surely get broken, which is the agent's only interest. The decisional overdetermination scenarios described above do not consider the energy lost by throwing stones, which would be a reason to twist the agent's indifference. Suzy should not throw in such cases. Anyway, these situations, far from being problematic, do not worry us now.

There are two reasons that could explain why the prescriptions given by a theory based on an intuitive notion of causation tend to the option of throwing, rather than omitting the throw. The first one, which is clearly captured in ranking causal decision theory, is the fact that two throws ensure the shattering of the bottle more than just one. Even if we assume that it is very likely that nothing will go wrong, a rational agent should not dismiss the mere possibility that something can go wrong. If just one thing goes wrong when both Suzy and Billy are supposed to throw the stones, the bottle will still break. But if just one thing goes wrong after Suzy decides to omit the throw, leaving the responsibility to Billy, the bottle won't shatter. If Suzy does throw, then two things have to fail to expect that the bottle won't break.

The second reason is the very realistic and practical fact that agents weight their beliefs about causal facts according to the *confidence* they give to the possible causes considered. If a decisional scenario ends up in indifference, the agent might evaluate the potential causes of the desired outcome in terms of confidence. To understand the difference between this notion and the degree of belief under which an agent evaluates the act's possible causal relation with the outcomes, it is important to notice that the former only applies directly to the other possible causes, which are different from the evaluated action itself. We can hardly give probabilities to actions (Spohn 1977, Levi 1986); we just suppose what they would cause, if

they were performed. But we can assign probabilities to alternative causes of the relevant outcomes. Confidence is expressed by such degrees of belief on the alternative causes and on the impossibility to attach probabilities to actions. Somehow, in decisional symmetric overdetermination scenarios, any rational agent will choose the actions that are related causally to his desired outcomes, in spite the fact that traditional versions of expected utility may initially promote indifference. If Billy and I are both great at smashing bottles, why should I put more confidence on his throw than on mine? Of course, this self-confident attitude must not always be assumed; it should be just considered as a strategic (and thus, rational) tie-breaking move in cases of initial indifference.

I have shown that decisional overdetermination is not problematic for traditional causal decision theory. Nevertheless, what seems to be threatening by such kind of scenarios is the notion of causation on which traditional definitions of expected utility are based. A version of expected utility based on an intuitive notion of causation differs from traditional accounts. This does not only confirm that the fundamental concepts considered in traditional accounts of expected utility, for example, counterfactual dependence, are far from a transparent definition of causation. It also suggests that a fundamental and general theory of causation should be searched with more independence from those concepts. Furthermore, a theory that tells us how to act rationally should not only consider the importance of an agent's causal influence on the relevant outcomes, but also be based on a notion of causation that clearly explains what is usually understood as a causal relation and what causation really is. An intuitive approach may be a crucial point in reconsiderations of the foundations of causal decision theory, although it should not be the last step.

**Esteban Céspedes**

J. W. Goethe-Universität Frankfurt a.M.
e.cespedes@stud.uni-frankfurt.de

## References

Arló-Costa, H. 2001: 'Bayesian epistemology and epistemic conditionals: On the status of the export-import laws', *Journal of Philosophy* 98, 555-593.

Beebee, H. 2004: 'Causing and Nothingness', in Paul, Hall & Collins (ed.): *Causation and Counterfactuals*. MIT Press.

Dowe, P. 2000: *Physical Causation*. Cambridge University Press.

Dowe, P. 2009: 'Would-cause semantics', *Philosophy of Science* 76, 701-711.

Gibbard, A. & Harper. W. 1978: 'Counterfactuals and Two Kinds of Expected Utility', in Hooker, Leach & McClennen (ed.): *Foundations and Applications of Decision Theory*. Dordrecht: Reidel.

Hall, N. 2004: 'Two Concepts of Causation', in Hall, Paul, & Collins (eds.): *Causation and Counterfactuals*. Cambridge: MIT Press.

Harré, R. & Madden, E. H. (1975). *Causal Powers*. Oxford: Blackwell.

Hitchcock, C. 1996: 'The role of contrast in causal and explanatory claims', *Synthese* 107 , 395 - 419.

Hitchcock, C. 2013: 'What is the 'Cause' in Causal Decision Theory?', Erkenntnis

Levi, I. 1986: *Hard Choices*. Cambridge University Press.

Lewis, D. 1973: 'Causation'. *Journal of Philosophy*, 70, 556-567.

Lewis, D. 1979: 'Counterfactual dependence and time's arrow', *Noûs* 13, 455-476.

Lewis, D. 1981: 'Causal Decision Theory', *Australasian Journal of Philosophy* 59, 5 – 30.

Lewis, D. 2000: 'Causation as influence', *Journal of Philosophy* 97, 182-197.

Maslen, C. 2012: 'Regularity Accounts of Causation and the Problem of Pre-emption: Dark Prospects Indeed', *Erkenntnis* 77, 419-434.

McGee , V. 1985: 'Counterexample to Modus Ponens', *Journal of Philosophy* 82, 462-471.

Mumford, S. & Anjum, R. L. 2011: *Getting Causes from Powers*. Oxford University Press.

Northcott, R. 2008: 'Causation and contrast classes', *Philosophical Studies* 139, 111 - 123.

Nozick, R. 1970: 'Newcomb's Problem and Two Principles of Choice', in Rescher (ed.): *Essays in Honor of Carl G. Hempel*. Dordrecht: Reidel.

Pearl, J. 2000: *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Salmon, W. 1984: *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schaffer, J. 2005: 'Contrastive causation', *Philosophical Review* 114, 327-358.

Skyrms, B. 1980: *Causal Necessity*. Yale University Press.

Spirtes, Glymour & Sheines. 2000: Causation, Prediction, and Search. MIT Press.

Spohn, W. 1977: 'Where Luce and Krantz do really generalize Savage's decision model', *Erkenntnis* 11, 113 - 134.

Spohn, W. 2006: 'Causation: An alternative', *British Journal for the Philosophy of Science* 57, 93-119.

Spohn , W. 2012a: 'Reversing 30 years of discussion: why causal decision theorists should one-box', *Synthese* 187, 95-122.

Spohn, W. 2012b: *The Laws of Belief*. Oxford University Press.

Stalnaker, R. 1972: 'Letter to David Lewis', in Harper, Stalnaker, and Pearce. *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: Reidel.

Stalnaker, R. 1968: 'A Theory of Conditionals', in Rescher. *Studies in Logical Theory*. Oxford: Blackwell.

Stone, J. 2009: 'Trumping the causal influence account of causation', *Philosophical Studies* 142, 153-160.

# Double Effect and Terror Bombing

Ezio Di Nucci

I argue against the Doctrine of Double Effect's explanation of the moral difference between terror bombing and strategic bombing. I show that the standard thought-experiment of Terror Bomber and Strategic Bomber which dominates this debate is underdetermined in three crucial respects: (1) the non-psychological worlds of Terror Bomber and Strategic Bomber; (2) the psychologies of Terror Bomber and Strategic Bomber; and (3) the structure of the thought-experiment, especially in relation to its similarity with the Trolley Problem. (1) If the two worlds are not identical, then it may be these differences between the two worlds and not the Doctrine of Double Effect to explain the moral difference; (2a) if Terror Bomber and Strategic Bomber have the same causal beliefs, then why does Terror Bomber set out to kill the children? It may then be this unwarranted and immoral choice and not the Doctrine of Double Effect that explains the moral difference; (2b) if the two have different causal beliefs, then we can't rule out the counterfactual that, had Strategic Bomber had the same beliefs as Terror Bomber, she would have also acted as Terror Bomber did. Finally, (3) the Strategic Bomber scenario could also be constructed so as to be structurally equivalent to the Fat Man scenario in the Trolley Problem: but then the Doctrine of Double Effect would give different answers to two symmetrical cases.

Since even before WWII[1], the discussion of the *Doctrine of Double Effect* (DDE[2]) has been intertwined with the discussion of *terror* bombing and *strategic* bombing.[3] The concepts of

[1] For the earliest examples known to me, see Ryan 1933 and Ford 1944. Gury also talks about the killing of non-combatants in the context of his seminal discussion of double effect (see Boyle 1980: 528-29).

[2] Here I will just assume previous knowledge of the Doctrine of Double Effect, and restrict my discussion of the actual principle to this footnote with the following representative definitions:

- McIntyre in the Stanford Encyclopedia of Philosophy: "sometimes it is permissible to bring about as a merely foreseen side effect a harmful event that it would be impermissible to bring about intentionally" (http://plato.stanford.edu/entries/double-effect/);
- Woodward in the Introduction to his standard anthology on DDE: "intentional production of evil... and foreseen but unintentional production of evil" (2001: 2);
- Aquinas, which is often credited with the first explicit version of DDE: "Nothing hinders one act from having two effects, only one of which is intended, while the other is beside the intention" (Summa II-II, 64, 7);
- Gury: "It is licit to posit a cause which is either good or indifferent from which there follows a twofold effect, one good, there other evil, if a proportionately grave reason is present, and if the end of the agent is honourable – that is, if he does not intend the evil effect" (Boyle's translation 1980: 528);
- Mangan: „A person may licitly perform an action that he foresees will produce a good and a bad effect provided that four conditions are verified at one and the same time: 1) that the action in itself from its very object be good or at least indifferent; 2) that the good effect and not the evil effect be intended; 3) that the good effect be not produced by means of the evil effect; 4) that there be a proportionately grave reason for permitting the evil effect" (1949: 43).

I have discussed other aspects of double effect elsewhere: Di Nucci (2012), Di Nucci (f0rthcoming, a), Di Nucci (forthcoming, b), Di Nucci (dx.doi.org/10.2139/ssrn.1930832), Di Nucci (submitted, a), Di Nucci (submitted, b), and Di Nucci (book manuscript).

[3] Here the terminology is a bit confusing: in modern philosophical discussions, the talk is always of 'terror' bombing and 'strategic' or 'tactical' bombing. Some (such as for example Cavanaugh 2006: xii) distinguish between 'strategic' and 'tactical' on historical grounds, finding the latter more appropriate. Others (such as for example Ford 1944: 263) object to both 'strategic' and 'tactical' and opt for 'precision' bombing. Other terms for 'terror' bombing are 'obliteration' bombing, 'area' bombing, and 'indiscriminate' bombing (Walzer 1991: 11). To make matters more confusing, the adjective 'strategic' is

'terror bombing' and 'strategic bombing' are, both in historical and philosophical context, quickly clarified by looking at how the British changed their directives to their pilots sometime in late 1940. Frankland writes that in June 1940 British authorities still "specifically laid down that targets had to be identified and aimed at. Indiscriminate bombing was forbidden." (1970: 24[4]) Here indiscriminate bombing is what has come to be known in the literature as terror bombing. And it has presumably acquired that name because the British soon changed their fighting ways: already in November 1940 "Bomber Command was instructed simply to aim at the center of a city... the aiming points are to be the built-up areas, not, for instance, the dockyards or aircraft factories" (1970: 24) And built-up areas here means residential areas, as the British did not care to hide: Churchill spoke in the Commons of the "the systematic shattering of German cities." (July 1943[5]); "the progressive destruction and dislocation of the German military, industrial and economic system and the undermining of the morale of the German people to the point where their capacity for armed resistance is fatally weakened." (joint British-American Casablanca conference); "To the RAF fell the task of destroying Germany's great cities, of silencing the iron heart-beat of the Ruhr, of dispossessing the working population, of breaking the morale of the people" (*Target: Germany*, an RAF official publication of that period). Finally they ended up calling it 'terror' bombing themselves: "Here, then, we have *terror* and devastation carried to the core of a warring nation." (Still from *Target: Germany* as quoted by Ford 1944: 294).

## 1. The Thought-experiment

The British started with what in contemporary literature we refer to as strategic bombing and then turned to so-called terror bombing. As we have seen (Ford 1944), the connection between these practices and the DDE was drawn already at the time. In the post-war period, the distinction between terror bombing and strategic bombing has evolved into a philosophical thought-experiment widely used to illustrate (and often also to defend) DDE. An influential example is Jonathan Bennett's discussion in his *Tanner Lectures on Human Values*:

> In this lecture I shall exhibit some difficulties about a certain distinction which is thought important by many moralists - namely that between what you intend to come about as a means to your end and what you do not intend although you foresee that it will come about as a by-product of your means to your end. This has a role in most defences of the Doctrine of Double Effect, and is one source for the view that terror bombing is never permissible though tactical bombing may sometimes be - i.e., that it is never right to kill civilians as a means to demoralizing the enemy country, though it may sometimes be right to destroy a munitions factory as a means to reducing the enemy's military strength, knowing that the raid will also kill civilians. In the former case - so the story goes - the civilian deaths are intended as a means; in the latter they are not intended but merely foreseen as an inevitable by-product of the means; and that is supposed to make a moral difference, even if the probabilities are the same, the number of civilian deaths the same, and so on. (1980: 95)[6]

The similarity between Bennett's characterization of terror bombing and the British directives from WWII is striking: "to kill civilians as a means to demoralizing the enemy country" is

---

sometimes used for 'terror' bombing as well. I stick to 'terror' bombing and 'strategic' bombing throughout because it is the most common usage in the literature (as a brief Google search revealed).

[4] Reference found in Walzer (1971: 11).

[5] This and the following quotes are taken from Ford 1944: 262 ff.

[6] To be sure: Bennett is a critic of DDE, but he has contributed decisively to the establishment of the thought-experiment as a standard one. See also Bennett 1995.

offered as an example of terror bombing; strategic bombing is described as "to destroy a munitions factory as a means to reducing the enemy's military strength, knowing that the raid will also kill civilians". The case we are asked to imagine is, supposedly, one in which a pilot is ordered to bomb a munitions factory, so as to reduce the enemy's military strength; she is also informed that there is a very high probability of civilian casualties as a result of the bombing of the munitions factory. The day after the same pilot is ordered to bomb civilians as a means to demoralize the enemy; she is informed that there is a very high probability (the same very high probability as yesterday) that the numbers of civilian deaths will be the same as yesterday.[7] Now the idea that DDE is supposed to defend is that it is permissible on the first day but not on the second day for the pilot to drop her bombs.[8]

Michael Bratman develops this very scenario as follows:

> Both Terror Bomber and Strategic Bomber have the goal of promoting the war effort against Enemy. Each intends to pursue this goal by weakening Enemy, and each intends to do that by dropping bombs. Terror Bomber's plan is to bomb the school in Enemy's territory, thereby killing children of Enemy and terrorizing Enemy's population. Strategic Bomber's plan is different. He plans to bomb Enemy's munitions plant, thereby undermining Enemy's war effort. Strategic Bomber also knows, however, that next to the munitions plant is a school, and that when he bombs the plant he will also destroy the school, killing the children inside. Strategic Bomber has not ignored this fact. Indeed, he has worried a lot about it. Still, he has concluded that this cost, though significant, is outweighed by the contribution that would be made to the war effort by the destruction of the munitions plant. Now, Terror Bomber intends all of the features of his action just noted: he intends to drop the bombs, kill the children, terrorize the population, and thereby weaken Enemy. In contrast, it seems that Strategic Bomber only intends to drop the bombs, destroy the munitions plant, and weaken Enemy. Although he knows that by bombing the plant he will be killing the children, he does not, it seems, intend to kill them. Whereas killing the children is, for Terror Bomber, an intended means to his end of victory, it is, for Strategic Bomber, only something he knows he will do by bombing the munitions plant. Though Strategic Bomber has taken the deaths of the children quite seriously into account in his deliberation, these deaths are for him only an expected side effect; they are not – in contrast with Terror Bomber's position – intended as a means... In saying this I do not deny that Strategic Bomber kills the children intentionally. (1987: 139-140)[9]

---

[7] The epistemic characterization is here important, but it can vary: we can talk of certainty, high probability, or even just possibility, as long as there is no epistemic gap between the two cases.

[8] As I already said, here I will not get into issues of interpretation of DDE. Let me just say that moral permissibility is both the strongest and most common interpretation of DDE (see Boyle 1980 for an argument as to why we should interpret DDE this way); alternative interpretations may involve different attributions of responsibility, excuse as opposed to justification, or different sentencing. At the other end of the spectrum we find the claim that not even the action-theoretical distinction upon which DDE is found is a legitimate one (this last possibility is discussed here too).

[9] From the point of view of military ethics in general and just war theory in particular, there is an important difference between talking about 'civilian casualties' in general, as Bennett does, and talking about school children, as Bratman does. The civilian casualties referred to by Bennett may very well be the munitions factory workers, and their moral status is controversial. On this, see debates on non-combatants, civilians-m, and civilians-w (where 'm' and 'w' distinguish between those civilians which provide military equipment such as munitions and those which provide welfare equipment such as food); in particular, see Fabre 2009 and McMahan 2009. While Bennett's reference to 'civilian casualties' may be a reference to civilians-m who may actually turn out to be liable to attack, Bratman's reference to school children simplifies the thought-experiment by providing a group (school children) which none of the contrasting views would consider liable to attack. That is why I shall stick to Bratman's school children throughout, which help identify the DDE debate on terror bombing and strategic bombing as independent from the non-combatant debate.

The philosophical discussion of terror bombing and strategic bombing starts with the intuition that there is a moral difference between them; indeed, the Doctrine of Double Effect is normally offered as an explanation of the moral difference between Terror Bomber and Strategic Bomber. Elsewhere (Di Nucci, submitted b), I have looked at this supposed moral intuition experimentally and found no evidence for it. Here I concentrate on theoretical considerations and offer three arguments against DDE's explanation of the thought-experiment. I show that, once the thought-experiment of terror bombing and strategic bombing is properly analysed, it should really be no surprise that there is no intuitive moral difference between Terror Bomber and Strategic Bomber: depending on how some crucial underdetermined aspects of the thought-experiment are interpreted, either the relevant differences around which the thought-experiment is constructed (such as intending/merely foreseeing and means/side-effects) do not explain the supposed moral differences or there are, indeed, no such moral differences – as the evidence from intuition suggests.

## 2. Bratman and Different Options

Reading Bratman's version of the thought-experiment, one may think that we are in a twin thought-experiment, where everything is identical apart from the plans of the two pilots. But what Bratman writes after a few pages indicates that this is not what he meant:

> ...this does not tell us whether or not Strategic Bomber would also go ahead and bomb if his bombing option were precisely that of Terror Bomber's. The difference between Strategic Bomber and Terror Bomber in the original case lies in the options with which they are presented; it need not involve a difference in inclination to plump for terror-bombing if that is the only bombing option available. (1987: 161)

Bratman's though-experiment, then, is not only different in the psychology of the two pilots; it is also different in the options available to them; which means, supposedly, that the difference between the Terror scenario and the Strategic scenario goes beyond psychological differences between Terror Bomber and Strategic Bomber. From the way both Bennett and Bratman describe the though-experiment it would have been legitimate to suppose, for example, that the consequences of the bombings would be identical: both Terror Bomber and Strategic Bomber destroy the munitions factory, both kill the same number of children. But actually there is no munitions factory in the world of Terror Bomber, otherwise we could not make sense of the above remark that "this does not tell us whether or not Strategic Bomber would also go ahead and bomb if his bombing option were precisely that of Terror Bomber's". That the difference between Terror Bomber and Strategic Bomber need not involve "a difference in inclination to plump for terror-bombing if that is the only bombing option available" suggests that DDE may have to argue for the permissibility of what Strategic Bomber does even in the case in which Strategic Bomber would have behaved exactly as Terror Bomber had he been faced with the options that Terror Bomber was faced with.[10] We will see in Section 4 that this is a problematic position to defend.

Let us take stock: we have identified the classic terror-strategic thought-experiment as being underdetermined in a first important respect: the options with which Terror Bomber and Strategic Bomber may be presented need not be identical, as long as they kill the same number of children (or some such). This is left open to the extent that Bratman, for example, allows for the possible counterfactual in which Strategic Bomber would admit that, had she been presented with the options Terror Bomber was presented with, she would have done just

[10] This point does not depend on claiming that there is no munitions' factory in the world of Terror Bomber. The same point can be made by supposing that there is a munitions factory but that Terror Bomber does not know that or that the orders Terror Bomber receives do not mention one (this fits Bratman's talk of 'options').

what Terror Bomber has done. This first point, then, can be summarized by saying that the thought-experiment is underdetermined as to the non-psychological differences between the two scenarios.

There is also an important underdetermination as to the psychological differences between the two agents, which I discuss in the next two sections: it may be that Terror Bomber and Strategic Bomber have the same causal beliefs; or it may be that they have different causal beliefs. Let us begin with discussing the variant in which the two pilots have the same causal beliefs.

## 3. Same Causal Beliefs

Let us suppose that the two agents, Terror Bomber and Strategic Bomber, have the same causal beliefs[11]: of the sixteen possible permutations resulting from combining the two agents with the two beliefs 'killing children will weaken enemy' and 'destroying munitions will weaken enemy' (and their respective negations), twelve involve at least one of the two agents in some form of irrationality – I will therefore disregard those even though some of them are such that the two agents have the same causal beliefs.[12] Of the remaining four, three are such that the two agents have different causal beliefs. So there is only one permutation such that neither of the agents is irrational and the two agents have the same causal beliefs, the following:

> Terror Bomber believes that killing children will weaken enemy and she believes that destroying munitions will weaken enemy.
>
> Strategic Bomber believes that destroying munitions will weaken enemy and she believes that killing children will weaken enemy.

Here there is both a cognitive problem and a normative problem. In brief, the cognitive problem is how we get a difference in intention out of the same motivation and the same causal beliefs.[13] The normative problem is why Terror Bomber sets out to kill the children. Both Terror Bomber and Strategic Bomber believe that killing the children will weaken enemy. Both Terror Bomber and Strategic Bomber believe that destroying munitions will weaken enemy. Their instrumental beliefs are the same, then. And their motivation is the

---

[11] Here my talk of causal beliefs does not presuppose causalism about action-explanation: I say that the beliefs are 'causal' to refer to their being beliefs about the causal structures of the world, such as the causal effectiveness of different strategies. Elsewhere I have criticized causalism in action theory (Di Nucci 2008, Di Nucci 2011a, and Di Nucci 2011b), but my argument here is supposed to be independent from the truth or falsity of causalism.

[12] Still, some of these irrational combinations may still play a role in the intuition that our moral judgement on Terror Bomber should be different from our moral judgement on Strategic Bomber. Take the following:

Terror Bomber does not believe that killing children will weaken enemy and she does believe that destroying munitions will weaken enemy. Strategic Bomber believes that destroying munitions will weaken enemy and she does not believe that killing children will weaken enemy.

This is a permutation in which Terror Bomber and Strategic Bomber have the same causal beliefs, but I have excluded it because it involves Terror Bomber in criticisable irrationality: why does she embark on the plan to kill the children in order to weaken enemy if she does not believe that killing children will weaken enemy? Still, maybe this possible combination of the two agents' beliefs may be at least a part of the intuition that Terror Bomber is morally criticisable while Strategic Bomber is not morally criticisable. But this would be seemingly unfair: the two, in such a case, have the same beliefs and cause the same amount of suffering. Can we possibly blame Terror Bomber more just because of her error of judgement? It seems not, because it was not an error of *moral* judgement (if it were, then Strategic Bomber would have committed the same error).

[13] This is, indeed, the core of Bratman's non-reductive planning theory of intention; and here I am not offering a general critique of Bratman's theory, which I have discussed at length elsewhere (Di Nucci 2008, Di Nucci 2009, and Di Nucci 2010).

same too: they both want to promote the war effort by weakening enemy. That is, they have the same motivating reasons or, if you will, pro attitudes. And the same beliefs too: they both believe that 'killing children' will satisfy their pro attitude towards 'weakening enemy' and they both believe that 'destroying munitions' will satisfy their pro attitude towards 'weakening enemy'. They also both know that they cannot destroy munitions without killing children (and that they cannot kill children without destroying munitions). Where does the difference in intention come from?

What we have, here, is a kind of Buridan case: both 'killing children' and 'destroying munitions' satisfy the agent's pro attitude, and the agent does not seem to have distinctive reasons to do one over the other. Still, the agent has overwhelming reasons to do one, and therefore we may suppose that she just picks one because of her overwhelming reasons to do one of the two things. But here we may think that from the motivating perspective this may be like a Buridan case, but from the normative perspective it is outrageous to talk about *picking* between 'killing children' and 'destroying munitions'. There are strong normative reasons to *choose* 'destroying munitions' over 'killing children'. And since there are no instrumental reasons to choose 'killing children' over 'destroying munitions' or to not choose 'destroying munitions' over 'killing children', then the agent ought to choose 'destroying munitions' over 'killing children'. And so we have already come to the normative problem: starting from a cognitive identity, we get a duty to choose 'destroying munitions' over 'killing children'. And Terror Bomber violates this duty to choose 'destroying munitions' over 'killing children'. But then, and this is the crucial point here, it is not DDE, but Terror Bomber's violation of her duty to choose 'destroying munitions' over 'killing children' – duty which Strategic Bomber has not violated – which explains the moral difference between Terror Bomber and Strategic Bomber.

The following plausible moral principle may be what is implicitly doing the work here: if you believe that both A and B satisfy your legitimate goal C, and you believe that A involves the death of no one while you believe that B involves the death of many children, then other things being equal you have a duty not to choose or do B. It is this very plausible moral principle, and not DDE, that may justify the distinction between Terror Bomber and Strategic Bomber if the two have the same causal beliefs.

Here it may be objected that this principle does not apply because both agents choose or do both A and B: but whether or not one wants to talk about 'choosings' or 'doings' in cases of merely foreseen side-effects (see next paragraph), the point stands: given that there is an obvious moral difference between A and B such that B is morally much worse than A, why does Terror Bomber settle on B instead of A when she believes that A would be just as effective in satisfying her goals? She may be ignorant of the obvious moral difference between A and B but then, given that Terror Bomber knows all too well what A and B are, her ignorance about their relative moral value would be itself a serious moral shortcoming on the part of Terror Bomber – and that moral shortcoming would be able to distinguish, morally, between what Terror Bomber does and what Strategic Bomber does. On the other hand, Terror Bomber may not be ignorant of the moral difference between A and B but just indifferent to it – but that's as serious a moral shortcoming as the previous one.

Here it could still be objected that my critique depends on being able to say that Terror Bomber 'settles' on B or 'chooses' B or 'does' B but does not do A; and that, in turn, we need DDE to be able to distinguish between Terror Bomber's attitude towards A and B. But that's just not true: DDE contains a distinction between intended means and merely foreseen side-effects which could be applied to distinguish between Terror Bomber's attitudes towards A and B. But, crucially, that distinction need not exhaust the difference between Terror Bomber's attitude to A and her attitude to B; and, more importantly, DDE claims that it is the distinction between intended means and merely foreseen side-effects which is, itself, morally relevant; while here we have shown that the moral work is being done by other

considerations. Notice, also, the advantage of my solution over the solution offered by DDE: DDE requires an is-ought gap in that it claims that a theoretical distinction in the psychology of the agent makes a moral difference; while my solution only appeals to normative distinctions, which are in themselves *basic* – as the simple moral principle I put forward.

Alternatively, it may be objected that we should not understand this interpretation of the thought-experiment as a Buridan case because the two agents may have different motivations despite having the same causal beliefs. The two agents may indeed be taken to have different moral motives in that they may be following different moral principles: but then, as in the argument already offered, it is the difference in the moral principles they are following and not the Doctrine of Double Effect that is doing the normative work: namely, nothing would depend on the difference between intended means and merely foreseen side-effects.

We have just shown that if we understand the thought-experiment in terms of same causal beliefs, then we can show why this thought-experiment does not support DDE – and this without even beginning to get into the usual arguments on DDE that dominate the literature. This, it may be argued, is a reason to think that we should not understand the thought-experiment in terms of Terror Bomber and Strategic Bomber having the same causal beliefs – even though such an understanding is compatible with the standard versions of the thought-experiment (as those by Bennett and Bratman that we have been following here): in the next section I discuss the alternative interpretation of the agents' psychologies according to which the two agents have different causal beliefs.

## 4. Different Causal Beliefs

Let us then look at the interpretations on which Terror Bomber and Strategic Bomber do not have the same causal beliefs. There are three permutations which do not involve either of the two agents in criticisable irrationality where the two agents do not have the same causal beliefs:

> A) Terror Bomber believes that killing children will weaken enemy and she believes that destroying munitions will weaken enemy.
>
> Strategic Bomber believes that destroying munitions will weaken enemy and she does not believe that killing children will weaken enemy.
>
> B) Terror Bomber believes that killing children will weaken enemy and she does not believe that destroying munitions will weaken enemy.
>
> Strategic Bomber believes that destroying munitions will weaken enemy and she believes that killing children will weaken enemy.
>
> C) Terror Bomber believes that killing children will weaken enemy and she does not believe that destroying munitions will weaken enemy.
>
> Strategic Bomber believes that destroying munitions will weaken enemy and she does not believe that killing children will weaken enemy.

Readings (A) and (B) share a problem with the interpretation on which Terror Bomber and Strategic Bomber have the same causal beliefs: namely, on (A) it is not clear why Terror Bomber chooses 'killing children' over 'destroying munitions' and on (B) it is not clear why Strategic Bomber chooses 'destroying munitions' over 'killing children'. The problem with (A) we have already discussed. The problem with (B) is symmetric, and may have a symmetric effect on morally preferring Strategic Bomber over Terror Bomber. Namely, we may morally prefer Strategic Bomber because, in the absence of instrumental reasons to choose between 'killing children' and 'destroying munitions', we assume that she must have had some moral reasons to prefer the morally superior alternative, namely 'destroying munitions'. But this

need not be the case: maybe, in the spirit of Buridan, Strategic Bomber flipped a coin; and then it would be difficult to morally prefer Strategic Bomber over Terror Bomber, after such a show of indifference towards the moral difference between 'destroying munitions' and 'killing children'.

Let us then leave (A) and (B) aside and focus on (C), which has clear advantages over the interpretation on which Terror Bomber and Strategic Bomber have the same causal beliefs. (C) explains, namely, why Terror Bomber sets out to kill children and not to destroy munitions. And (C) explains, also, why Strategic Bomber sets out to destroy munitions and not to kill children. Terror Bomber opts for the plan of killing children over the plan of destroying munitions because she believes that killing children will weaken enemy and she does not believe that destroying munitions will weaken enemy. And Strategic Bomber opts for the plan of destroying munitions over the plan of killing children because she believes that destroying munitions will weaken enemy and she does not believe that killing children will weaken enemy. And this leaves open the crucial possibility that, had Strategic Bomber had the same beliefs as Terror Bomber, she would have also chosen as Terror Bomber (and *vice versa*). This counterfactual is importantly different from the counterfactual – mentioned also by Bratman – about what Strategic Bomber would have done had she been presented with the same options as Terror Bomber. That counterfactual was about non-psychological options; this counterfactual is about the beliefs of Terror Bomber and Strategic Bomber, not the strategic options offered by their worlds. Still, both counterfactuals generate similar problems for DDE.

Reading (C) leaves open both the possibility that Terror Bomber, had she had Strategic Bomber's beliefs, would have acted as Strategic Bomber did; and the possibility that Strategic Bomber, had she had Terror Bomber's beliefs, would have acted as Terror Bomber did. And one may think that this is going to be a problem for those who want to offer different moral judgements for what Terror Bomber and Strategic Bomber did. On the other hand, it may be objected, what is at issue are moral judgements over actions (for example, the permissibility of killing the children in the case of Strategic Bomber) and not moral judgements over agents, and suggest that therefore not being able to distinguish, morally, between the two agents does not imply that we will not be able to distinguish, morally, between the two actions.

The symmetrically opposite position is often put forward as a softer version or last resort of DDE: namely, that in the impossibility of distinguishing, morally, between the two actions, we may at least distinguish, morally, between the two agents – for example talk about differences in character between the two agents; or talk about "the way the agent went about deciding what to do" (Scanlon 2008: 36). Without discussing the merits of this position, it illustrates the difficulties of its symmetrical opposite: if we can't even find moral differences in the agents, where are the moral differences in the actions going to come from, given that what actually happens in the world is identical in both cases? So interpreting the thought-experiment as supposing that Terror Bomber and Strategic Bomber have different causal beliefs is problematic because then we can't even distinguish, morally, between Terror Bomber and Strategic Bomber as we do not have any reason to think that Strategic Bomber would have acted differently from Terror Bomber had she had her beliefs. There is another problem with tracing back the moral difference to a difference of belief, which I shall just mention here briefly: it exposes the normative judgement to too much luck, and agents should be judged for their actions and inclinations, and not for their causal beliefs.

Let us take stock: we have here analysed another way in which the thought-experiment is underdetermined, namely the beliefs of the two agents. We have shown that there are important differences between interpreting the two agents as having the same causal beliefs and interpreting the two agents as having different causal beliefs. In both cases, though for different reasons, the thought-experiment is shown not to support DDE: in the former case because there is a much more basic moral principle which explains the moral difference; in

the latter case because there is no moral difference – which was also the problem with Bratman's allowing for the two pilots being confronted with different options.

## 5. Structural Similarity with the Trolley Problem

There is another, important, variable. Before discussing it, it helps to introduce the other classic thought-experiment in the DDE literature, the Trolley Problem (Foot 1967; Thomson 1976, 1985, and 2008). In one of the infamous thought-experiments of analytic philosophy, a runaway trolley is about to kill five workmen who cannot move off the tracks quickly enough; their only chance is for a bystander to flip a switch to divert the trolley onto a side-track, where one workman would be killed. In a parallel scenario, the bystander's only chance to save the five is to push a fat man off a bridge onto the tracks: that will stop the trolley but the fat man will die. This is how Thomson introduces the two cases, *Bystander at the Switch* and *Fat Man*:

> In that case you have been strolling by the trolley track, and you can see the situation at a glance: The driver saw the five on the track ahead, he stamped on the brakes, the brakes failed, so he fainted. What to do? Well, here is the switch, which you can throw, thereby turning the trolley yourself. Of course you will kill one if you do. But I should think you may turn it all the same (1985: 1397).

> Consider a case - which I shall call Fat Man - in which you are stand-ing on a footbridge over the trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exits from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He is leaning over the railing, watching the trolley; all you have to do is to give him a little shove, and over the railing he will go, onto the track in the path of the trolley (1985: 1409).

Briefly, DDE is often used to argue that in Bystander at the Switch it is morally permissible to intervene because the killing of the one workman is just a side-effect of saving the five while in Fat Man it is not morally permissible to intervene because the killing of the Fat Man is a means to saving the five. Roughly, then, Bystander at the Switch should be paired with Strategic Bomber and Fat Man should be paired with Terror Bomber. There are some obvious differences between the Trolley thought-experiment and the Terror-Strategic thought-experiment: in the Trolley Problem, there are definite non-psychological differences between the two scenarios. In Fat Man there is a bridge, in Bystander at the Switch there is no bridge, for example. Secondly, in the Trolley Problem there is no talk of intentions, we rather talk of 'means' and 'side-effects'. This suggests that, borrowing respectively from the other thought-experiment, we could analyse the Trolley Problem and the Terror-Strategic thought-experiment as follows: we can say that Terror Bomber kills the children as a means to weakening enemy, while Strategic Bomber's killing of the children is just a side-effect of weakening enemy. Similarly, we can say that, in Bystander at the switch, the bystander does not intend to kill the one workman; and we will say that on the other hand in Fat Man the bystander does intend to kill the fat guy.

We have introduced the Trolley Problem because the thought-experiment of Strategic Bomber and Terror Bomber is underdetermined also with respect to its structural similarity with the Trolley Problem. Suppose that the munitions are kept under the school's ground[14]; that is, supposedly, why we cannot destroy the munitions without killing the children. That

[14] Delaney (2008) proposes a similar scenario. I criticize Delaney in Di Nucci (submitted, a).

the munitions be geographically located under the school is compatible with the way in which the thought-experiment is normally told (see Bennett and Bratman above, for example) and it presents a structural similarity with Fat Man as opposed to Bystander at the Switch, as the children are now physically between the bombs and the munitions just as the poor fat guy will find himself physically between the trolley and the five workmen. The bombs will hit the school and then, and only then, hit the munitions; the same way in which the trolley will hit the fat guy and then, and only then, stop; while in Bystander at the Switch we may say that the five are saved before the trolley kills the one, as it is enough that the trolley is deviated on the side-track.

Now we know where the munitions and the school are located, but nothing is supposed to hinge on this. We will still say that Terror Bomber's plan is to kill the children in order to weaken the enemy, and that she knows that in killing the children she will also destroy the munitions. Similarly, we will say that Strategic Bomber's plan is to destroy the munitions in order to weaken the enemy, and that she knows that she will also kill children. The proposed analysis is that Terror Bomber intends to kill children and merely foresees that she will destroy munitions; and that Strategic Bomber intends to destroy munitions and merely foresees that she will kill children.[15]

We can see that the above structure is supposed to make no difference to Bratman's analysis of Terror Bomber's intention to kill the children, which Strategic Bomber lacks. The three roles of intention individuated by Bratman (1987: 140-143) are: (i) 'posing problems for further reasoning', (ii) 'constraining other intentions', and (iii) 'issuing in corresponding endeavouring'. As these roles are applied to Terror Bomber's intention, Bratman says that Terror Bomber's intention will (i) pose the problem of how he is going to kill the children: "Terror Bomber must figure out, for example, what time of day to attack and what sorts of bombs to use" (1987: 141). (ii) Terror Bomber's intention will also be incompatible with other possible strategies. Terror Bomber may not, for example, implement a plan to deploy some troops if this deployment would result in the enemy evacuating the children: "So Terror Bomber's prior intention to kill the children stands in the way of his forming a new intention to order the troop movement (1987: 141). (iii) Terror Bomber will also guide his conduct so as to cause the death of the children: "If in midair he learns they have moved to a different school, he will try to keep track of them and take his bombs there" (1987: 141-142).

Bratman claims that these three roles are not true of Strategic Bomber's attitude towards killing the children: Strategic Bomber will not engage in practical reasoning about how to kill the children; if further intentions of Strategic Bomber should be incompatible with killing the children, that will not be a prima facie reason to disregard them; and, to put Bratman's point crudely, if the children move, Strategic Bomber will not follow them. These three claims are independent of the three underdetermined elements that we have so far identified: (a) whether or not there is a munitions factory in the world of Terror Bomber, his attitude towards killing the children will have these three roles and Strategic Bomber's attitude will not have these three roles; (b) whether Terror Bomber and Strategic Bomber have the same causal beliefs (for example about the efficacy of killing children to weaken the enemy) will not

[15] Let me here note that even though I have imported the structure of the trolley problem, the two thought-experiments remain different in that in the trolley problem there are obvious non-psychological differences (the bridge, for example) which need not be the case in the terror-strategic thought-experiment. Also, it may be argued that there is a further difference in that the agent in Fat Man physically uses the fat guy for her purposes, while the agent in Strategic Bomber does not physically use the children for her purposes – the difference being, supposedly, that the agent in Fat Man physically pushes the fat guy while the agent in Strategic Bomber does not have any such contact with the children. Here I would be worried that we would then be just talking, as in Harris's irony, about the difference between throwing people at trolleys and throwing trolleys at people (or throwing bombs at people and throwing people at bombs). But for those who take this challenge more seriously, see my critique of Quinn (1989) in Di Nucci (submitted, a).

alter the three roles of Terror Bomber's attitude towards killing the children. And the same goes (c) for the structure of the scenario, so that even if the munitions are hidden under the school, then it will still be the case that Strategic Bomber will have to engage in practical reasoning which has to do with, say, the sorts of bombs that will penetrate deep enough in the ground while that element will not play a role in Terror Bomber's reasoning.

The problem for DDE is that, apparently, borrowing the structure of Fat Man from the Trolley Problem does not make any difference to the attribution of the relevant intention to Strategic Bomber. But then we have two structurally similar scenarios, Fat Man and Strategic Bomber, to which DDE gives different answers, as it says that it is not morally permissible to kill the fat guy in Fat Man while it says that it is morally permissible to kill the children in Strategic Bomber: and this latter claim seems in turn less plausible if the munitions are hidden under the children – think of the case of human shields.

## 6. The Three Roles of Intention

What happens if we apply Bratman's analysis of the three roles of intention to the Trolley Problem? As we said, the comparison between the Trolley Problem and the Terror-Strategic thought-experiment is complicated by the use of different terminologies in discussing the two cases: for the Trolley Problem the talk is of side-effects as opposed to means, for the Terror-Strategic thought-experiment the talk is of intended as opposed to merely foreseen. But if both thought-experiments are to be explained by DDE, then there must be an available common reading.[16] There are, in fact, two common readings: we can either talk in both cases of side-effects and means, or we can talk in both cases of intended and merely foreseen. The outcome is that we would say, of Bystander at the Switch, that the bystander does not intend to kill the one workman and that the killing of the one workman is just a side-effect of the bystander's rescue of the five. Of Fat Man, we would on the other hand say that the bystander does intend to kill the fat guy and that the bystander's killing of the fat guy is a means to the bystander's end of saving the five. Of Terror Bomber, we will say that she intends to kill the children and that killing the children is a means to Terror Bomber's end of weakening the enemy. Finally, we will say of Strategic Bomber that she merely foresees the killing of the children without intending it, and that killing the children is, for Strategic Bomber, merely a side-effect of her destruction of the munitions factory.

With this common understanding in place, we can test Bratman's three roles of intention on the attribution of the relevant intentions to the Trolley scenarios. Let us for example take the bystander's intention, in Fat Man, to kill the fat guy. This can be compared to the bystander's intention to stop the Trolley. Defenders of DDE have, traditionally, difficulties in explaining why in these cases we may not just say that the agent only intended to stop the trolley but did not intend to kill the fat guy.[17] We can look in the Fat Man scenario for the three roles identified by Bratman: (i) posing problems for further reasoning; (ii) constraining other

[16] As it is quickly shown that means and side-effects must be understood intensionally and not extensionally (see, for example, Davis 1984 or Roughley 2007), I will not repeat here arguments for the equivalence of the side-effect/means reading with the merely foreseen/intended reading. It is commonly accepted in the literature that means are intended while side-effects are not. For an exception, see Kamm 2000 & 2007.

[17] This is the so-called problem of closeness, already identified by Foot (1967) and which has since played a major role in the debate on DDE. I have discussed closeness elsewhere (Di Nucci submitted, a), so here I shall just mention some representative major contributions to this particular stream of the debate: Foot 1967, Bennett 1980, Quinn 1989, Fischer/Ravizza/Copp 1993, McMahan 1994, McIntyre 2001, and Wedgewood 2011. My discussion of the problem of closeness here is very brief and only focuses on Bratman's three roles of intention because in Di Nucci (submitted, a) I go in much more detail by looking at ten different recent proposals to deal with the problem of closeness in order to rescue DDE: I find each of these ten recently suggested solutions wanting.

intentions; and (iii) issuing in corresponding endeavouring. Does the bystander's attitude towards killing the fat guy have the following three roles? If it does not have these three roles, it is no intention, and then we cannot say, at least on Bratman's understanding of intention, that the bystander intended to kill the fat guy. And this would be particularly damaging for DDE, as Bratman's understanding – as we have seen – is meant to be sympathetic to DDE.

Let us start with the first role, posing problems for further reasoning. I think we can here contrast the supposed intention to kill the fat guy with the intention to stop the trolley, and see that only the latter attitude has the role for further reasoning identified by Bratman, and that therefore only the latter attitude is an intention. The bystander will have to reason about whether the fat man is heavy enough, for example; because if the fat man is not heavy enough to stop the trolley, then it will not make any rational sense to throw him off the bridge. But the bystander will not have to reason about a way of throwing him off the bridge so as to increase the chances that the trolley will hit head on the fat guy's vital organs, so as to guarantee the fat guy's death. The sort of further reasoning that the bystander will have to engage in has to do, then, only with how the fat guy will ensure that the trolley will be stopped; and not with the actual death of the fat guy.[18] Similar points can be made about the other two roles of intention: if the trolley happens to stop, for example, before I have pushed the fat guy, then I will no longer endeavour to throw him off the bridge – just as, in Bratman's discussion, Strategic Bomber will not pursue the children in case they leave the school. But now Fat Man looks like Strategic Bomber and not like Terror Bomber, so that we would say that the bystander in Fat Man intends to stop the trolley but merely foresees the death of the fat guy (or, in the other terminology, that the killing of the fat guy is a side-effect of the bystander's stopping of the trolley, which is in turns a means to his end of saving the five).

We have here shown, then, that in the debate on DDE one cannot just isolate an action-theoretical part of the argument: the problem of closeness has been here applied to Bratman's three roles of intention to show that, even if we understand the concept of intention as suggested by Bratman, in the application of this concept to the relevant thought-experiments we still have the usual difficulties. This also implies that we have offered a critique of Bratman's account of intention: but a general critical engagement with Bratman is not within the scope of this work.

## 7. Historical Closeness

I want to discuss one further aspect of the problem of closeness, which brings us back where we started, namely to the explicit character of the directives and language of the British Bomber Command in WWII. We have seen how, in the course of 1940, the British went from explicitly forbidding indiscriminate bombing ("specifically laid down that targets had to be identified and aimed at. Indiscriminate bombing was forbidden" (1970: 24)) through ordering that bombs be directed at inhabited areas ("Bomber Command was instructed simply to aim at the center of a city... the aiming points are to be the built-up areas, not, for instance, the dockyards or aircraft factories" (1970: 24)) to, finally, referring themselves to their own operations as *terror* ("Here, then, we have *terror and devastation* carried to the core of a warring nation" (as quoted by Ford 1944: 294)). This is interesting because it gives a very clear historical context to the Terror-Strategic thought-experiment that we have been examining throughout. But there is a further element of interest in the directives and

[18] Just to be clear, I do not pretend to be offering original kinds of arguments, as those familiar with the discussion of closeness will recognise at least some of these kinds of arguments. What I want to show is just that the three roles of intention identified by Bratman are not independent of the classic problem of closeness.

language of the British: they never actually explicitly talk about killing civilians. The following Air Ministry directif and Air Staff paper from 1941 are illustrative (cited by Harris):

> to focus attacks on the morale of the enemy civil population, and, in particular, of the industrial workers. (1995: 7)
>
> The ultimate aim of the attack on a town area is to break the morale of the population which occupies it. To ensure this we must achieve two things: first, we must make the town physically uninhabitable and, secondly, we must make the people conscious of constant personal danger. The immediate aim, is therefore, twofold, namely, to produce (i) destruction, and (ii) the fear of death. (1995: 7)

Let me first note that the civil servants of the time appear to have read Bentham, when they talk of the *immediate* aim of the operation.[19] More importantly, they explicitly talk about destruction, fear of death, and of making German cities "physically uninhabitable". It could not be any clearer that the targets are no longer only military or industrial sites; but, interestingly, the British stop short of writing down the obvious, in their orders: namely that the immediate aim is to terrorize civilians ("constant personal danger" and "fear of death") by killing as many of them as possible. That is, the order of killing civilians is nowhere explicitly formulated. And this shows a surprising sensibility of the British authorities to the problem of closeness: we only intend to scare them; we only intend to destroy their houses and workplaces; but we do not intend to actually kill them. As we bomb their houses at night while they sleep in them, we only intend to destroy the houses, not to kill those sleeping inside. Yeah, right! Here I am not making an historical point: firstly, for all I know, it may be that there are documents that talk about the explicit killing of the civilian population; secondly, I am not even claiming that the killing of the civilian population was deliberately left implicit in the formulation of orders.[20] All I am saying is that we can find in those historical documents all the argumentative force of the problem of closeness, as the challenge for DDE is exactly to spell out why an agent who intentionally bombs houses she knows to be full of people cannot be said to merely intend to destroy the houses without also intending to kill the inhabitants.[21]

## 8. Conclusion

Summarizing, we have here identified three different ways in which the terror-strategic thought-experiment is underdetermined, all of which are crucial to its support of DDE: (1) the thought-experiment is underdetermined as to whether the non-psychological worlds of Terror Bomber and Strategic Bomber are identical: but if the two worlds are not identical, then it may be those differences and not DDE which explain the supposed moral difference between Terror Bomber and Strategic Bomber. The thought-experiment is also

[19] Bentham (*An Introduction to the Principles of Morals and Legislation*) famously distinguishes between *oblique* intention and *direct* intention. The talk of immediacy takes us back to Gury and Mangan's definitions of double effect. To see that immediacy, when talking about double effect, is a way of denying that the relevant effect is a means, think of its German translation 'unmittelbar', which already includes in itself the denial of 'means' (mittel).

[20] This is indeed very plausible, but it is an historical hypothesis and this is not the place to defend it.

[21] Another possibly interesting question about DDE with relation to the history of WWII is whether we find it plausible that DDE offers different judgements for what the RAF did in the first part of 1940 and for what they did in the second part of 1940 – even though we must be careful in not overstating the application of DDE to WWII: the technology was such that only 22% of bombers dropped their load within 5 miles of the target (Walzer (1971: 15) and Frankland (1970: 38-39)); with this kind of success rate it is likely that we are not even allowed to talk of intentions in the first place, as we would violate belief constraints: the bombers' attitude towards their targets was at the time more one of hope than one of intention, at least if we accept rational constraints on intention (Bratman 1984 & 1987; McCann 1991, 2010, and 2011; Di Nucci 2009 & 2010).

underdetermined as to (2) whether the psychologies of Terror Bomber and Strategic Bomber are identical: we have shown that, whether or not we interpret Terror Bomber and Strategic Bomber as having the same causal beliefs, DDE has a problem: if the two agents have the same causal beliefs, then why does Terror Bomber choose killing the children over destroying munitions? Terror Bomber's choice is morally problematic in the absence of a difference in causal beliefs; but then it may be Terror Bomber's dubious moral choice, and not DDE, that explains the moral difference between Terror Bomber and Strategic Bomber. And if the two have different causal beliefs, then we can't rule out the counterfactual that, had Strategic Bomber had the same beliefs as Terror Bomber, she would have also acted as Terror Bomber did. But then how are we to morally distinguish between the two? And if we can't distinguish, morally, between the two agents, and the two worlds are identical, then where is the moral difference going to come from? Finally, the thought-experiment is also underdetermined as to (3) its structural similarity with the Fat Man scenario of the Trolley Problem: what if, namely, the munitions were kept under the school? Here DDE would give different answers to two structurally symmetrical cases, Strategic Bomber and Fat Man.

**Ezio Di Nucci**

Universität Duisburg-Essen
ezio.dinucci@uni-due.de

## References

Bennett, J. (1980), Morality and Consequences. The Tanner Lectures On Human Values.

Bennett, J. (1995). *The Act Itself*. Cambridge UP.

Boyle, J.M. (1980), ´Toward Understanding the Principle of Double Effect', Ethics 90: 527-538.

Bratman, M. (1984), 'Two Faces of Intention', Philosophical Review 93: 375-405.

Bratman, M. (1987), Intention, Plans, and Practical Reason. Cambridge, Mass.: Harvard University Press.

Cavanaugh, T.A. (2006). Double-Effect Reasoning. Oxford UP.

Davis, N. (1984). The Doctrine of Double Effect: Problems of Interpretation. Pacific Philosophical Quarterly 65: 107-123.

Delaney, N. (2006). Two cheers for closeness: terror, targeting, and double effect. *Philosophical Studies* 137: 335-367.

Di Nucci, E. (2008), *Mind Out of Action*. VDM Verlag.

Di Nucci, E. (2009), 'Simply, false', *Analysis* 69/1: 69-78.

Di Nucci, E. (2010), 'Rational Constraints and the Simple View', *Analysis* 70/1: 481-486.

Di Nucci, E. (2011a), 'Frankfurt versus Frankfurt: a new anti-causalist dawn', *Philosophical Explorations* 14 (1): 117-131.

Di Nucci, E. (2011b), 'Automatic Actions: Challenging Causalism', *Rationality Markets and Morals* 2 (1): 179-200.

Di Nucci, E. (2012), 'Double Effect and Assisted Dying'. *British Medical Journal* (letter, 7.2.2012).

Di Nucci, E. (forthcoming, a), 'Self-Sacrifice and the Trolley Problem', *Philosophical Psychology*.

Di Nucci, E. (forthcoming, b). 'Embryo Loss and Double Effect'. *Journal of Medical Ethics*.

Di Nucci, E. 'The Doctrine of Double Effect and the Trolley Problem'.

dx.doi.org/10.2139/ssrn.1930832

Di Nucci, E. 'Too Close to Call: Double Effect' (submitted, a).

Di Nucci, E. ´Moral Intuitions on Collateral Damage and Double Effect´ (submitted, b).

Di Nucci, E. *Ethics without Intention* (book manuscript).

Fabre, C. (2009), 'Guns, Food, and Liability to Attack in War', Ethics 120/1: 36-63

Fischer, J.M., Ravizza, M. & Copp, D. (1993), ´Quinn on Double Effect: The Problem of Closeness', Ethics 103: 707-725.

Foot, P. 1967. The problem of abortion and the doctrine of the double effect. Oxford Review 5: 5–15.

Ford, J.C. (1944), 'The Morality of Obliteration Bombing', Theological Studies 5: 261-309.

Frankland, N. (1970), Bomber Offensive. Ballantine Books.

Harris, A.T. (1995), Despatch on War Operations. Frank Cass & Co.

Mangan, J.T. (1949). An Historical Analysis of the Principle of Double Effect. Theological Studies 10: 41-61.

McCann, H. 1991. Settled objectives and rational constraints. American Philosophical Quarterly 28: 25–36.

McCann, H. 2010. Di Nucci on the Simple View. Analysis 70: 53–59.

McCann, H. (2011), 'The Simple View again: a brief rejoinder', Analysis 71/2: 293-295.

McIntyre, A. (2001), 'Doing Away With Double Effect', Ethics 111/2: 219-255.

McMahan, J. (1994), 'Revising the Doctrine of Double Effect', Journal of Applied Philosophy 11: 201-212.

McMahan, J. (2009), Killing in War. Oxford UP.

Quinn, W. (1989), 'Actions, Intentions, and Consequences: The Doctrine of Double Effect', Philosophy and Public Affairs 18/4: 334-51.

Petrinovich, L., and O'Neill, P., (1996). Influence of wording and framing effects on moral intuitions. Ethology and Sociobiology 17: 145-171.

Roughley, N. (2007), ´The Double Failure of Double Effect', in Nannini, S. & Lumer, C. (eds.), Intentionality, Deliberation, and Autonomy. Ashgate.

Ryan, J.K. (1933), Modern War and Basic Ethics. Bruce.

Scanlon, T. (2008), Moral Dimensions. Harvard UP.

Sinnott-Armstrong, W. (2008). Framing Moral Intuitions in W. Sinnott –Armstrong (Ed.) Moral Psychology, Volume 2: The Cognitive Science of Morality, (pp. 47-76). Cambridge, MA: MIT Press.

Swain, S., Alexander, J. and Weinberg, J. (2008). The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp. Philosophy and Phenomenological Research 76: 138-155.

Thomson, J.J. 1976. Killing, letting die, and the trolley problem. The Monist 59: 204–17.

Thomson, J.J. 1985. The trolley problem. The Yale Law Journal 94: 1395–415.

Thomson, J.J. 2008. Turning the trolley. Philosophy and Public Affairs 36: 359–74.

Walzer, M. (1971), 'World War II: Why Was This War Different?', Philosophy & Public Affairs 1/1: 3-21.

Wedgwood, R. (2011). Defending Double Effect. Ratio (forthcoming).

Wiegmann, A., Okan, Y., Nagel, J. (2010). Order Effects in Moral judgment. Philosophical Psychology (forthcoming).

Woodward, P.A. (ed.) (2001). The Doctrine of Double Effect. University of Notre Dame Press

# Counterfactuals and Two Kinds of *Ought*

Daniel Dohrn

I discuss Caspar Hare's solution to a new variant of Parfit's Non-Identity Problem. Hare's solution rests on distinguishing two kinds of ought: *The Ought of Omniscient Desire:* what you ought$^{OD}$ to do is what an omniscient, rational creature with appropriate interests would want you to do. *The Ought of Most Reason:* what you ought$^{MR}$ to do is what there is most reason to do. I argue that the distinction does not dissolve the problem. Moreover, I show that Hare's proposal to spell out his distinction in terms of an embedded counterfactual (if you had not done what you did, then, if you had done what you did, what would the consequence have been?) is flawed.

Derek Parfit's notorious Non-Identity Problem still stirs debate. Recently, David Velleman (2006) and Caspar Hare have given the problem an interesting new twist. I use an example of Hare's to illustrate the original problem (cf. Hare 2007: 499):

> Mary knowingly causes her child to have heart conditions. For she ignores her doctor's advice and conceives a child while recovering from German Measles. She gives birth to a child, John, who suffers from a life-long heart condition but leads a life otherwise worth living.

Intuitively, Mary did something wrong. Had she waited, her child would in all probability have been healthy. Yet she would not have conceived *John* but a person with a different genome. John prefers living with a heart condition to not existing at all. So how could Mary have done something wrong? What she did has benefited John.

## 1. The New Non-Identity Problem

Here is the new twist, assuming you are Mary:

> A related, but distinct puzzle, which has received less attention, is to make sense of your attitude, years later, towards John's birth. Years later you have become very attached to John. You are not faking it, when you celebrate his birthday. You are glad that you ignored your doctor's advice and brought John into the world. And so you *should* be. Good parents have this attitude towards their children. And yet you recognize that you ought not to have conceived John. How can this be? How can you simultaneously be glad you did something, and recognize that you ought to be glad that you did it, and recognize that you ought not to have done it? (Hare 2011: 202)

It is important to see the normative claim: it is not just morally permissible for Mary to be glad that she conceived John, say because her feeling has no negative moral consequences. Mary *ought* to be glad that she did conceive John. I prefer to state the point a bit differently: from our independent viewpoint, Mary ought not to have conceived John. However, Mary should adopt certain feelings towards John once he is born. In light of these feelings, *she* should judge that she ought to have conceived John. The problem is how to reconcile this intuition with the intuition that Mary ought not to have conceived John.

In my interpretation, Hare's challenge is the following. From our independent viewpoint, after John has been born, we endorse the following valuations:

(i) Mary ought not to have conceived John.

(ii) Upon giving birth to John, it is appropriate for Mary to adopt John-centred values. Guided by these values, Mary should accept that she ought to have conceived John.

Our verdicts (i) and (ii) are time-invariant. *We* accept (i) at any time. And we accept (ii) at any time (anticipating John), although *Mary* should only adopt John-centred values *after* John has been born. So her judgement should shift: before John has been born, she should judge that she ought not to conceive a child earlier, after John has been born, she should judge that she ought to have conceived him.

Hare endorses an even stronger claim:

(iii) Even after his birth, Mary herself should accept that she ought not to have conceived John.

I am not so sure about this claim. Thus, I will concentrate on (i) and (ii). But in case (iii) seems convincing, I offer the following explanation: there are two different readings for 'I ought to have conceived John' as judged from Mary's viewpoint. There is a default reading: her evaluative outlook guiding the present judgement is the present one. At $t_2$, after she has given birth to John, she should judge that she ought to have conceived John. But there is a non-default reading of 'Mary ought to have conceived John': in that reading, the point in time at which she accepts 'I (Mary) ought not to have conceived John' and the point *for which* she evaluates this judgement come apart: she so to speak takes the vantage point of her former self. Judging from a point of evaluation and everyone's (including her own) values *at* $t_1$, before she has given birth to John, she ought not to have conceived John. For simplicity, I mostly disregard this refinement of the case.

## 2. Hare's Distinction of Two Kinds of *Ought*

I now turn to Hare's solution to the new problem. Hare distinguishes a subjective and an objective sense of *ought*:

> what you ought to do in the objective sense has to do with the merits and de-merits of the options available to you, while what you ought to do in the subjective sense has to do with the merits and de-merits of the options available to you, *from your epistemic position* ...
>
> *The Ought of Omniscient Desire:* What you ought$^{OD}$ to do is what an omniscient, rational creature with appropriate interests would want you to do.
>
> *The Ought of Most Reason:* What you ought$^{MR}$ to do is what there is most reason to do. (Hare 2011: 190, first emphasis mine)

Ought$^{OD}$ is determined by all facts, past and future ones. Ought$^{MR}$ is determined by facts which are available to the agent when she makes a decision. The difference between both kinds of *ought* is presented as an epistemic one. Ought$^{MR}$ is confined to reasons within your purview as a finite cognizer, ought$^{OD}$ comprises reasons within the purview of an omniscient ideal cognizer. Hare goes on to flesh out the distinction in terms of counterfactuals. I think that the counterfactual part of his reasoning is flawed. I will come to it in a moment Nevertheless, at first glance the idea that the epistemic distinction bears on the Non-Identity Problem is appealing: to Hare, Mary ought$^{OD}$ to have conceived John, yet she ought$^{MR}$ not to have conceived John. She could not know that she would conceive John. So she could not feel

attached to *John* before John came into existence. But from an epistemic position which includes her attachment to John, she ought to have conceived John.

However, I do not think the relevant distinction is an epistemic one, at least epistemic in the sense intended by Hare. It is not Mary's *learning* something about John that brings about her John-centred feelings. It is no peculiar intrinsic property of John as opposed to some other possible child of Mary's which makes him the appropriate target of her emotions. Moreover, Mary could have perfectly anticipated that whoever would be her child would be the subject of her feelings. So it is neither her ignorance of John's specific properties nor her ignorance of the nature of her attachment that explains why she ought$^{MR}$ not to have conceived John.

I use Hare's own idea of an omniscient observer to shed further doubt on the idea that the problem is an epistemic one. The omniscient observer knows everything about the consequences of Mary's actual actions. Assume Mary (like Frank Jackson's Mary) is an extremely efficient learner. Before she conceives John, the omniscient observer has told her everything about her future child John, including Mary's special affection for John and so on. Say she is provided with a complete description of the future situation. If we frame the distinction of two kinds of ought as an epistemic one, Mary's position at t1 (even before she has conceived John) should now be characterized by ought$^{OD}$. Still I do not think that this enforces a change in Mary's obligations or values. Judging from her own vantage point before she conceives John, she still ought to have waited, although she is in the very same *epistemic* position she will be after giving birth to John. In sum, Hare's epistemically based distinction of two kinds of *ought* does not dissolve the puzzle.

Before proceeding with my criticism of Hare, I briefly sketch my own preferred solution: the change in Mary's valuation is not brought about by her learning something but by becoming *emotionally* acquainted with John (cf. Velleman 2006: 269). Acquaintance is not understood in informational terms but in the sense of intimacy between mother and child. There is nothing Mary could not have known in advance about her attachment to John, except, perhaps, what it is like to feel acquaintanted with John. We cannot feel the same for John. Thus we retain our judgement. But we acknowledge that it is appropriate for Mary to adopt her new feelings upon becoming acquainted with John. In contrast to Velleman, I do not claim Mary to judge at $t_2$ that she should not have conceived John (except in the non-default reading exposed above). Thus, I do not have to accept inconsistent valuations in claiming Mary to judge at $t_2$ that she should not have conceived John.

## 3. Criticism of Hare's Counterfactual Account

I criticize Hare's counterfactual account of the different senses of 'ought' he distinguishes.

Assume you may spin a wheel of fortune. If it were spun with a force of 15.88348N, it would show red, if it were spun with 15.88349N, it would show black. You do not spin the wheel. Arguably it is indeterminate what would have happened if you had spun the wheel. Hare contends that in general, such indeterminacy is *resilient under embedding*. Now consider a varied case: let S stand for 'the wheel is spun', R for 'it shows Red'. S $\Box\!\!\rightarrow$ R is by assumption indeterminate. Assume S and R obtain. Moreover, outcome R actually leads to a loss but the expected value of spinning was positive. As a test what one ought$^{OD}$ to have done (whether one ought to have spun the wheel), Hare imagines an omniscient being who knows that S actually leads to the unfortunate outcome R. From the perspective of this being, one ought$^{OD}$ not to have spun the wheel. In order to figure out what one ought$^{MR}$ to have done, Hare suggests, one should ask what would have happened if one had not spun the wheel by an *embedded counterfactual*. If one had not spun, then if one had spun, would the outcome have been positive? Since S $\Box\!\!\rightarrow$ R is indeterminate, Hare argues, so is $\neg$S $\Box\!\!\rightarrow$(S $\Box\!\!\rightarrow$ R). As a consequence, to figure out what one ought$^{MR}$ to have done, one must resort to a weaker

counterfactual that represents the expected value of the spinning: if one had spun, the expected value attained would have been such and such. Since the expected value is positive, the same goes for the embedded counterfactual. One should have spun. Ought$^{OD}$ and ought$^{MR}$ come apart.

This reasoning is flawed: for a counterfactual to be indeterminate, it must be genuinely contrary-to-fact: S does not obtain. For in the special case where S and R actually obtain, S $\Box\!\!\rightarrow$ R is ipso facto true (by the centring axiom which forms part of the standard account used by Hare) and not indeterminate. Yet Hare's case is the special case. S and R *do* obtain. Hare might argue that centring does not matter here: the embedded counterfactual is true iff, judged from the closest $\neg$S–world, S $\Box\!\!\rightarrow$ R is true. Yet what precludes that the actual (S-) world outdoes any other S-world in closeness to the closest $\neg$S-world? Moreover, Hare's solution depends on rejecting a principle which many find attractive: import-export: (P $\Box\!\!\rightarrow$(Q $\Box\!\!\rightarrow$ R)) $\leftrightarrow$ (( P & Q) $\Box\!\!\rightarrow$ R). Applied to the spinning wheel-example, $\neg$S $\Box\!\!\rightarrow$(S $\Box\!\!\rightarrow$ R) leads to the vacuous case ($\neg$S & S) $\Box\!\!\rightarrow$ R. The result is that the embedded counterfactual is true and not indeterminate.

Anyway Hare's criterion is plausible only as far as it tracks the epistemic criterion (what is in the purview of the agent when she deliberates her action?) Disregard for the sake of argument all my qualms about the semantics of counterfactuals. The counterfactual criterion tracks the epistemic criterion as far as it rules out certain contingent consequences of the action under consideration. Those consequences are known to an omniscient observer but not to the agent deliberating her action. If all facts ruled out by counterfactually undoing the action are of this sort, the counterfactual criterion tracks a necessary condition of ought$^{MR}$.

Nevertheless the point of discerning ought$^{MR}$ and ought$^{OD}$ is missed when the real epistemic issue is replaced by the counterfactual criterion. I use the Non-Identity problem to show this. Hare's solution to the new variant of the Non-Identity problem boils down to the following: while Mary ought$^{OD}$ to have conceived a child earlier, Mary ought$^{MR}$ not. Given Mary's John-centred values, from the perspective of the omniscient observer who knows that Mary will conceive John, there is a reason for her to conceive a child earlier. For that child will be John, and John will be valued by Mary. But Mary ought$^{MR}$ to have conceived a child earlier only if the following is true:

> if Mary had not acted such as to conceive a child earlier, then if Mary had acted such as to conceive a child earlier, Mary would have conceived John.

And since the embedded counterfactual is indeterminate (there are many different possible children which could have been the result of Mary's action under the supposition that the action is undone, or so I grant), the embedding counterfactual is not true but indeterminate. It is not the case that Mary would have conceived John. Thus Mary must resort to the *expected quality of life* of a child conceived earlier. Since that quality is lower than the quality of life of a child conceived later, Mary ought$^{MR}$ not to have conceived a child earlier.

There are two counterarguments. Firstly, I have imagined a situation where the omniscient being at $t_1$ tells Mary everything about future John. Ought$^{MR}$ has been introduced by reasons within Mary's purview. Now the reasons within Mary's purview are the same that are within the omniscient being's purview. From t1 onwards, ought$^{MR}$ and ought$^{OD}$ fall together. But we (including Mary) intuitively judge at t1 that Mary ought not to conceive a child earlier. Perfect *knowledge* of John does not have to change one's valuations. This shows that Hare's epistemically based distinction does not apply. But it also shows how the epistemic criterion and the counterfactual criterion can come apart. If Hare is right about how the counterfactual criterion works, the latter does not yield the same solution to Non-Identity as the epistemic criterion: judging from the counterfactual criterion, Mary ought not to have conceived a child

earlier. For although Mary knows that she *actually* will conceive John, it is not determinately the case that she would conceive John under the counterfactual supposition.

To develop a second argument, I present a different variation of Hare's Non-Identity scenario: just as in Hare's original scenario, Mary does not know what the omniscient being knows. At t1, she does not yet know that she will conceive John. Assume the omniscient and very mighty being *unbeknownst to Mary* has fixed a condition. There is a law-like connection between Mary's conceiving a child earlier and her conceiving *John*. Under these circumstances, arguably

> if Mary had not acted such as to conceive a child earlier, if Mary had acted such as to conceive a child earlier, Mary would have conceived John.

The counterfactual criterion yields that Mary ought$^{MR}$ to have conceived a child earlier. But by Hare's own lights, judging from reasons within Mary's purview, she still ought not to have done so: Mary could not anticipate that she would conceive John. But Mary could anticipate that conceiving a child later would lead to the child having a better life. Admittedly my second argument somewhat stretches the boundaries of sound thought experiment.[1] But so does Hare's original idea of counterfactually undoing an action.

**Daniel Dohrn**

Humboldt-Universität zu Berlin
dohrndan@hu-berlin.de

## References

Hare, C. 2007: 'Voices from Another World: Must we Respect the Interests of People Who Do Not, and Will Never, Exist', *Ethics* 117, 3, 498–523.

— 2011: 'Obligation and Regret When there is No Fact of the Matter About What Would Have Happened if You Had not Done What You Did', *Noûs* 45, 190-206.

Parfit, D. 1989 (originally 1984): *Reasons and Persons*. Oxford: Oxford University Press.

Velleman, D. 2008: 'Love and Non-Existence', *Philosophy and Public Affairs* 36, 266-288.

[1] And there are some uncertainties how law-like conditions of the sort considered behave under counterfactual suppositions.

# Thomas Buddenbrook und der Vorrang der Moral

Martin Hoffmann

Handelnde sollen in ihrem Handeln moralischen Gründen gegenüber anderen praktischen Gründen den Vorrang geben. – Diese Auffassung ist in der Ethik weithin akzeptiert. In der philosophischen Tradition ist Kant einer ihrer herausragenden Vertreter; in der modernen Metaethik hat sie Richard Hare an prominenter Stelle verteidigt. In jüngerer Vergangenheit ist die allgemeine Geltung der These vom Vorrang der Moral aber aus verschiedenen Gründen bezweifelt worden. Mein Ziel besteht darin, diese Kritik zu entkräften und eine plausible Version dieser These zu explizieren. Zunächst wird der Status der Vorrangthese diskutiert. Es wird dafür argumentiert, dass die Vorrangthese eine Behauptung ist, die in unserem alltäglichen Sprechen und Urteilen über moralische Gründe tief verankert ist. Vor diesem Hintergrund wird dann die Frage gestellt, ob das in der Vorrangthese zum Ausdruck kommende Gebot ein allgemeines *Vernunft*gebot oder ein *moralisches* Gebot ist. Ich werde am Beispiel eines in Thomas Manns *Buddenbrooks* gestalteten Handlungsszenarios zeigen, dass sich die Vorrangthese sinnvoll als moralische Forderung interpretieren lässt. Abschließend werden zwei Fallbeispiele diskutiert, die die Kritiker der Vorrangthese vorbringen, um ihre Unplausibilität zu belegen. Es wird dafür argumentiert, dass diese Fallbeispiele entgegen dem ersten Anschein keine schlagkräftigen Belege gegen die Geltung des Vorrangs der Moral darstellen.

## 1. Einleitung

Die Auffassung, dass Handelnde in ihrem Handeln moralischen Gründen gegenüber anderen praktischen Gründen den Vorrang geben sollen – im Folgenden: die Vorrangthese –, ist in der Ethik weithin akzeptiert. In der philosophischen Tradition ist Kant einer ihrer herausragenden Vertreter. So schreibt er in der *Grundlegung zur Metaphysik der Sitten*:

> Der hypothetische Imperativ sagt also nur, daß die Handlung zu irgendeiner *möglichen* oder *wirklichen* Absicht gut sei. Im erstern Falle ist er ein ***problematisch***-, im zweiten ***assertorisch***-praktisches Prinzip. Der kategorische Imperativ, der die Handlung ohne Beziehung auf irgend eine Absicht, d. i. auch ohne irgend einen anderen Zweck, für sich als objektiv nothwendig erklärt, gilt als ein ***apodiktisch***-praktisches Prinzip. (Kant 1785: 414–415)

Und in der *Kritik der praktischen Vernunft* heißt es:

> Die Maxime der Selbstliebe (Klugheit) *räth* bloß an; das Gesetz der Sittlichkeit *gebietet*. Es ist aber doch ein großer Unterschied zwischen dem, wozu man uns *anräthig* ist, und dem, wozu wir *verbindlich* sind. (Kant 1788: 36)

Offensichtlich schreibt Kant hier dem kategorischen Imperativ (dem Gesetz der Sittlichkeit) gegenüber hypothetischen Imperativen (bloßen Maximen der Selbstliebe) einen kategorial anderen Geltungsmodus zu, der moralischen Forderungen stets eine prioritäre Verbindlichkeit sichert. In jüngerer Vergangenheit ist die allgemeine Geltung der Vorrangthese aber aus verschiedenen Gründen bezweifelt worden. Zu nennen sind in diesem Zusammenhang vor allem Philippa Foot (1978a, 1978b), John L. Mackie (1977: 99–102), David Copp (1997) sowie – im deutschen Sprachraum – Dieter Birnbacher ([2]2007: 40–42) und Héctor Wittwer (2010: 357–61; 2011).

Das Ziel meiner Argumentation besteht darin, diese Kritik zu entkräften und eine plausible Version der Vorrangthese zu verteidigen. Dabei gehe ich in fünf Schritten vor: Ich stelle zunächst einige Überlegungen zum Status der Vorrangthese an und stelle fest, was daraus in Bezug auf ihre Begründung folgt (2). Daraufhin werden die beiden gegenwärtig meistdiskutierten Explikationsansätze der Vorrangthese voneinander abgegrenzt. Ich werde mich dagegen aussprechen, die Vorrangthese als allgemeines Vernunftgebot aufzufassen, und dafür votieren, sie als ein moralisches Gebot zu verstehen (3). Diesen Vorschlag werde ich dann gegen zwei wichtige Einwände verteidigen: gegen den Trivialitätseinwand von Foot (4) und den Unverständlichkeitseinwand von Wittwer (5). Abschließend werden zwei Fallbeispiele diskutiert, die in der Debatte häufig angeführt werden, um die Unplausibilität der Vorrangthese zu erweisen. Es wird gezeigt, dass diese bei näherer Analyse nicht gegen den Vorrang der Moral sprechen (6).

## 2. Der Status der Vorrangthese

Kants Behauptung einer kategorial prioritären Verbindlichkeit moralischer Forderungen vor allen Regeln der Klugheit ist systematisch fundiert in seiner Theorie der Autonomie des Willens und der Heteronomie der Willkür. Während letztere durch Naturgesetze und Neigungen „pathologisch“ bestimmt wird, gewinnt erstere ihre unbedingte Geltung aus der reinen (nicht empirisch affizierten) praktischen Vernunft allein (Kant 1788: 33ff.). Kants Vorrangthese ist somit tief eingebettet in seine Metaphysik der Moral. Sie hat im Rahmen dieser Konzeption lediglich den Status eines Korollars, das aus grundlegenden Vernunftprinzipien ableitbar ist.

Daraus folgt aber nicht, dass sich die Vorrangthese *nur* vor dem Hintergrund von Kants Theorie der praktischen Vernunft oder eine vergleichbar anspruchsvollen Metaphysik der Moral verteidigen ließe. Ich werde im Folgenden eine weitaus bescheidenere Statusbestimmung der Vorrangthese vertreten, die sich an Richard Hares Rekonstruktion der Moralsprache orientiert. Hare nimmt an, dass unsere Moralsprache primär durch drei Bestimmungen zu charakterisieren sei: Neben Präskriptivität und Universalisierbarkeit moralischer Urteile verweist er auch auf deren Vorrang, den er unter dem Schlagwort „overridingness“ prominent vertreten hat (Hare 1981: 53–60). In diesem Zusammenhang grenzt Hare moralische Prinzipien von anderen handlungsrelevanten Prinzipien definitorisch ab wie folgt:

> A man's moral principles, in this sense, are those which, in the end, he accepts to guide his life by, even if this involves breaches of subordinate principles such as those of aesthetics and etiquette. (Hare 1963: 169)

Bevor ich auf den Status der Vorrangthese eingehen kann, ist vor dem Hintergrund der Zitate von Kant und Hare ein kurzer terminologischer Exkurs nötig. Denn offensichtlich besteht keine Einigkeit hinsichtlich des genauen Gehalts der These. Dies wird augenfällig, wenn man danach fragt, *welchen Gegenständen* eigentlich der Vorrang zugeschrieben werden soll. Hare schreibt ihn an der zitierten Stelle moralischen Prinzipien, an anderer Stelle moralischen Urteilen zu (Hare 1981: 55). Kant schreibt ihn Prinzipien, Imperativen oder Maximen zu. All diese Optionen sind problematisch: Kants Begrifflichkeit wird nur vor dem Hintergrund seiner Metaphysik der Moral verständlich, die im Folgenden nicht als notwendige Bedingung für die Geltung der Vorrangthese vorausgesetzt werden soll. Hares Prinzipienbegriff ist zu eng, weil gemeinhin nicht nur moralischen Regeln einer hohen Allgemeinheitsstufe (eben: Prinzipien), sondern auch spezialisierten Konkretisierungen moralischer Regeln Vorrang vor anderen speziellen Regeln der Klugheit oder Etikette zugeschrieben wird. Die Bezugnahme auf Moralurteile ist dagegen zu weit, weil dabei auch gänzlich unbegründete oder falsche Moralurteile einbezogen würden. All diese Probleme werden durch eine theoretische Option

vermieden, die in der Debatte z. B. Copp (1997) und Wittwer (2011) vertreten haben und die ich meinen Überlegungen zugrundelege: Im Folgenden wird *moralischen Gründen* der Vorrang vor nichtmoralischen Gründen zugeschrieben. Moralische Gründe werden dabei als eine Sorte von Handlungsgründen charakterisiert. Diese zeichnen sich dadurch aus, dass sie zum einen motivationale Kraft haben (d. h. Handlungen motivieren können, aber nicht in jedem Einzelfall eine Handlung motivieren), und zum anderen zur Rechtfertigung von Handlungen angeführt werden.

Fasst man nun die Vorrangthese als eine These über moralische Gründe auf, dann lässt sich ihr Status im Anschluss an Hare wie folgt bestimmen: Sie ist eine *semantische These*, deren Gehalt in unserer Alltagssprache der Moral verankert ist. Moralische Gründe zeichnen sich demnach *per definitionem* dadurch aus, dass sie in unserem Handeln gegenüber nicht-moralischen Gründen (z. B. der Ästhetik oder der Etikette) vorrangig befolgt werden sollten. Genau dies scheint mir Hare im oben angeführten Zitat auszusagen: Hat ein Handelnder die Vorrangthese akzeptiert und sieht einen Handlungsgrund *h* in einer spezifischen Handlungssituation (in der sonst nur nichtmoralische Handlungsgründe relevant sind) als nicht für sein Handeln verbindlich an, so impliziert dies, dass er *h* nicht als *moralischen* Handlungsgrund ansieht. Demnach stellt die Vorrangthese eine analytische Wahrheit dar, die sich aus der Bedeutung unseres normalsprachlichen Begriffs moralischer Gründe ergibt.

Diese Statusbestimmung der Vorrangthese hat mehrere wichtige Konsequenzen für ihre argumentative Verteidigung. Erstens lege ich mich aufgrund des Verweises auf die Verankerung in unserer Moralsprache nicht darauf fest, dass die Geltung der Vorrangthese in allen überhaupt denkbaren oder zumindest in allen plausiblen Theorien der Moral akzeptiert werden müsste. Vielmehr gestehe ich explizit zu, dass sie möglicherweise nicht mehr Bestandteil eines verbesserten oder idealen Verständnisses von Moral wäre (das z. B. durch eine umfassende Moral- und Rationalitätstheorie begründet werden könnte). Wichtig ist in diesem Zusammenhang nur, dass eine solche verbesserte Auffassung von Moral in relevanter Weise von unserem alltäglichen Begriff von Moral abweichen müsste. Der Unterschied, auf den es an dieser Stelle ankommt, lässt sich an der Distinktion von deskriptiver und revisionärer Metaphysik erläutern, den Strawson (1959) eingeführt hat. Die Vorrangthese ist insofern Bestandteil unserer *deskriptiven Metaphysik* der Moral, als sie Teil des Begriffsschemas (*conceptual scheme*) ist, mit dem wir die Alltagsmoral konzeptualisieren. Nun gibt es, wie Stroud (1998) gezeigt hat, durchaus Moraltheorien, die der Vorrangthese widersprechen. Dieser Befund ist genau dann verträglich mit der hier vertretenen Position, wenn es sich bei diesen Moraltheorien um *revisionäre* Moraltheorien handelt, die grundlegende Bestandteile unserer Normalsprache der Moral nicht abbilden.

Aus dieser Statusbestimmung folgt – zweitens –, dass die Vorrangthese nicht direkt, d. h. durch das Anführen guter rationaler Gründe gestützt werden kann. Man kann allenfalls indirekt für sie argumentieren – und dies primär durch den Verweis auf Intuitionen, die unser alltägliches Moralverständnis bestimmen. Dieses Begründungsverfahren bezeichne ich als „indirekt“, weil die These dadurch nicht positiv gestützt wird, sondern sich ihre Plausibilität nur durch den Nachweis ergibt, dass ihre Leugnung zu Widersprüchen mit Intuitionen führt, die sich aus unserer alltagssprachlichen Verwendung moralischer Grundbegriffe ergeben. Dies scheint zunächst ein schwaches, weil nur negatives Begründungsverfahren zu sein. Es sei jedoch daran erinnert, dass auch andere analytische Wahrheiten über die Moralsprache – wie deren Präskriptivität und Universalisierbarkeit – nur schwerlich positiv durch gute Gründe gestützt werden können. Was wäre ein gutes Argument dafür, dass Moralurteile tatsächlich universalisierbar sind? Es scheint doch vielmehr so zu sein, dass die Universalisierbarkeit so fest in unserer Moralsprache verankert ist, dass wir über jemandem, der sie leugnet, aussagen würden, dass er nicht mehr unseren Begriff der Moral teilt, sondern eine revisionäre Theorie der Moral vertritt. Meines Erachtens verhält es sich bei der Vorrangthese genau so. Deshalb wird sie indirekt gestützt durch ihre

Verträglichkeit mit den sprachlichen und den moralischen Intuitionen, die für unser alltägliches Verständnis der Moral wichtig sind – und es spräche gegen sie, wenn sie mit diesen Intuitionen in Konflikt geriete.

Wie aus dem soeben Gesagten hervorgeht behaupte ich – drittens –, dass die Vorrangthese nicht nur durch sprachliche Intuitionen zum Begriff der Moral indirekt gestützt wird, sondern auch durch ihre Verträglichkeit mit *moralischen Intuitionen*. Dies steht nun *prima facie* im Widerspruch zum Neutralitätsgebot der Metaethik, demzufolge eine metaethische These (und eine These zur Bedeutung moralsprachlicher Grundbegriffe ist eine metaethische These) keine Antworten auf ethisch-normative Fragen beinhalten darf.[1] Gerät man nicht unweigerlich in Konflikt mit dem Neutralitätsgebot, wenn man einerseits vertritt, dass die Vorrangthese eine *semantische* These zur Moralsprache sei, andererseits aber auch behauptet, dass sich ihr Gehalt in normativen Moralurteilen ausdrückt? Obwohl ich beides vertrete, ergibt sich meines Erachtens kein Widerspruch zum Neutralitätsgebot. Denn ich behaupte weder, dass die Annahme der Vorrangthese spezifische ethisch-normative Festlegungen *voraussetzt*, noch behaupte ich, dass aus der Vorrangthese *allein* solche folgen würden. Ich lege mich lediglich auf die bescheidenere Annahme fest, dass sich aus der Vorrangthese im Verbund mit anderen Moralurteilen ethisch-normative Moralurteile ableiten lassen, die ohne die Annahme der Vorrangthese nicht ableitbar wären. Weil dieser Gedanke für die Entkräftung eines Einwands gegen die Vorrangthese wichtig ist (s. Abschnitt 4), möchte ich ihn an einem Beispiel illustrieren. Ich wähle dazu eine semantische These zur Moralsprache, an deren Analytizität gemeinhin nicht gezweifelt wird: die bereits erwähnte Universalisierbarkeit moralischer Urteile. Diese These besagt in der Explikation von Jörg Schroth (2001: 52):

> **Universalisierbarkeitsthese**
>
> Wenn die Handlung *h* gut ist, so ist auch jede Handlung gut, die *h* in moralisch relevanter Hinsicht gleicht.

Aus diesem Konditional folgt für sich genommen kein Moralurteil. Wenn man aber als zweite Prämisse das Moralurteil „Handlung *h* ist gut." hinzufügt, so lässt sich ableiten, dass jede Handlung, die *h* in moralisch relevanter Hinsicht gleicht, gut ist. Dieses inhaltlich stärkere Moralurteil, das sich nun auf *alle* Handlungen erstreckt, die *h* in moralisch relevanter Hinsicht gleichen, würde ohne die Universalisierbarkeitsthese nicht folgen. Ableitbar ist es erst aus dem Moralurteil „Handlung *h* ist gut." *im Verbund mit* der Universalisierbarkeitsthese.[2] An diesem einfachen Beispiel wird deutlich, dass es durchaus Bedingungen gibt, unter denen aus semantischen Thesen zur Moralsprache (im Verbund mit weiteren Annahmen) Moralurteile folgen. Und diese ethisch-normativen Folgerungen können in Abhängigkeit von ihrer Verträglichkeit mit unseren moralischen Intuitionen indirekt für oder gegen die Plausibilität der jeweiligen semantischen These sprechen. Wie noch deutlich wird, verhält es sich bei der Vorrangthese genau so.

Viertens schließlich ist festzuhalten, dass man sich mit der Annahme der Vorrangthese tatsächlich auf *eine* metaethische Vorannahme festlegt, die gegenwärtig durchaus umstritten ist. Wer nämlich moralischen Gründen gegenüber nichtmoralischen Gründen einen Vorrang zuschreiben will, der muss zunächst einmal über die begrifflichen Mittel verfügen, moralische Gründe von nichtmoralischen abzugrenzen, d. h. er muss moralische Gründe als Handlungsgründe *sui generis* auszeichnen. Über diese begrifflichen Mittel aber verfügt nicht jede Rationalitätstheorie. So wird praktische Rationalität in der Humeschen Tradition häufig ausschließlich als Zweck-Mittel-Rationalität konzipiert, die nur instrumentelle Gründe zur Erreichung arationaler Zwecke oder zur Verwirklichung arationaler Wünsche bereitstellt. Vor

[1] S. für eine differenzierte Diskussion des Neutralitätsgebots der Metaethik Hallich (2008: 55–65).

[2] Dies ist der Grund, warum die Universalisierbarkeitsthese, obwohl analytisch, für die moralische Argumentation nicht irrelevant ist, wie Schroth (2001: 113–7) überzeugend nachgewiesen hat.

dem Hintergrund einer solchen Rationalitätskonzeption lässt sich nur schwer verständlich machen, wodurch sich spezifisch moralische Gründe auszeichnen, denen dann ein Vorrang vor anderen Handlungsgründen eingeräumt werden könnte. Deshalb folge ich in dieser Hinsicht der Kantischen Tradition und lege mich auf die These fest, dass man zwischen moralischen Handlungsgründen und anderen rationalen, aber nichtmoralischen Handlungsgründen unterscheiden kann – wobei für die hier verfolgten argumentativen Zwecke offen bleiben kann, aufgrund welcher Kriterien diese Unterscheidung erfolgt. Denn nur unter dieser Vorannahme ist die Vorrangthese sinnvoll und verständlich (vgl. Wittwer 2011: 329–31).

## 3. Zwei konkurrierende Explikationsansätze

Die Vorrangthese besagt, man solle moralischen Gründen vor anderen praktischen Gründen den Vorrang einräumen. Damit ergibt sich die Frage, welche Sorte normativer Verbindlichkeit an dieser Stelle mit dem deontischen Ausdruck „sollen“ angesprochen wird. In der gegenwärtigen Kontroverse lassen sich grundsätzlich zwei konkurrierende Explikationsansätze voneinander unterscheiden:

**Zwei Explikationsansätze der Vorrangthese**

(A) Es ist *moralisch geboten*, moralischen Gründen vor allen anderen praktischen Gründen den Vorrang einzuräumen.

(B) Es ist *durch die Vernunft geboten*, moralischen Gründen vor allen anderen praktischen Gründen den Vorrang einzuräumen.

Die zentrale Streitfrage zwischen den Vertretern von (A) und von (B) lautet somit wie folgt: Thematisiert die Vorrangthese ein moralisches Gebot oder ein Vernunftgebot? Copp (1997) und Wittwer (2011) votieren dafür, diese Frage zugunsten von (B) zu entscheiden, und entwickeln daraufhin überzeugende Argumente gegen einen als Vernunftgebot verstandenen Vorrang der Moral.

(B) aber kann nicht die Vorrangthese sein, der ich hier auf der Spur bin. Denn wie Wittwer in Anlehnung an ein Argument von Frankfurt (2000: 261–3) selbst feststellt, ist (B) unverträglich mit einigen grundlegenden Intuitionen zur Alltagsmoral. Diese Intuitionen betreffen die Emotionen oder die Haltungen, mit denen wir auf den Bruch eines Vernunftgebots und den Bruch eines moralischen Gebots reagieren. Während wir jemanden, der gegen ein Vernunftgebot verstößt für dumm, töricht, unüberlegt oder nachlässig halten, reagieren wir auf den Bruch eines moralischen Gebots mit der Verurteilung der Handlung als unmoralisch, d. h. je nach Kontext als böse, verletzend, hinterhältig, empörend etc. Mithin unterscheiden wir in unseren Reaktionen zwischen unvernünftigen und unmoralischen Handlungen. Ist nun der Verstoß gegen den Vorrang moralischer Handlungsgründe unvernünftig oder unmoralisch? Betrachten wir ein einfaches Beispiel: Peter hat seinem Freund Markus versprochen, ihm beim Umzug zu helfen. Kurzfristig erhält er aber die Gelegenheit, an dem fraglichen Samstag an einem attraktiven Ausflug teilzunehmen. Peter sagt Markus, er habe einen unerwarteten, aber wichtigen Geschäftstermin und nimmt am Ausflug teil. Er bricht also sein Versprechen und belügt Markus (moralische Gründe) und handelt nach seinen Wünschen (prudentielle Gründe). Später erfährt Markus vom Ausflug und von Peters Verhalten. Wie wird er reagieren? Er wird Markus' Verhalten doch nicht als unvernünftig, dumm oder töricht bewerten, sondern als verletzend und unaufrichtig. Unsere Intuitionen stützen also Explikationsvorschlag (A): Wer nichtmoralischen Gründen den Vorrang einräumt, der begeht einen moralischen, keinen rationalen Fehler. Dieser Befund spricht dafür, dass (A) und nicht (B) die Version der Vorrangthese ist, die in unserem

alltäglichen Moralverständnis verankert ist.[3] Bei (B) kann es sich demnach allenfalls um eine These handeln, die in ein revisionäres Verständnis von Moral integriert werden könnte (Wittwer 2011: 328–9).

Trotzdem sind sich Philippa Foot, Harry Frankfurt, David Copp, Héctor Wittwer und andere einig, dass – wenn überhaupt – (B) ein aussichtsreicher Kandidat für eine Verteidigung des Vorrangs der Moral sei. Sie kommen zu dieser Überzeugung, weil sich gegen (A) zumindest zwei fundamentale Einwände formulieren lassen, die (A) entweder als tautologisch oder als unverständlich erweisen sollen. In den folgenden beiden Abschnitten diskutiere und entkräfte ich diese beiden Einwände.

## 4. Der Trivialitätseinwand von Philippa Foot

Wittwer widmet dem Explikationsvorschlag (A) nur eine kurze Bemerkung: Er hält ihn für „tautologisch", weil er lediglich besage, „dass es, vom moralischen Standpunkt aus betrachtet, geboten ist, moralischen Gründen stets den Vorzug vor anderen Gründen zu geben" (Wittwer 2011: 324). Damit schließt er sich Philippa Foots Auffassung an, derzufolge die Frage, ob man moralischen Gründen folgen *solle* oder nicht, in das folgende Dilemma münde:

> There are difficulties about saying that one 'ought' to take account of the general [ethical] good. For either the 'ought' means 'morally ought' or 'ought from the moral point of view' or else it does not. If it does we have a tautological principle. If it does not the problem is to know what is being said. (Foot 1978a: 169)

Gestehen wir zunächst zu, dass Foot und Wittwer die Bezeichnung „Tautologie" hier in einem weiten Sinn verwenden, d. h. nicht logische Wahrheiten im strengen Sinne damit meinen, sondern analytische Wahrheiten – Sätze also, deren Wahrheit sich allein aus der Bedeutung der darin verwendeten Wörter ergibt. Worin aber besteht dann der Einwand? Einige der größten Debatten in der Philosophie kreisen um die Geltung analytischer Wahrheiten: Stellt „‚p' ist genau dann wahr, wenn p." eine Definition des Wahrheitsbegriffs dar? Ist Wissen nichts anderes als „wahre, gerechtfertigte Meinung"? Impliziert jedes moralische Sollen ein Können? Immer geht es um die Aufdeckung analytischer Zusammenhänge. Die Feststellung, dass die Vorrangthese analytisch wahr ist, spricht demnach nicht gegen, sondern für sie.

Diese Erwiderung würde Foots und Wittwers Anliegen – vielleicht mutwillig – missverstehen. Eine Formulierung Wittwers hilft dabei, die Kernintention dieses Einwands gegen (A) besser zu erfassen: So meint Wittwer, dass die Explikation (A) „der Auffassung [dass moralischen Handlungsgründen *in moralischer Hinsicht* Priorität zukommt] nichts hinzufügen" würde (Wittwer 2011: 324). An dieser Stelle wird deutlich, dass es nicht um einen Tautologie- bzw. Analytizitätseinwand geht, sondern um einen *Trivialitätsverdacht*. Ist der moralische Standpunkt erst einmal eingenommen, so meinen Foot und Wittwer, dann ergibt sich die prioritäre Geltung moralischer Gründe trivialerweise, weil ja alle anderen Handlungsgründe bereits aus der Betrachtung ausgeklammert sind. Die folgende Reformulierung von (A) macht den trivialen Charakter der These, an die Foot und Wittwer offensichtlich denken, explizit:

---

[3] Selbst Kant als ethischer Rationalist scheint diese Intuition zu teilen wenn er schreibt: „Der Mensch (auch der Beste) [ist] nur dadurch böse, daß er die sittliche Ordnung der Triebfedern in der Aufnehmung derselben in seine Maximen umkehrt: das moralische Gesetz zwar neben dem der Selbstliebe in dieselbe aufnimmt, da er aber inne wird, daß eines neben dem anderen nicht bestehen kann, sondern eines dem anderen als seiner obersten Bedingung untergeordnet werden müsse, er die Triebfeder der Selbstliebe und ihre Neigungen zur Bedingung der Befolgung des moralischen Gesetzes macht, da das letztere vielmehr als die **oberste Bedingung** der Befriedigung der ersteren in die allgemeine Maxime der Willkür als alleinige Triebfeder aufgenommen werden sollte" (Kant 1793/94: 36). Auch Kant bewertet also den Menschen, der den Triebfedern der Selbstliebe gegenüber denen des moralischen Gesetzes fälschlich den Vorrang einräumt, als *böse* und nicht als *unvernünftig*.

**Triviale Explikation der Vorrangthese**

Wenn **ausschließlich** nach moralischen Gründen gehandelt werden soll, dann ist es moralisch geboten, moralischen Gründen vor allen anderen praktischen Gründen den Vorrang einzuräumen.

Wenn der Explikationsvorschlag (A) wirklich in dieser Weise aufzufassen wäre, müsste man Foot und Wittwer zustimmen: Werden alle anderen Handlungsgründe aus der Betrachtung ausgeschlossen, so wäre (A) tatsächlich trivial. Dies lässt sich daran veranschaulichen, dass ja die triviale Explikation der Vorrangthese nicht nur für moralische Gründe, sondern auch für jede andere Sorte von praktischen Gründen reformuliert werden könnte und analytisch wahr bliebe: Betrachtet man ausschließlich rechtliche Gründe, sollte rechtlichen Gründen der Vorrang gegeben werden, bei der Berücksichtigung ausschließlich ästhetischer Gründe den ästhetischen etc.

Ich denke aber, dass man (A) nicht mit der trivialen Explikation identifizieren *muss* und dass es eine Lesart von (A) gibt, die nicht trivial ist. Wie wir bereits gesehen haben, folgt daraus, dass die Vorrangthese eine analytische Wahrheit ist, keineswegs, dass sie für moralisches Argumentieren irrelevant sei. In Abschnitt 2 habe ich am Beispiel der Universalisierbarkeitsthese vorgeführt, dass sich aus analytischen Thesen zur Moralsprache im Verbund mit anderen Moralurteilen ethisch-normative Moralurteile ableiten lassen, die ohne die Annahme der analytischen These nicht ableitbar wären. Aus diesem Grund ist auch eine analytische Wahrheit wie die Universalisierbarkeitsthese nicht trivial – eben weil sie in unserem Diskurs über ethisch-normative Fragen eine wichtige Rolle spielt. Genauso verhält es sich mit der Vorrangthese: Auch diese hat, obwohl sie selbst eine analytische Wahrheit ist, ethisch-normative Implikationen. Die wichtigste normative Implikation, die im Folgenden noch als Testfall für moralische Intuitionen verwendet wird, sei hier expliziert:

**Normative Implikation der Vorrangthese nach Explikationsansatz (A)**

Wenn

(1) in einer Situation eine Entscheidung zwischen zwei Handlungsoptionen getroffen werden muss,

(2) nur eine der beiden Handlungsoptionen verwirklicht werden kann,

(3) für die eine Handlungsoption auch moralische Gründe und für die andere Handlungsoption nur nichtmoralische Handlungsgründe sprechen

(4) und die moralischen Gründe **berücksichtigt** werden,

dann

ist es moralisch geboten, die Handlungsoption zu verwirklichen, für die die moralischen Gründe sprechen.

Im Antecedens des Konditionals werden die Bedingungen spezifiziert, unter deren Annahmen aus der Vorrangthese ethisch-normative Implikationen folgen.[4] (1) greift hierbei nur den Standardfall heraus; die Bedingung ließe sich auch für Szenarien mit mehr als zwei Handlungsoptionen formulieren. Wichtig ist aber (2), dass nur eine dieser Optionen realisiert werden kann und (3) nur für eine der Optionen moralische Gründe sprechen. Wird dann weiterhin vorausgesetzt, dass (4) moralische Gründe berücksichtigt werden, so folgt aus der Vorrangthese das moralische Gebot, dass der (und nur der) Handlungsoption zu folgen ist, für die die moralischen Gründe sprechen. Der entscheidende Unterschied zur trivialen

[4] Damit tatsächlich normative Implikationen ableitbar sind, müssen die im Antecedens eingebetteten Bedingungen (1) bis (4) noch als eigene, nicht eingebettete Prämissen hinzugefügt werden. Ich verzichte hier aus Platzgründen auf die vollständige Aufzählung der Prämissen.

Explikation besteht darin, dass nicht *ausschließlich* moralische Gründe betrachtet werden, sondern dass lediglich gefordert ist, moralische Gründe zu *berücksichtigen*. Einerseits wird die These dadurch abgeschwächt: Normative Implikationen ergeben sich aus der Vorrangthese nur, wenn die Berücksichtigung moralischer Gründe als gegeben vorausgesetzt wird. Das bedeutet, dass mit dem Verweis auf die Vorrangthese niemand davon überzeugt werden kann, moralisch zu sein. Nur wer bereits die Verbindlichkeit der Moral für sich akzeptiert hat, hat sich damit auch darauf festgelegt, keinen anderen Handlungsgrund einem moralischen vorzuziehen. Dies ist notwendigerweise so, weil die Vorrangthese selbst eine moralische Forderung darstellt.

Im Unterschied zur trivialen Explikation lässt sich dieser Vorrang aber nicht dadurch trivialisieren, dass er auf beliebige Sorten von Handlungsgründen übertragen werden kann. Rechtliche Gründe genießen, sofern man sie nicht ausschließlich in die Betrachtung einbezieht, sondern nur (neben moralischen Gründen) berücksichtigt, diesen Vorrang nicht. Dies lässt sich am Fall des gesetzten Unrechts erläutern. Gesetztes Unrecht besteht genau dann, wenn juridische Setzungen mit moralischen Normen in Konflikt geraten. In unserer Alltagsmoral ist nun die Auffassung tief verankert, dass im Konfliktfall der moralischen Norm Folge zu leisten ist und nicht der positiven Rechtssetzung eines Unrechtsstaats. Weiterhin werden die meisten zustimmen, dass selbst ein künstlerisches Genie sich nicht über grundlegende moralische Normen hinwegsetzen darf: Die Verwirklichung noch so großer ästhetischer Werte kann es nicht legitimieren, moralische Regeln zu brechen.

Damit ist der erste Einwand ausgeräumt: Im Kern handelt es sich bei diesem nicht um einen Tautologie-, sondern um einen Trivialitätseinwand. Der Anschein der Trivialität ergibt sich aber nur aufgrund einer zu restriktiven Explikation von (A), die zugegebenermaßen wirklich trivial ist. Es gibt aber eine schwächere Interpretation, die nicht trivial ist, weil aus ihr ethisch-normative Implikationen folgen können. Diese Implikationen stehen zudem im Einklang mit unseren moralischen Intuitionen. Wenden wir uns der Diskussion des zweiten Einwands zu.

## 5. Der Unverständlichkeitseinwand von Héctor Wittwer: Das Thomas Buddenbrook-Beispiel

Dem zweiten fundamentalen Einwand zufolge ist Explikationsvorschlag (A) schlicht unverständlich. Wenn man annimmt, dass die Vorrangthese eine begriffliche Eigenschaft moralischer Gründe formuliere, so meint Wittwer, ergibt sich das folgende Problem:[5] Die Vorrangthese soll einen unbedingten Vorrang moralischer Gründe vor allen anderen Handlungsgründen festschreiben. Dies kann sie aber nur dann in begründeter Weise leisten, wenn es eine Entscheidungsregel gibt, mit der moralische Gründe auf der einen Seite und prudentielle, instrumentelle, ästhetische Gründe auf der anderen Seite gegeneinander abgewogen werden können. Dies ist aber nur dann möglich, wenn es ein Metakriterium gibt, mit dessen Hilfe sich die normative Verbindlichkeit dieser verschiedenen Typen von Gründen zueinander ins Verhältnis setzen lassen.

> Die Rede vom Vorrang einer Norm vor einer anderen oder eines Handlungsgrundes vor einem anderen ist überhaupt nur verständlich, wenn man dabei ein bestimmtes Kriterium angibt, in Bezug auf welches der Vorrang bestehen soll. Dass etwas vor etwas *schlechthin* den Vorrang habe, ist hingegen eine unverständliche Aussage, weil in

[5] Wittwer formuliert dies als Einwand gegen Explikationsvorschlag (B). Da er sich aber in seiner Diskussion explizit auf eine Version der Vorrangthese bezieht, die den Vorrang als „ein begriffliches, also ein wesentliches Merkmal moralischer Gründe“ ausweist (Wittwer 2011: 332), lässt sich dieser Einwand meines Erachtens auch gegen (A) wenden.

> diesem Falle nicht deutlich wird, *warum* dem einen der Vorrang vor dem anderen gebühren sollte. (Wittwer 2011: 334)

Zunächst ist hier problematisch, dass Wittwers ein zu starkes Kriterium für Verständlichkeit zugrundelegt. Er behauptet, dass die Vorrangthese unverständlich sei, weil nicht begründet werde, warum dem moralischen Grund ein Vorrang gegenüber dem nichtmoralischen Grund gebühren sollte. Nun ist es aber ohne Zweifel möglich, eine Aussage zu verstehen, ohne sie begründen zu können. Ich kann den Pythagoräischen Lehrsatz verstehen, ohne seinen mathematischen Beweis zu kennen. Verständlichkeit und Begründungsfähigkeit sind also zweierlei.

Wittwer legt seinen Überlegungen ganz offensichtlich dieses starke Kriterium für Verständlichkeit zugrunde, weil er annimmt, dass ein Vertreter der Vorrangthese gute, positive Gründe für deren Geltung vorbringen muss. Ich hatte im Abschnitt 2 bereits zugestanden – und am Beispiel der Universalisierbarkeitsthese plausibilisiert –, dass dies im Falle der Vorrangthese deshalb schwierig ist, weil analytische Wahrheiten über die Bedeutung von Grundbegriffen unserer Moralsprache generell nur schwerlich direkt begründet werden können. Alternativ, so hatte ich eingeräumt, lassen sich diese Wahrheiten meistens nur durch den Verweis auf unseren Sprachgebrauch und unsere moralischen Intuitionen in indirekter Weise begründen. Dies möchte ich nun anhand eines ausführlicher entwickelten Fallbeispiels versuchen. Meine Leitfrage lautet dabei: Handelt es sich bei der Vorrangthese überhaupt um eine verständliche These, über deren Wahrheit oder Falschheit man sinnvoll nachdenken kann?

Thomas Mann hat in seinem Roman *Buddenbrooks* eine Szene gestaltet, in der der Protagonist sich mit genau diesem Problem beschäftigen muss. So wird im achten Teil der *Buddenbrooks* dem Senator Thomas Buddenbrook von seiner Schwester Tony das Angebot unterbreitet, die Ernte eines Bekannten „auf dem Halm“ zu kaufen, weil dieser schnellstmöglich 35.000 Courantmark braucht. Zunächst lehnt Thomas entrüstet ab – obwohl Tony ihm den Handel mit allen Mitteln schmackhaft zu machen versucht.

> [Tony] „[...] Dir wird die Gelegenheit geboten, eine gute That zu thun und gleichzeitig das beste Geschäft deines Lebens zu machen ...“
> [Thomas] „Ach was, meine Liebe, du redest lauter Unsinn!“ rief der Senator und warf sich sehr ungeduldig zurück. „Verzeih, aber du kannst einen mit deiner Unschuld in Harnisch jagen! Du begreifst also nicht, daß du mir zu etwas höchst Unwürdigem, zu unreinlichen Manipulationen rätst? Ich soll im Trüben fischen? Einen Menschen brutal ausbeuten? Die Bedrängnis dieses Gutsbsitzers benützen, um den Wehrlosen übers Ohr zu hauen? Ihn zwingen, mir die Ernte eines Jahres gegen den halben Preis abzutreten, damit ich einen Wucherprofit einstreichen kann?“ (Mann 1901: Bd. 2, 93 [Achter Teil, 2. Abschnitt])

Offensichtlich sieht Tony keinerlei Konflikte: Man könne einem Freund helfen (moralischer Grund) und zugleich ein gutes Geschäft machen (prudentieller Grund). Thomas hingegen beschreibt die Situation so, dass die Vorrangthese die im vorigen Abschnitt 4 explizierte normative Implikation zeitigt: Auf der einen Seite mögen prudentielle Gründe (geschäftliche Interessen) stehen, die moralischen Gründe aber sprechen gegen den Handel, weil hier ein in Not Geratener rücksichtslos ausgebeutet würde. Ein moralisch integrer Kaufmann darf sich aber auf einen solchen Handel nicht einlassen. Offensichtlich folgt Thomas hier zunächst der sich in der Vorrangthese ausdrückenden moralischen Forderung und lehnt den Handel ab.

Auf den folgenden Seiten wird dann ausführlich ein Reflexionsprozess von Thomas geschildert, der zu einer gravierenden Überzeugungsänderung führt. Am Ende entscheidet er sich *für* den Kauf der Ernte. Seine Überlegungen lassen sich als eine Reflexion darüber lesen, ob moralischen Gründen *in diesem Fall* der Vorrang zu geben sei. Ich zitiere nur zwei knappe Ausschnitte aus dem Text:

> [...] War Thomas Buddenbrook ein Geschäftsmann, ein Mann der unbefangenen That oder ein skrupulöser Nachdenker?
> Oh ja, das war die Frage; das war von jeher, solange er denken konnte, seine Frage gewesen! Das Leben war hart, und das Geschäftsleben war in seinem rücksichtslosen und unsentimentalen Verlaufe ein Abbild des großen und ganzen Lebens.
> [...]
> „Ich werde es thun!"
> Es war ja wohl das, was man einen Coup nennt? Eine Gelegenheit, ein Kapital von, sagen wir einmal, vierzigtausend Courantmark ganz einfach – und ein wenig übertrieben ausgedrückt – zu verdoppeln? ... Ja, es war ein Fingerzeig, ein Wink, sich zu erheben! Es handelte sich um einen Anfang, einen ersten Streich, und das Risiko, das damit verbunden war, ergab nur eine Widerlegung mehr aller moralischen Skrupeln. (Mann 1901: Bd. 2, 114, 120–1 [Achter Teil, 4. Abschnitt])

Was führt zu diesem Entschluss? Ich denke, dass Thomas weder den moralischen Standpunkt aufgibt, noch sich Tonys Argumentation zu eigen macht, er tue eine moralisch gute Tat. Vielmehr gewichtet er nun die Gründe anders als zu Beginn und folgt dabei nicht mehr der Vorrangthese. Sei er ein „Mann der unbefangenen Tat oder ein skrupulöser Nachdenker", fragt sich Thomas, und weist damit die Priorisierung moralischer Gründe zurück. Es gebe eben andere Werte (Initiative und Unbefangenheit im Handeln, Risikobereitschaft), die auf der anderen Waagschale lägen und die moralischen Gründe in diesem Fall überwiegen würden. Abschließend stellt er sogar fest, dass gerade „das Risiko, das [mit dem Geschäft] verbunden war, nur eine Widerlegung mehr aller moralischen Skrupeln" ergab. Die Übernahme des Risikos also ist es (und damit ein prudentieller Grund), der letztlich alle moralischen Gründe überwiegt.

Habe ich damit aber nicht ein Fallbeispiel präsentiert, das letztlich gegen mein hier verfolgtes Argumentationsziel spricht? Dies ist nicht der Fall. Zunächst habe ich das Fallbeispiel angeführt, um gegen den Unverständlichkeitseinwand zu argumentieren. Und dafür ist es ohne Zweifel geeignet: Es macht deutlich, dass man über die Geltung der Vorrangthese sinnvoll nachdenken und somit auch dann über ein Verständnis vom Vorrang der Moral verfügen kann, wenn man keine direkte Begründung dafür vorlegen kann. Allerdings zeigt es auch, dass man durch bestimmte Überlegungen zu dem Ergebnis kommen kann, die Vorrangthese zurückzuweisen. Dies spricht aber nicht gegen das hier vertretene Verständnis vom Vorrang der Moral. Ich habe im Abschnitt 2 ausgeführt, dass die hier vertretene Vorrangthese nur für unser normalsprachlich verankertes Moralverständnis einschlägig ist und dass es Moraltheorien gibt, für die sie nicht gilt. Vor dem Hintergrund dieser Statusbestimmung lässt sich Thomas Buddenbrooks Überzeugungswechsel in einer Weise erklären, die mit der von mir vertretenen Auffassung verträglich ist: Thomas beginnt seine Überlegungen mit einem – in Strawsons Terminologie – *deskriptiven* Verständnis und beendet sie mit einem *revisionären* Verständnis von Moral. Dies wird am moralischen Gehalt seiner Überlegungen deutlich: So meint er am Ende mit dem Risiko, das mit einem Geschäft verbunden ist, den Bruch moralischer Regeln rechtfertigen zu können. Diese Auffassung mag auch gegenwärtig ihre Vertreter finden – sie widerspricht aber unserem alltäglichen Verständnis von Moral.

Die Interpretation des Fallbeispiels hilft dabei, Wittwers Gründe für seine Ablehnung der Vorrangthese besser zu verstehen. Ich binde mich bei meiner Verteidigung der Vorrangthese an unser alltägliches Moralverständnis und an unsere semantischen Intuitionen zur Moralsprache. Wittwer dagegen meint, dass der Verweis auf den Alltagsbegriff der Moral für die Begründung der Vorrangthese irrelevant sei.

> Der Verweis auf die übliche Verwendung des Begriffs „Moral" [wäre] selbst dann keine schlüssige Begründung der Vorrang-These, wenn der Anspruch auf unbedingte Vorrangigkeit tatsächlich zur Intension des Begriffs gehörte. In diesem Fall müsste

> nämlich nachgewiesen werden, dass die Überzeugung, die in dem Begriff zusammengefasst wäre, wahr ist. Sie wäre es genau dann, wenn der Vorrang der Moral auch ohne Rekurs auf den Sprachgebrauch nachgewiesen werden könnte. (Wittwer 2011: 336)

Wittwer verlangt also nach einer Begründung der Vorrangthese, die unabhängig von unserem Moralverständnis formuliert werden kann und die die Wahrheit des Vorrangs der Moral für *jede* rationale Theorie der Moral erweist – sei sie nun deskriptiv oder revisionär. Die Begründung einer so anspruchsvoll konzipierten Version der Vorrangthese scheint mir tatsächlich aussichtslos zu sein, wie Wittwer (2010: 357ff.) überzeugend nachgewiesen hat. Dies schließt aber nicht aus, dass eine schwächere Version der Vorrangthese gemäß Explikationsvorschlag (A) plausibilisiert werden kann, sofern man ihre Geltung nur für unser alltägliches Moralverständnis behauptet. Diese These, so habe ich in diesem Abschnitt zu zeigen versucht, ist vollkommen verständlich und sie ist verträglich mit einigen wichtigen Intuitionen zur Alltagsmoral. Damit stellt sich die Frage, ob sich nicht andere Intuitionen anführen lassen, die Belege gegen die Geltung der so verstandenen Vorrangthese darstellen.

## 6. Zwei Fallbeispiele

Ist die Vorrangthese tatsächlich verträglich mit allen wichtigen moralischen Intuitionen, die unser alltägliches Moralverständnis ausmachen? Einige Autoren haben Gegenbeispiele präsentiert, die hier abschließend diskutiert werden sollen. Generell sind diese Beispiele so konstruiert, dass sie die vier Bedingungen erfüllen, die in der normativen Implikation der Vorrangthese im Abschnitt 4 formuliert worden sind. Etwas verkürzt formuliert sind sie so konzipiert, dass zwei Handlungsoptionen *a* und *b* unterschieden werden, Option *a* durch moralische Gründe geboten, die mit *a* unverträgliche Option *b* nur durch nichtmoralische Gründe geboten ist, und wir intuitiv befürworten, gegen die moralischen Gründe zu handeln und *b* zu tun. Meines Erachtens gibt es zwei Typen von Gegenbeispielen, die die geeigneten Intuitionen generieren: Der erste Typ bezieht sich auf Notsituationen, der zweite auf Situationen, in denen nur marginale moralische Werte auf dem Spiel stehen.

Dieter Birnbacher führt ein Beispiel des ersten Typs an, wenn er schreibt:

> So kann etwa in Notzeiten der „Kartoffelklau“ aus Selbsterhaltungsgründen die Option der Wahl sein, ohne dass deshalb die Handlung des Diebstahls von Grundnahrungsmitteln ihre in der Regel geltende moralische Bedenklichkeit verlieren muss. (Birnbacher [2]2007: 40)

Birnbachers Idee ist offensichtlich, dass in diesem Fall auf der einen Seite das moralische Gebot steht, den Diebstahl zu unterlassen, und auf der anderen Seite das *rein prudentiell begründete Gebot*, für die Selbsterhaltung zu sorgen. Intuitiv beurteilen wir es als moralisch legitim, den Diebstahl in der geschilderten Situation zu begehen. Diesem Urteil werden sich zumindest im Ergebnis viele anschließen. Ergibt sich dadurch ein Problem für den Vertreter der Vorrangthese?

Dies ist nicht der Fall. Zunächst ist nämlich zweifelhaft, ob für die Selbsterhaltung tatsächlich nur prudentielle Gründe sprechen. So kann z. B. die Mutter, die für sich und ihre Kinder Kartoffeln stiehlt, nicht nur auf prudentielle Gründe verweisen, sondern auch auf die moralische Sorgeverpflichtung gegenüber ihren Kindern. Wie steht es aber, wenn keine Pflichten gegenüber Dritten bestehen? Gestehen wir zu, dass sich in diesem Fall keine moralischen Gründe für den Diebstahl anführen lassen. Allerdings ist hier eine weitere Erwägung von Belang: So spricht Birnbacher vom Kartoffelklau „in Notzeiten“. Dass das Vorliegen von Not einen Ausnahmezustand generiert, der moralische Verpflichtungen in besonderer Weise restringiert, ist tief in unserer Alltagsmoral verankert, was sich z. B. in dem Ausspruch „Not kennt kein Gebot!“ ausdrückt. Der vorliegende Notstand schränkt das

generell geltende Diebstahlsverbot ein und bietet dadurch demjenigen, der in akuter Not für seine Selbsterhaltung sorgen muss, eine moralisch fundierte Legitimationsbasis für sein Handeln. Damit ist nicht unbedingt gesagt, dass der Kartoffeldieb *moralisch richtig* handelt. Vielmehr wird kontrovers darüber gestritten, ob der Notstand hier rechtfertigend oder lediglich entschuldigend wirkt (also ob der Dieb tatsächlich das moralische Recht hat zu stehlen, oder ob er das Recht bricht und lediglich nicht schuldhaft handelt). Wie auch immer dieser Streit ausgehen mag: Im vorliegenden Kontext ist entscheidend, dass das unter Normalbedingungen geltende Diebstahlverbot in Notsituationen nicht ohne Weiteres einen moralischen Grund dafür bietet, dass nicht gestohlen werden darf. Damit aber steht dem Diebstahl kein moralisches Verbot entgegen; vielmehr ist er vielleicht sogar erlaubt. Mithin spricht dieses Beispiel nicht gegen die Vorrangthese, weil der prudentiell gebotenen Handlung kein moralisches Verbot entgegensteht und sich somit kein handlungsrelevanter Konflikt zwischen moralischen und nichtmoralischen Gründen ergibt.

Wenden wir uns dem zweiten Typ von Gegenbeispielen zu. Beispielen dieses Typs ist gemeinsam, dass einem marginalen moralischen Regelbruch ein erheblicher prudentieller Gewinn gegenübersteht. Auf der einen Seite steht z. B. die persönliche Kränkung, die man einer Person zufügt, wenn man es unterlässt, ihr die Hand zu schütteln, wenn man jemandem eine zugesagte Gefälligkeit verweigert etc.; auf der anderen Seite steht ein großer Lottogewinn, den man nicht mehr bekäme, würde man ihn nicht sofort realisieren. Fordern wir wirklich von einem Handelnden, auf den Lottogewinn zu verzichten, nur um das marginale moralische Gut zu verwirklichen? Philippa Foot formuliert ihre Antwort auf diese Frage wie folgt:

> Of course no one expects him to. In face of a sizeable financial consideration a small moral consideration often slips quietly out of sight. (Foot 1978b: 184)

Tatsächlich sind die Intuitionen, mit denen wir auf Gegenbeispiele dieses Typs reagieren, *prima facie* unentschieden. Wirkt es doch auf den ersten Blick unnachgiebig, vielleicht sogar unmenschlich, von jemandem tatsächlich den Verzicht auf große prudentielle Vorteile zu fordern, wenn in moralischer Hinsicht nur Marginales auf dem Spiel steht. Ich möchte einige Überlegungen gegen die Plausibilität dieses Typs von Gegenbeispielen anführen. Diese Überlegungen werde zeigen, dass auch dieser Beispieltyp keinen überzeugenden Beleg gegen die Geltung der Vorrangthese bietet.

Erstens ist Foots Antwort eher als empirisch-psychologische denn als ethisch-normative Auskunft formuliert: Niemand „erwarte“, dass die Person gegen ihre prudentiellen Interessen handle, ein kleines moralisches Opfer gerate im Angesicht großer finanzieller Gewinne „stillschweigend aus dem Blick“. Dies mag durchaus wahr sein, sofern man es als empirischen Bericht über eine Verhaltensbeobachtung auffasst. Vermutlich haben die meisten Menschen Verständnis dafür, dass man durch Nachlässigkeit eine persönliche Kränkung ausdrückt oder eine Zusage bricht, wenn dafür ein Lottogewinn winkt. Diese Tatsache betrifft aber nicht die Geltung der Vorrangthese, die ja festschreibt, was wir tun *sollen*, nicht was wir faktisch tun oder für welche Handlungen wir Verständnis aufbringen. Wenn Foots Antwort also relevant für das hier verhandelte Problem sein soll, müssen wir sie wie folgt reformulieren: *Zu Recht* erwarten wir nicht, dass gegen die prudentiellen Interessen gehandelt wird; es ist *moralisch richtig*, moralisch Marginales bei großem finanziellen Gewinn zu vernachlässigen.

Dies führt mich zu einer zweiten Überlegung: Lässt sich diese explizit ethisch-normative Intuition in unserer Alltagsmoral wiederfinden? Ich gestehe auch dies zu. Trotzdem liegt damit kein intuitiver Befund gegen die Geltung der Vorrangthese vor. Vielmehr verdankt sich diese Intuition der Tatsache, dass wir es als rigide und lebensfremd ansehen würden, in einer solchen Situation das *sensu stricto* moralisch Geforderte zu realisieren. Unnachgiebige Rigidität und Lebensferne in moralischen Fragen stehen aber ihrerseits der Verwirklichung moralischer Werte entgegen. Schließlich sind nicht alle moralischen Fehler *schwere* oder

*unverzeihliche* moralische Fehler. Und die Forderung an den Handelnden, den Lottogewinn zu opfern, um einem Bekannten die Hand zu schütteln (um dessen Kränkung aufgrund dieser Unhöflichkeit zu vermeiden), wäre so wenig verhältnismäßig, dass sie sich in unserer Alltagsmoral nicht wiederfindet – insbesondere deshalb, weil dieser Normverstoß leicht durch eine spätere Erklärung und Entschuldigung ausgeglichen werden kann. Etwas pointiert formuliert: Von einem Freund ernsthaft zu fordern, auf große persönliche Vorteile zu verzichten, damit ein Gruß entboten oder eine kleine Zusage eingehalten werde, scheint ein Ausmaß an Rigidität auszudrücken, das moralisch fragwürdig ist. Ist man nicht moralisch dazu verpflichtet, dem Freund den moralischen Lapsus zuzugestehen, wenn ihm ein solcher Gewinn winkt? Die Gegenbeispiele dieses zweiten Typs scheitern demnach daran, dass ihnen die Eigenschaft fehlt, eine Handlungsoption zu beinhalten, *für die ausschließlich nichtmoralische Gründe sprechen*. Es stehen sich hier zwei Entscheidungsoptionen gegenüber, für die beide moralische Gründe sprechen. Somit ergibt sich auch in diesem Fall kein Widerspruch zur Vorrangthese, weil diese keine normativen Implikationen für moralinterne Konflikte (Konflikte zwischen divergierenden moralischen Regeln) hat.

Zusammengefasst: Beide Beispieltypen können nicht als Belege gegen die Vorrangthese angesehen werden, weil im ersten Fall gar keine konfligierenden Handlungsgründe bestehen und im zweiten Fall ein moralinterner Konflikt vorliegt, für den sich aus der Vorrangthese keine normativen Implikationen ergeben.

## 7. Fazit

Ich habe in dieser Untersuchung dafür argumentiert, dass die Vorrangthese einen Bedeutungsaspekt des Begriffs moralischer Gründe expliziert, der in unserem Alltagsverständnis der Moral fest verankert ist. Diese Vorstellung vom Vorrang der Moral ist dabei weder trivial (denn sie hat normative Implikationen und deshalb Folgen für unser moralisches Argumentieren) noch ist die Vorrangthese unverständlich (denn man kann sinnvoll darüber streiten, ob moralischen Gründen wirklich der Vorrang gegenüber nichtmoralischen einzuräumen sei oder nicht). Außerdem vermag die Vorrangthese in der hier vertretenen Form viele moralische Intuitionen zu integrieren, mit denen wir auf konkrete Situationen moralischen Entscheidens reagieren.

Schließlich ist meine Argumentation mit der Tatsache verträglich, dass in revisionären Moraltheorien moralischen Gründen häufig kein Vorrang vor nichtmoralischen Gründen eingeräumt wird. Dies muss nicht immer die tragischen Konsequenzen haben, die es für Thomas Buddenbrook hatte. Ein Vertreter einer solchen Theorie der Moral muss aber die Frage beantworten, welche guten Gründe dafür sprechen, die Vorrangthese zurückzuweisen, die – vergleichbar der Universalisierbarkeit moralischer Urteile – die alltägliche Verwendung unserer moralischen Grundbegriffe entscheidend mitbestimmt.[6]

**Martin Hoffmann**
Universität Hamburg
martin.hoffmann@uni-hamburg.de

[6] Diese Untersuchung verdankt sich der intensiven Auseinandersetzung mit der inspirierenden Studie von Héctor Wittwer (2011). Für wertvolle Hinweise danke ich außerdem Ali Behboud, Stefan Waller, Fabian Wendt und Annette Wolf.

## Literatur

Birnbacher, D. ²2007: *Analytische Einführung in die Ethik*. Berlin – New York: Walter de Gruyter.

Copp, D. 1997: „The Ring of Gyges: Overridingness and the Unity of Reason“, *Social Philosophy and Policy* 14, 86–106.

Foot, Ph. 1978a: „Morality as a System of Hypothetical Imperatives“, in *Virtues and Vices and Other Essays in Moral Philosophy*, Oxford: Basil Blackwell, 157–73.

Foot, Ph. 1978b: „Are Moral Considerations Overriding?“, in *Virtues and Vices and Other Essays in Moral Philosophy*, Oxford: Basil Blackwell, 181–88.

Frankfurt, H. G. 2000: „Rationalism in Ethics“, in M. Betzler und B. Guckes (Hrg.): *Autonomes Handeln. Beiträge zur Philosophie von Harry G. Frankfurt*, Berlin: Akademie Verlag, 259–73.

Hallich, O. 2008: *Die Rationalität der Moral. Eine sprachanalytische Grundlegung der Ethik*. Paderborn: Mentis.

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

Hare, R. M. 1981: *Moral Thinking. Its Levels, Method and Point*. Oxford: The Clarendon Press.

Kant, I. 1785 [1911]: *Grundlegung zur Metaphysik der Sitten*. zitiert nach *Akademie-Ausgabe* (AA), Band IV, Berlin: Georg Reimer.

Kant, I. 1788 [1913]: *Kritik der praktischen Vernunft*. AA, Band V.

Kant, I. 1793/94 [1907]: *Die Religion innerhalb der Grenzen der bloßen Vernunft*. AA, Band VI.

Mackie, J. L. 1977: *Ethics. Inventing right and wrong*. Harmondsworth: Penguin Books.

Mann, Th. 1901: *Buddenbrooks. Verfall einer Familie*. 2 Bde. Berlin: S. Fischer Verlag.

Schroth, J. 2001: *Die Universalisierbarkeit moralischer Urteile*. Paderborn: Mentis.

Strawson, P. F. 1959: *Individuals. An Essay on Descriptive Metaphysics*. London – New York: Routledge.

Stroud, S. 1998: „Moral Overridingness and Moral Theory“, *Pacific Philosophical Quarterly* 79, 170–89.

Wittwer, H. 2010: *Ist es vernünftig, moralisch zu handeln?* Berlin – New York: Walter de Gruyter.

Wittwer, H. 2011: „Der vermeintliche Vorrang der Moral“, *Zeitschrift für philosophische Forschung* 65 (3), 323–45.

# Practical Knowledge

David Horst

In her book *Intention* Elizabeth Anscombe famously claims that acting intentionally essentially involves knowledge of one's action. Doing something intentionally is doing it knowingly. I shall offer an interpretation and defense of her claim. In doing so, I will mainly argue for two points: first, I show that, if there is such an essential connection between knowledge and intentional action, the relevant kind of knowledge cannot be *observational* or *inferential* knowledge; rather, the knowledge involved in intentional action must be *practical* knowledge. Second, I argue that it is impossible to understand the relevant notion of practical knowledge if we presuppose a compositional view of practical thought, according to which such thought is composed of content and causal force as independent elements. I end by sketching an alternative conception of practical thought, which, I claim, is able to account for the relevant notion of practical knowledge.

Paul is in the kitchen when his wife comes in and asks him, "What are you doing?" "I am preparing a salad", says Paul. To this, his wife responds: "Are you sure? You are cutting tomatoes. You might be doing all kinds of things. Maybe you are not preparing a salad, but making a gazpacho." And to that Paul replies: "Oh my god, you are right! Maybe I am making a gazpacho. Who knows. I guess I have to wait and see what happens next."[1]

This sort of exchange certainly strikes us as absurd. But why? Why is it absurd to imagine that someone is in the process of making a gazpacho, but doesn't know that he is? After all, there is any number of things that might be true of Paul without him knowing that they are. For example, Paul might be developing a tumor without knowing that he is. Or, again, it's possible that Paul is a father and yet does not know that he is. Tragic or unusual as these cases may be, there is nothing absurd about them. So what makes the case in which Paul is preparing a gazpacho so special?

Elizabeth Anscombe in her book *Intention* had a straightforward answer to this question: doing something intentionally is doing it knowingly (Anscombe 2000). That is, if someone performs an intentional action – like making a gazpacho –, then he knows what he is doing. This is the reason why the imagined exchange is indeed absurd. For, if Paul doesn't know that he is making a gazpacho, then whatever he is doing it is *not* the preparation of a gazpacho. This is what distinguishes intentional action from other things that may be true of Paul, like having a tumor or being a father. By taking a paternity test, Paul may find out that he is indeed a father. But the fact that he is a farther in no sense presupposes or implies that he knows this fact. By contrast, if Paul doesn't know that he is making a gazpacho, then there is no such action to be known in the first place. This is what I want to call the *knowledge-requirement* on intentional action: someone who acts intentionally necessarily knows what he is doing.

This knowledge-requirement is not uncontroversial.[2] But I think it is in agreement with our pre-philosophical intuitions about action; and, for the sake of this paper, I will assume that it

[1] I owe this sort of example to John Gibbons (2010). For a similar example compare also Elizabeth Anscombe (2000: 51).

[2] Famously, Donald Davidson considers a case that, at least on first sight, appears to constitute a counterexample to the knowledge-requirement (Davidson 1971: 50). But even Davidson concedes that,

is true. Instead, what I want to focus on is the question of how it *can* be true. That is, given that a satisfactory understanding of intentional action *must* account for the necessary presence of knowledge in intentional action, what I will consider is how such an account is *possible*. Here, I will mainly argue for two claims: First, that if knowledge is necessary for intentional action, the relevant knowledge can be neither *observational* nor *inferential* knowledge. Rather, it must be what Anscombe calls *practical* knowledge, that is: knowledge that is the cause of what it understands. Second, and more importantly, I will argue that a successful account of practical knowledge – and, thus, of the knowledge-requirement – is only possible if we give up a certain widespread conception of practical thought, according to which such thought is composed of content and causal force as independent elements. Instead, what I will suggest is that understanding the relevant notion of practical knowledge requires the specification of a rather special sort of practical thought, namely one that is defined by the essential unity of content and causal force.

Let me begin with a general remark on what it is to act intentionally. I take it that, at least in the fundamental case, if my doing A is an intentional action, then I have the intention to do A and this intention is the cause of my doing A.[3] Where "cause" just means that my intention explains the *existence* of my doing A. That much, I think, should be uncontroversial. Adding the knowledge-requirement, we get the idea that, in doing A because I intend to do A, I know that I am doing A. Now, if knowledge really is necessarily present in intentional action, it should be clear that the relevant knowledge can not be by observation. For, if I only knew by observation what I am doing intentionally, it might very well be that I am doing A because I intend to do A without knowing what I am doing – I might have just failed to observe my behavior. Thus, on an observational conception of knowledge, it would be quite mysterious why the connection between such knowledge and intentional action should be anything but *contingent*. For the same reason, the relevant knowledge cannot be inferential knowledge either. For, if I only knew of my action by an inference from, e.g., my intention and my empirically based belief that my intention usually issues into a corresponding action, it might very well be that I am doing A because I intend to do A without knowing what I am doing – for, again, I might have just failed to draw the relevant inference. So, neither an observational nor an inferential conception of knowledge can account for the truth of the knowledge-requirement.

The general shape of the problem is this. In both cases, my knowledge that I am doing A depends on an act of mind that it *not* already contained in the fact that I am doing A because I intend to do A – in the one case, my knowledge depends on the observation of my behavior and, in the other, it depends on an inference based on my intention and a belief about how my intention is connected to my behavior. But if, *in addition* to acting on my intention to do A, there is a further condition that I have to meet in order to know what I am doing, then it is always possible for these two things to come apart. That is, it is always possible that I am doing A because I intend to do A without knowing that I am doing so, because I failed to meet the relevant additional condition for knowledge. From this, I think we can conclude what must be the case if knowledge is indeed necessarily present in intentional action: I must know that I am doing A *in virtue of* the fact that I am doing A because I intend to do A. Intentional action itself must be a source of knowledge.

---

when someone is acting intentionally, there is *some* description under which she knows what she is doing (Davidson 1971: 50).

[3] The qualification "in the fundamental case" is to make room for the possibility of cases in which an agent is doing A intentionally without having the intention to do A – Michael Bratman's arguments against the Simple View presumably establish such a possibility (Bratman 1987). Still, even if it is possible to do A intentionally without intending to do A, this is so because doing A is a foreseen or desired consequence of an intentional action that is intended by the agent. Thus, we can speak of the case in which doing A intentionally is the execution of the intention to do A as the *fundamental* case.

But how is this possible? I think this is possible only if there is what Anscombe, at one point in *Intention*, calls "knowledge in intention" (Anscombe 2000: 57). Basically, this is the idea that, in acting on an intention, one's knowledge of what one is doing is constituted by one's intention itself. That is, my intention to do A is a thought through which I know that I am doing A; and, since my intention is the cause of my doing A, this means that the relevant thought is the cause of what is known through it. This is why, according to Anscombe, the knowledge of my intentional action is *practical* knowledge, which she defines as knowledge that is the cause of what it understands. If this idea could be rendered intelligible, it would indeed explain why knowledge in intentional action is no accident. For, if the thought through which I know what I am doing is the cause of my doing it, then the existence of my action includes my knowledge of this action. So, the basic idea is that, in order to account for the knowledge-requirement, what we have to understand is how the cause of an intentional action can be, as such, a source of knowledge of this action. However, the difficulty certainly is to understand how this idea of practical knowledge is to be spelled out in any detail. What I want to argue now is that, given a certain widespread conception of practical thought, it is indeed not obvious that there is a satisfactory way of spelling out this idea.

The conception of practical thought that I have in mind is defined by the assumption that a practical thought is basically composed of two independent elements: a conceptual content and a certain causal force. On this view, my intention to do A has a kind of content that it shares with other kinds of thought (for example, with belief); a kind of content that is often taken to be expressible in a proposition like "I am doing A" (at least if, as I assume, it is an intention in action). To this kind of content different attitudes may be attached, which in turn are to be distinguished in terms of their different causal roles. Very roughly, the idea is that my intention is a *practical* thought in that it is the cause of what it represents, whereas a belief is a *theoretical* thought in that it is causally dependent on what it represents (however mediated this causal dependency might be). So, on this view, it is possible to understand what it is for an agent's intention to be a *representation* of the relevant movement without thereby implying an understanding of the intention as being the cause of the movement. And, accordingly, it is possible to understand what it is for the intention to be the *cause* of the movement without thereby implying an understanding of the intention as being a representation of the movement. This is what I want to call the *compositional* view of practical thought.

What's crucial is that, on this view, the fact that my intention is a representation of my doing A does not include, and is not included in, the fact that it is the cause of my doing A. Its being a representation of my doing A and its being the cause of my doing A are two independent features of my practical thought. But if this is so, it is difficult to see how my intention can be a thought that is, as such, knowledge of its own effect – as the knowledge-requirement demands. To see the problem, consider that for my practical thought to be knowledge of its own effect, there must be a *non-accidental* connection between my thought's being the cause of my doing A and its being a representation of my doing A. But, on the compositional view of practical thought, representation and causal efficacy are independent of one another: taken as such, there is nothing in the idea of causality that requires the presence of representation; and, correspondingly, there is nothing in the idea of representation that requires the presence of causality. So, it seems that there is nothing in the relevant notions of causality and representation that explains why it is *not* just an accident if my thought represents what it causes. Consequently, it seems that there is nothing that can account for my practical thought's qualification as knowledge of its own effect.

The problem is not just that my intention might fail to be causally efficacious. Rather, the problem is that, if the causal efficacy of my intention is independent of its content, it seems to be an accident that what my intention causes and what it represents come together in the right way, namely so that the effect of my intention is the movement it represents. And if

that's so, it is hard to see how my intention, just by itself, can account for its being knowledge of its own effect.

In response to this, one may want to introduce a certain *further* condition, one that accounts for some sort of non-accidental connection between my practical thought's efficacy and its content. For example, one may want to stipulate that my thought is a *reliable* cause of what it represents; and that I know from past experience that there is such a reliable connection between my thought and my behavior. On this proposal, my practical thought's qualification as knowledge depends on a *further* thought, one that is based on my experience that, upon thinking a practical thought, something tends to happen that satisfies its content. But, even assuming that awareness of reliability is in general enough for knowledge, this proposal is not good enough to account for the relevant knowledge-requirement. For, as was argued above, if the presence of knowledge in intentional action really is necessary, then this knowledge cannot depend on a further thought, distinct from the one that figures as the cause of my action. Rather, I must know of my action through the thought that is the cause of this action. The point is that, in order to account for the knowledge-requirement, it is not enough that there is *some* connection between my practical thought's causal efficacy and its being a representation of this efficacy, for this does not imply that the connection is known *in* the practical thought itself.

So, in short, my suspicion is that the following two claims do not go together: first, the claim that practical thought is composed of content and causality as *independent* elements; and, second, the claim that practical thought is, as such, a source of *knowledge* of its own effect.

If this suspicion is correct, then there are two possibilities: either we give up the knowledge-requirement or we come up with an alternative to the compositional conception of practical thought. Although I have not given a real argument for the validity of the knowledge-requirement here, it strikes me as a rather intuitive condition on intentional action. For, if knowledge were *not* necessary for intentional action, it would be hard to explain why there should be anything strange in imagining a case in which someone is making a gazpacho without knowing that he is. In other words, there would be nothing wrong with the sort of exchange with which we started. But there obviously is. So, giving up the knowledge-requirement seems to be no option. This leaves us with the second possibility: developing an alternative to the compositional conception of practical thought. Of course, this is not a proper task for the remaining pages of this paper. Instead, what I will, at least, try to do is to briefly point out what such an alternative conception of practical thought would have to render intelligible if it is to provide an account of the knowledge-requirement.

According to the compositional conception, practical thought is composed of content and causality as independent elements. If the foregoing considerations are correct, this is what stands in the way of understanding how practical thought can be, as such, knowledge of its own effect. So, an alternative to the compositional view would have to be a conception of practical thought according to which content and causality are *not* two independent elements, but rather constitute an essential unity. More precisely, we have seen that, if the causal efficacy of my practical thought is independent of its content, it is a mere accident if what my thought represents is what it causes. So, in order to exclude the relevant sort of accident, the efficacy of my thought must be constituted by its content: my thought must be efficacious in virtue of being a representation of its effect.

Now, in order to bring out what this means, consider, first, what it is for my practical thought to be a representation of what I am doing. I take it that, in thinking the practical thought "I am doing A", my thought is a representation of what I am doing in that it is an act of predicating the concept "doing A" to myself. So, if, on the alternative view, my practical thought is supposed to be efficacious in virtue of being a representation of its effect, this must mean that practical thought is distinguished by a specifically *practical* way of predicating the concept "doing A". The basic idea is this: in thinking the relevant practical thought, I deploy

the concept "doing A" in such a way as to *cause* the reality of this concept. Consequently, on this view, what it is for my practical thought to represent my doing A is not independently intelligible of its being the cause of my doing A. In fact, both features are inseparable: to think a practical thought is to predicate the concept "doing A" – and, thus, to represent my doing A – in such a way as to realize this concept. By contrast to the compositional view, this means that, here, the causal force of my practical thought does not attach to a kind of representation that is independently intelligible of being combined with such a force. Rather, on the present view, the notion of causality already enters into the specification of what it is for my thought to be a representation of what I am doing. And this is why we can say that it is essential to the causality of my thought that it is a representation of its own effect.

The point is that, on this alternative view, a subject capable of practical thought is someone who has learned to deploy concepts in a certain way, namely in such a way as to realize theses concepts in action. A thought for which this holds true is *necessarily* a representation of its own effect. And, as this is a character of the way in which the thought represents what one is doing, it will be something one understands just by thinking this thought. Moreover, and more importantly, this conception of practical thought as practical predication seems to be able to account for a practical thought's qualification as knowledge. For, as there is no gap between the thought's representation of my doing A and its being the cause of my doing A, there is also no room for my practical thought's falling short of being knowledge of its own effect. To be sure, this doesn't mean that there is not plenty of room for things to go wrong. All sorts of things may happen and prevent me from successfully carrying out my practical thought. But the basic idea is that if there is practical thought then there is efficacy – however incipient or qualified it may be. And, if this is correct, then it seems that a conception of practical thought in terms of practical predication is what we need in order to do justice to the relevant knowledge-requirement.

No doubt, these sketchy remarks do no more than point to an idea that would have to be developed in far more detail.[4] Also, I am aware that this view of practical thought has many potentially controversial consequences. However, my present point is just that, if what I have been arguing so far is on the right track, then there is reason to think that a satisfactory account of the relevant knowledge-requirement does in fact depend on our ability to make sense of this alternative conception of practical thought. More precisely, I have argued for the following claim: if it is correct that (a) an account of the necessary presence of knowledge in intentional action requires an understanding of *practical* knowledge and (b) that an account of practical knowledge is *not* available as long as the compositional conception of practical thought is in place, then what we need is in fact an account of practical thought along the lines of the notion of practical predication.

**David Horst**

Hebrew University of Jerusalem
davho76@yahoo.de

## References

Anscombe, E. 2000: *Intention*. Cambridge (MA): Harvard University Press.

Bratman, M. 1987: *Intention, Plans and Practical Reason*. Cambridge (MA): Harvard University Press.

[4] I have tried to do this in David Horst (2012).

Davidson, D. 1971: 'Agency', in *Essays on Actions and Events*, Oxford: Oxford University Press 1980, 43–61.

Gibbons, J. 2010: 'Seeing What You're Doing', in T. S. Gendler and J. Hawthorne (eds.): *Oxford Studies in Epistemology 3*. Oxford: Oxford University Press, 63–85.

Horst, D. 2012: *Absichtliches Handeln*. Paderborn: Mentis Verlag.

# Sollen, Können und Versuchen

Michael Kühler

Soll ich ein ertrinkendes Kind *retten* oder soll ich nur *versuchen*, es zu retten? Üblicherweise gehen wir davon aus, dass wir den Erfolg unserer Handlungen nicht vollständig unter unserer Kontrolle haben. Der Handlungserfolg hängt vielmehr auch von Umständen ab, die wir bestenfalls partiell kontrollieren können. Wir können demnach lediglich entsprechende Handlungsversuche unternehmen. Zudem gehen wir gemäß dem Prinzip „Sollen impliziert Können" davon aus, dass wir das, was wir tun sollen, auch tun können. Wenn unser Können jedoch lediglich Handlungsversuche umfasst, dann wäre der Inhalt von Sollensansprüchen konsequenterweise entsprechend einzuschränken. Wir wären somit generell nurmehr zu (erfolgsunabhängigen) Handlungsversuchen aufgefordert. Dieser Argumentation steht unsere Alltagspraxis entgegen, in der das *Gesollte* sehr wohl etwa in der (erfolgreichen) Rettung des Kindes besteht und nicht nur in meinem gegebenenfalls erfolglosen Versuch. In Auseinandersetzung mit den drei wesentlichen handlungstheoretischen Positionen, das Konzept des Versuchens zu explizieren (volitionaler Ansatz, instrumentalistischer Ansatz und Fähigkeitsansatz), argumentiere ich zugunsten unserer Alltagspraxis dafür, dass eine grundsätzliche Einschränkung des Inhalts von Sollensansprüchen auf Handlungsversuche nicht zu überzeugen vermag und das Prinzip „Sollen impliziert Können" keineswegs so streng gilt, wie es zunächst den Anschein hat.

## 1. Einleitung

Soll ich ein ertrinkendes Kind *retten* oder soll ich nur *versuchen*, es zu retten? Soll der Stürmer einer Fußballmannschaft Tore *schießen* oder soll er dies nur *versuchen*? Unserer Alltagspraxis zufolge hat der Stürmer den an ihn adressierten Sollensanspruch keineswegs bereits durch erfolglos bleibende Versuche, ein Tor zu schießen, erfüllt. Der Stürmer *soll* sehr wohl – erfolgreich – Tore *schießen* und dies nicht nur versuchen. Auch im Falle des ertrinkenden Kindes besteht das *Gesollte* üblicherweise in dessen erfolgreicher *Rettung* und nicht nur in meinem gegebenenfalls erfolglosen Versuch – selbst wenn ich im Nachhinein nicht dafür getadelt werde, wenn ich das ertrinkende Kind nicht zu retten vermochte.

Dieser Alltagspraxis steht folgende kritische Überlegung gegenüber: Erstens gehen wir üblicherweise davon aus, dass wir den Erfolg unserer Handlungen nicht vollständig unter unserer Kontrolle haben, sei es mit Blick auf die Durchführung der Handlung selbst oder hinsichtlich der Herbeiführung eines bestimmten Sachverhalts. Der Handlungserfolg hängt vielmehr auch von äußeren Umständen ab, die wir bestenfalls partiell kontrollieren oder beeinflussen können. Als Handelnde *können* wir demnach nicht mehr tun, als entsprechende *Handlungsversuche* zu unternehmen. Zweitens gehen wir gemäß dem Prinzip „Sollen impliziert Können" davon aus, dass wir das, was wir tun sollen, auch tun können.[1] Wenn unser Können jedoch lediglich Handlungsversuche umfasst, dann wäre der Inhalt von Sollensansprüchen konsequenterweise entsprechend einzuschränken. Wir wären somit generell nurmehr zu – *erfolgsunabhängigen* – Handlungsversuchen aufgefordert. Entsprechend wäre ich nur aufgefordert zu *versuchen*, das Kind zu retten, und der Stürmer sollte ebenso nur *versuchen*, Tore zu erzielen.

[1] Dabei ist es hier unerheblich, ob das Prinzip nun im Sinne einer begrifflichen Implikation, einer kolloquialen Implikatur oder im Sinne einer normativen Forderung interpretiert wird. Siehe hierzu Kühler 2013, Kap. 2.

Im Folgenden werde ich diese Argumentation zugunsten einer Einschränkung von Sollensansprüchen auf Handlungsversuche kritisch auf den Prüfstand stellen und dafür argumentieren, dass sie nicht zu überzeugen vermag. Die entscheidende Folgerung lautet dann, dass das Prinzip „Sollen impliziert Können“ keineswegs so streng gilt, wie es zunächst den Anschein hat.

## 2. Die Argumentation zugunsten einer Einschränkung von Sollensansprüchen auf Handlungsversuche

Die Argumentation zugunsten einer Einschränkung lässt sich zunächst in präzisierter Form folgendermaßen fassen:

P1: Akteure haben den Erfolg ihrer Handlungen nicht vollständig unter ihrer Kontrolle.

P2: Vollständig unter Kontrolle haben Akteure ausschließlich ihre Handlungsversuche.

P3: „Sollen impliziert Können“. Forderungen an Akteure richten sich sinnvollerweise nur auf das von Akteuren Kontrollierbare.

K1: Sollensansprüche können grundsätzlich nur Handlungsversuche zum Gegenstand haben.

K2: Unsere Alltagspraxis ist entweder entsprechend zu ändern oder entspricht implizit bereits K1.

Wie überzeugend ist nun diese Argumentation?

## 3. Zu Prämisse 1: Kontrolle und Handlungserfolg

P1 lässt sich schwerlich leugnen. Möchte ich beispielsweise das Licht im Zimmer einschalten und betätige hierfür den Lichtschalter, so hängt der Erfolg der Handlung des Licht-Einschaltens von einer ganzen Reihe von Faktoren ab, z.B. von einer funktionierenden Glühbirne, einer funktionierenden Verkabelung und nicht zuletzt von einer gewährleisteten Stromversorgung. Über zumindest einige dieser Faktoren habe ich im Handlungsmoment des Einschaltens klarerweise keine Kontrolle. So könnte beispielsweise die Stromversorgung ausgefallen sein, selbst wenn ich meine Stromrechnung pünktlich beglichen habe. Der Erfolg einer Handlung hängt insofern in der Tat zum guten Teil von Bedingungen ab, die sich dem Einfluss des Akteurs entziehen.

## 4. Zu Prämisse 2: Kontrolle und Handlungsversuche

Um die Plausibilität von P2 einschätzen zu können, ist zunächst zu klären, was genau unter Handlungsversuchen zu verstehen ist. Die Rede von Versuchen beinhaltet üblicherweise die Möglichkeit des Scheiterns. Denn in Fällen, in denen der Erfolg oder Misserfolg mit Sicherheit eintritt, könnte man sich den Verweis schlicht sparen.[2] In der handlungstheoretischen Debatte finden sich im Wesentlichen drei konkurrierende Analysen:[3]

[2] Vgl. hierzu etwa Heath/Winch 1971, Hornsby 1980, 34, und Hunter 1987.

[3] Für knappe Übersichten siehe Hornsby 1995, Brand 1995, 546, und Wilson/Shpall 2012, Kap. 1.2. Für allgemeine Diskussionen um die Reichweite des Verständnisses von Versuchen siehe Heath/Winch 1971, Hunter 1987 und Schroeder 2001.

(1) ein *volitionaler Ansatz*, der Versuche als Volitionen, d.h. als rein mentale Akte, interpretiert, die die entsprechenden Handlungen kausal herbeiführen;[4]

(2) ein *instrumentalistischer Ansatz*, der Versuche als eigenständige untergeordnete Handlungen interpretiert, die als Mittel durchgeführt werden, um eine komplexere Handlung zu realisieren;[5]

(3) ein *Fähigkeitsansatz*, der Versuche als Aktualisierung geeigneter Fähigkeiten interpretiert, um bestimmte Handlungen erfolgreich auszuführen, wenn der Akteur annimmt, dass es hierbei Schwierigkeiten zu überwinden gilt.[6]

Die drei Ansätze teilen die Ansicht, dass vor allem komplexe Handlungen nicht automatisch erfolgreich sind. Handeln wird deshalb grundsätzlich im Sinne von Versuchen seitens des Akteurs verstanden. Diese können dann erfolgreich sein oder eben nicht. Eine Handlung *x* ist insofern stets zunächst nichts anderes als der Versuch, *x* (erfolgreich) zu tun.

### *4.1 Der volitionale Ansatz*

Das entscheidende Argument für den *volitionalen Ansatz* besteht darin, dass es Versuche gibt, die *völlig* scheitern.[7] Versuche ich beispielsweise meinen Arm zu bewegen, so kann dieser Versuch, wenn mein Arm vollständig gelähmt ist, völlig scheitern und also keinerlei Körperbewegungen umfassen, nicht einmal Ansätze davon. Dennoch kann ich offenbar zutreffend behaupten, dass ich immerhin versucht habe, ihn zu bewegen. Gesteht man angesichts dieses Beispiels zu, dass es Fälle gibt, in denen zwar keinerlei Körperbewegungen vorliegen, wir aber dennoch zutreffend behaupten können, der Akteur habe versucht zu handeln, so sind Versuche offenbar zunächst nur rein mentale Akte, zu denen darüber hinaus der Handelnde einen privilegierten Zugang hat.

Der volitionale Ansatz geht jedoch zudem davon aus, dass, wenn *alle* Handlungen zunächst Versuche sind oder solche zumindest beinhalten, auch *alle* Versuche *nur* mentale, volitionale Akte sind. Die Handlungen werden dann nurmehr kausal von diesen mentalen Akten hervorgebracht – oder bei einem völligen Scheitern eben auch nicht.[8]

Erstens folgt aus der bloßen Möglichkeit rein mentaler Akte des Versuchens allerdings nicht, dass *alle* Versuche *nur* mentale Akte sind.[9] Und zweitens führt die Annahme einer *eigenständigen* Handlung des Versuchens entweder in einen Regress, oder aber rein mentale Akte des Versuchens können niemals scheitern. In einen Regress führt die Konzeption, wenn ein Akteur versucht, *x* zu tun, und dieser Versuch selbst als – mentale – Handlung aufzufassen ist. Denn dann muss der Akteur offenbar auch versuchen können zu versuchen, *x* zu tun, usw. Jennifer Hornsbys wenig überzeugender Versuch, diesem Regress zu entgehen, läuft denn auch schlicht darauf hinaus, Versuche als diesbezüglich in Frage kommende Handlungen einfach dogmatisch auszuschließen.[10]

Soll der Regress vermieden werden, so hat dies jedoch zur Folge, dass rein mentale Akte des Versuchens niemals scheitern können. Wenn ich nicht – mental – versuchen kann, einen

---

[4] Siehe hierfür v.a. Hornsby 1980, Kap. III und IV, sowie O'Shaughnessy 1973 und O'Shaughnessy 2008, Kap. 12. Hornsby hat sich später von einem volitionalen Ansatz distanziert. Vgl. Hornsby 1995.

[5] Siehe hierfür Taylor 1966, Kap. 6, bes. 79ff.

[6] Siehe hierfür Hornsby 1995 und Brand 1995.

[7] Für das Folgende vgl. Hornsby 1980, 40ff., O'Shaughnessy 2008, 386f., und Grünbaum 2008.

[8] Pointiert dargestellt bei Schroeder 2001, 213-216, O'Shaughnessy 2008, 385, und Grünbaum 2008, 68.

[9] Vgl. Schroeder 2001, 214 und 216.

[10] Vgl. Hornsby 1980, 63ff. Später hat sie denn auch die Vorstellung einer eigenständigen Handlung des Versuchens als verfehlt zurückgewiesen und „Versuchen" vielmehr einen adverbialen Status zugesprochen. Vgl. Hornsby 1995, 527.

ebenfalls rein mentalen Akt des Versuchens zu versuchen, dann ist mein derart eng verstandenes Handeln – der Versuch als Versuch – stets erfolgreich.[11]

Dies wiederum hat mit Blick auf Sollensansprüche die absurde Konsequenz, dass deren Nichterfüllung nicht mehr vorstellbar ist. Denn wenn sich Sollensansprüche inhaltlich nurmehr auf Handlungsversuche beziehen und diese wiederum lediglich in mentalen Akten bestehen, kann sie jeder Adressat problemlos erfüllen – zumal dies aufgrund der privilegierten Zugangsweise des Akteurs zu den eigenen rein mentalen Handlungen ohnehin von niemand anderem überprüft werden kann. Eine gesellschaftliche normative Praxis wäre vollkommen witzlos. Der volitionale Ansatz führt in dieser Strenge folglich zu kaum akzeptablen, wenn nicht gar absurden Konsequenzen.

### *4.2 Der volitionale Ansatz: Handlungsversuche und Erfolgsbezug*

Für die Argumentation zugunsten einer Einschränkung von Sollensansprüchen auf Handlungsversuche ist aber noch eine weitere Schwierigkeit des volitionalen Ansatzes entscheidend. Denn die Annahme einer *eigenständigen* Handlung des Versuchens führt nicht nur in einen Regress, sondern auch zu dem handlungstheoretischen Problem, diese Handlung überhaupt angemessen zu explizieren.[12] Worin genau sollte sie bestehen? Es müsste sich um einen volitionalen Akt, d.h. um ein Wollen, handeln, das keinen weiteren intentionalen, propositionalen Gehalt hat, d.h. um ein in dem Sinne „reines“ Wollen, als es bewusst ohne Bezug darauf, *was* gewollt wird, konzipiert wäre. Ein so verstandenes „reines“ Wollen ohne intentionale Gerichtetheit aber ist nicht sinnvoll vorstellbar. Wir wollen immer *etwas*. Versteht man Wollen im Kern als optativische Einstellung, so enthält es stets einen intentionalen, propositionalen Gehalt, d.h. dass *etwas* der Fall sein möge.[13] Der volitionale Ansatz kann ein Versuchen folglich nicht explizieren ohne einen Bezug darauf, *was* denn versucht wird.[14] Dieser Bezug aber schließt notwendig Erfolgsbedingungen ein, da andernfalls unklar bleibt, *was* denn versucht wird und wann ein Versuch als gescheitert oder eben erfolgreich zu beurteilen ist. Von einer eigenständigen Handlung des Versuchens kann deshalb nicht sinnvoll die Rede sein.

### *4.3 Der instrumentalistische Ansatz*

Lässt man nun die These des volitionalen Ansatzes fallen, dass Versuche *ausschließlich* in mentalen Akten bestehen, so kann man zum einen die Reichweite der Rede von Versuchen so erweitern, wie sie auch im Alltag mit Blick auf beobachtbares, willentliches Tun verwendet wird, und man kann zum anderen immerhin daran festhalten, dass es Grenzfälle des Versuchens gibt, die in der Tat nicht über einen rein mentalen Akt hinausgelangen. In diesem Sinne lässt sich der *instrumentalistische Ansatz* deuten,[15] der Versuche zwar ebenfalls als Handlungen versteht, diese jedoch nicht auf rein mentale, eigenständige Akte des Versuchens einschränkt, sondern beliebige Handlungen einschließt, sofern diese als Mittel für die Realisierung übergeordneter, komplexerer Handlungen aufgefasst werden. Zudem wird kein kausaler Zusammenhang angenommen, sondern Handlungsversuche werden unternommen,

---

[11] Vgl. Heath/Winch 1971, 202f.

[12] Vgl. bereits Taylor 1966, 78. Für diese und weitere handlungstheoretische Kritik am volitionalen Ansatz siehe zudem etwa Seebaß 1993, 255ff., Schroeder 2001 und Grünbaum 2008.

[13] Vgl. Seebaß 1993, 66, 69 und 249f.

[14] Zudem sind Wollen und Versuchen keineswegs identisch, insbesondere wenn man die alltägliche Verwendung von Versuchen mit Blick auf komplexe Handlungen, die wiederum – ganz im Sinne des instrumentalistischen Ansatzes – untergeordnete Handlungen und damit auch beobachtbares Verhalten umfassen, mit berücksichtigt. Vgl. hierzu Seebaß 1993, 55ff.

[15] Taylor 1966, 80f., allerdings weist rein mentale, volitionale Akte des Versuchens auch als lediglich möglichen Grenzfall zurück.

*indem* man untergeordnete Handlungen ausführt.[16] Versuche ich beispielsweise, den Konstanzer Uni-Kater Sammy zu streicheln, so könnte ich ihn mit etwas Futter anlocken, mich langsam und ruhig auf den Boden setzen und schließlich vorsichtig den Arm nach ihm ausstrecken, wenn er nah genug ist.

Die untergeordneten Handlungen beinhalten allerdings allesamt eine willentliche Komponente und können selbst wiederum erfolgreich sein oder nicht: das Futter könnte Sammy gänzlich kalt lassen, beim Hinsetzen könnte ich ausrutschen und ihn durch meine hektische und lautstarke Reaktion vertreiben und schließlich könnte erneut die Armbewegung völlig scheitern. Der instrumentalistische Ansatz führt für die Analyse der jeweils immer weiter untergeordneten Handlungen folglich zu demselben handlungstheoretischen Dilemma wie der volitionale Ansatz. Entweder die Analyse beinhaltet einen Regress oder es müssen Handlungen als Versuche angenommen werden, die nicht scheitern können. Auch wenn der instrumentalistische Ansatz somit die Reichweite der Rede von Versuchen in plausibler Weise erweitert, so hilft er hinsichtlich der handlungstheoretischen Explikation von Versuchen doch nicht entscheidend weiter.

### *4.4 Der Fähigkeitsansatz*

Der *Fähigkeitsansatz* schließlich interpretiert Versuche deshalb so, dass es sich um direkt intendierte Aktualisierungen geeigneter Fähigkeiten seitens des Akteurs handelt, sofern dieser annimmt, dass bei der Ausführung der anvisierten Handlung Schwierigkeiten zu überwinden sind.[17] Versuche gelten dem Fähigkeitsansatz nach dabei zwar noch immer als Handlungen, nicht mehr aber als eigenständige Handlungen, die sich ohne Verweis darauf, was versucht wird, verständlich machen lassen.[18] Die Rede von Versuchen hängt schlicht von der Wahrnehmung des Akteurs ab. Ein Akteur *versucht* zu handeln, wenn er zwar eine erfolgreiche entsprechende Handlung intendiert, sich aber gewisser Schwierigkeiten bewusst ist, die ihn an diesem Erfolg hindern könnten. Damit ist die Möglichkeit des Scheiterns ausdrücklich eingeräumt.[19]

Der Fähigkeitsansatz vermag damit nicht nur untergeordnete Handlungen innerhalb komplexer Handlungen als Versuche zu erfassen, sondern auch diejenigen Grenzfälle, in denen Versuche völlig scheitern und nicht über einen rein mentalen Akt hinausgelangen. Insofern scheint sich der Fähigkeitsansatz als der plausibelste Kandidat hinsichtlich der Explikation des Versuchens herauszustellen.

Selbst wenn man diese Frage jedoch angesichts der allzu kursorisch gebliebenen Diskussion lieber offenlassen mag, so bleibt zumindest festzuhalten, dass die Rede von einer eigenständigen Handlung des Versuchens abzulehnen ist und die Explikation von Versuchen damit nicht ohne einen Erfolgsbezug darauf auskommt, *was* versucht wird. Was aber bedeutet dies nun für die mögliche Einschränkung des Inhalts von Sollensansprüchen auf

---

[16] Vgl. nochmals Taylor 1966, Kap. 6, bes. 79ff., und Brand 1995, 542.

[17] Vgl. v.a. Hornsby 1995 und Brand 1995.

[18] Vgl. Hornsby 1995, 527. Darüber hinaus kann ein Akteur, der durch Gründe motiviert ist, etwas zu versuchen, nur diejenigen Handlungen versuchen durchzuführen, für die er geeignete Fähigkeiten besitzt und sich deren auch bewusst ist. Vgl. Hornsby 1995, 533f., wobei sie einräumt, dass man sich bestimmte (praktische) Fähigkeiten auch allererst aneignen kann, indem man versucht, die entsprechenden Tätigkeiten auszuführen. Vgl. auch Brand 1995, 543.

[19] Vgl. Brand 1995, 544. Man könnte nun auch hier kritisch nachfragen, ob der Akteur denn versuchen kann, eine bestimmte Fähigkeit zu aktualisieren, und ob dieser Versuch als willentliches Tun oder gar als rein mentaler Akt nicht wiederum als Handlung aufzufassen ist. Denn falls ja, so würde auch der Fähigkeitsansatz darauf hinauslaufen, eine eigenständige Handlung des Versuchens postulieren zu müssen. Dies wird jedoch gerade explizit abgelehnt. Es bleibt damit bei der direkten willentlichen Aktualisierung einer Fähigkeit, die den Handlungsversuch darstellt. Dieser wiederum kann – als insofern bereits aktualisierte Fähigkeit – scheitern, wenn sich die vom Akteur wahrgenommenen Schwierigkeiten tatsächlich als erfolgsverhindernd herausstellen.

Handlungsversuche und damit speziell für P3, d.h. für das Prinzip „Sollen impliziert Können“?

## 5. Zu Prämisse 3: Die Einschränkung von Sollensansprüchen auf Handlungsversuche

Für eine Einschränkung des Inhalts konkreter Sollensansprüche auf Handlungsversuche hat in exemplarischer Weise Elinor Mason argumentiert.[20] Ihr zufolge geht es in der Moral um die Formulierung handlungsleitender Sollensansprüche, die letztlich von der subjektiven Perspektive der Adressaten abhängen. Der Inhalt handlungsleitender Sollensansprüche verdankt sich damit dem, was ein Akteur *glaubt* in einer bestimmten Situation tun zu sollen. Und dies besteht nach Mason lediglich darin *zu versuchen*, eine bestimmte Handlung auszuführen. *Versuchen* versteht sie dabei so allgemein wie möglich, ohne sich auf einen der drei genannten Ansätze festzulegen: Ein Akteur versucht, *x* zu tun, wenn er glaubt, dass *x* logisch und naturgesetzlich möglich ist, und wenn er das tut, von dem er glaubt, dass es am ehesten dazu geeignet ist, *x* herbeizuführen. So zählt in Masons Beispiel etwa auch Alans Beschwörungsgesang als Versuch, sein Auto zu reparieren, wenn Alan glaubt, dies sei angesichts göttlicher Eingriffe das erfolgversprechendste Mittel.[21] Wie überzeugend aber ist im Anschluss daran der weitere Gedanke, dass Alan hiermit den Sollensanspruch erfüllt, sein Auto zu reparieren?

Mason gesteht zunächst den in diesem Zusammenhang wichtigsten Punkt der handlungstheoretischen Diskussion ausdrücklich zu: Versuche sind keine eigenständigen Handlungen, sondern stets Versuche, *etwas* zu tun. Von Forderungen zu eigenständigen, „reinen“ Versuchen ist also keine Rede. Stattdessen hängen subjektive Pflichten,[22] etwas zu versuchen, ausdrücklich von einem Bezug auf *erfolgreiche* Versuche ab.[23] Ein Akteur hält einen Versuch zu handeln nur dann für seine Pflicht, wenn er glaubt, dass dieser Versuch eine hinreichende Chance auf Erfolg hat. Das aber heißt nichts anderes, als dass der Inhalt von Sollensansprüchen letztlich von erfolgreichen Handlungen bestimmt wird.[24] Im Zentrum des

[20] Siehe Mason 2003, bes. 323-327. Dass sie hierbei vor dem Hintergrund einer utilitaristischen Stoßrichtung argumentiert, ist für die hier verfolgte allgemeine Fragestellung nicht weiter von Bedeutung.

[21] Vgl. Mason 2003, 323. Für die Bedeutung der subjektiven Perspektive des Akteurs und seinen Glauben an einen möglichen Erfolg seiner Versuche siehe auch Halberstam 1979, bes. 118, Adams 1995, bes. 550f., sowie kritisch Ludwig 1995.

[22] D.h. Pflichten, deren Bestehen und Inhalt von der subjektiven Perspektive des Akteurs abhängen.

[23] Vgl. Mason 2003, 325f.

[24] William Frankena, auf den Mason ausdrücklich verweist (vgl. Mason 2003, 326, und für das Folgende Frankena 1950, 163), hat argumentiert, dass eine subjektive Pflicht zu Versuchen erst vor dem Hintergrund einer subjektiven Pflicht zu erfolgreichem Handeln verständlich wird, und zwar – ironischerweise – ausdrücklich auf der Basis des Prinzips „Sollen impliziert Können“. Denn wenn ich angesichts der obigen Überlegung etwas nur versuchen, nicht aber erfolgreich tun kann, so entfällt mit der subjektiven Pflicht, es zu tun, offenbar auch die subjektive Pflicht, es zu versuchen. Denn die subjektive Pflicht, es zu versuchen, orientiert sich, wie von Mason zugestanden, an erfolgreichen Versuchen. Mason zufolge aber würde das Prinzip „Sollen impliziert Können“ mit Blick auf die Versuche nicht greifen, da der Akteur diese stets unternehmen kann. Es wäre weiterhin von einer subjektiven Pflicht zu Versuchen auszugehen, auch wenn dem Akteur klar ist, dass die Versuche nicht erfolgreich sein werden. Dies aber widerspricht offenbar ihrer vorherigen These, dass ein Akteur Versuche nur dann für seine Pflicht hält, wenn diese Aussicht auf Erfolg haben. Mason versucht diese Kritik zu entkräften, indem sie zwischen einer Pflicht zu handeln und Gründen zu handeln unterscheidet. Kann ein Akteur nicht erfolgreich handeln, so entfällt nicht nur die subjektive Pflicht, erfolgreich zu handeln, sondern es entfallen auch die Gründe, so zu handeln und entsprechende Handlungsversuche zu unternehmen. Damit wiederum entfällt auch die subjektive Pflicht zu Handlungsversuchen. Vgl. Mason 2003, 326. Bemerkenswert an dieser Stelle ist, dass Mason ausdrücklich subjektive Pflichten zu erfolgreichem Handeln eingesteht, was ihrer ursprünglichen Eingrenzung auf Handlungsversuche

Inhalts von Sollensansprüchen stehen also gerade nicht Handlungsversuche als solche, sondern vielmehr *erfolgreiche* Versuche und also erfolgreiche Handlungen. Diese fungieren als normativer Standard, auf den die handlungsleitende Funktion von Sollensansprüchen aufsetzt. Von einer ernsthaften Einschränkung des Inhalts von Sollensansprüchen auf Handlungsversuche, deren Erfolg oder Scheitern keine weitere Rolle spielen würde, kann insofern jedenfalls keine Rede sein.

Dass Masons Position letztlich nicht zu überzeugen vermag und immer wieder auf den Bezug zu erfolgreichen Versuchen zurückgeworfen ist, wird auch in ihrer weiteren Argumentation deutlich.[25] Die Pflicht zu weiteren Versuchen nach zunächst erfolglosen Versuchen lässt sich eingestandenermaßen nur vor dem Hintergrund des anvisierten Erfolgs der Versuche verständlich machen.[26] Des Weiteren bleibt einmal mehr entweder unklar, wie ein Akteur eine Pflichterfüllung verfehlen kann, wenn er nur zu Versuchen verpflichtet ist, oder Mason ist zu der These genötigt, dass derartige Pflichten stets erfüllt werden (können). Schließlich wirft auch die Beurteilung des Akteurs und seines Versuchens aus der Außenperspektive Schwierigkeiten auf, wenn es um die Frage geht, ob er seiner Pflicht, Versuche zu unternehmen, in geeigneter Weise bzw. „gut genug" nachkommt.[27]

## 6. Zu Prämisse 3: Sollensansprüche, geeignete Versuche und die praktischen Fähigkeiten der Adressaten

Ob ein Akteur allerdings überhaupt geeignete Versuche unternehmen kann, hängt nicht zuletzt von seinen praktischen Fähigkeiten ab. In dem von Mason vorgebrachten Beispiel ließe sich Alans Beschwörungsgesang als ungeeigneter Versuch, das Auto zu reparieren, kritisieren. Dies wiederum führt zu einem weiteren wichtigen Punkt mit Blick auf die zur Debatte stehende Argumentation. Denn nun stellt sich gemäß P3, d.h. dem Prinzip „Sollen impliziert Können", die Frage, ob sich der Inhalt von Sollensansprüchen nicht an den besonderen praktischen Fähigkeiten der Adressaten orientieren muss oder zumindest sollte.[28]

Angenommen, Alan verfügt nicht über die praktische Fähigkeit, ein Auto zu reparieren. Er kann somit auch keine geeigneten Versuche unternehmen. Allenfalls könnte er *willkürlich irgendetwas* tun mit der Intention, das Auto zu reparieren.[29] Vor dem Hintergrund des Prinzips „Sollen impliziert Können" dürften sich Sollensansprüche nun nur an Akteure richten, die über die nötige Fähigkeit verfügen, geeignete und erfolgversprechende Versuche zu unternehmen. Alan zur erfolgreichen Autoreparatur oder zu einem entsprechend geeigneten Versuch aufzufordern wäre somit von vornherein verfehlt.

Selbst eine Einschränkung des Inhalts von Sollensansprüchen auf geeignete Versuche aber führt einmal mehr zu der Frage, inwieweit erfolglos bleibende Versuche auch bei vorliegender Fähigkeit als Erfüllung von Sollensansprüchen gelten können. Denn es ist zwar in der Tat

---

offensichtlich zuwiderläuft und ihre Argumentation bereits an dieser Stelle schwächt. Die Überlegung kommt damit allerdings nicht ohne einen Bezug zu erfolgreichem Handeln aus. Denn es ist erst die Überzeugung der Erfolglosigkeit von Versuchen, die dazu führt, von der subjektiven Pflicht zu Handlungsversuchen abzusehen.

[25] Siehe hierzu Mason 2003, 326f.

[26] Vgl. Mason 2003, 326f., und McConnell 1989, 444f, der diesen Einwand bereits zuvor formuliert hat und auf den Mason an dieser Stelle eingeht.

[27] Vgl. Mason 2003, 327, sowie McConnell 1989, 445, der erneut diese Schwierigkeit bereits zuvor herausgestellt hat und auf den Mason an dieser Stelle nochmals eingeht.

[28] Erneut kann hier zunächst offenbleiben, welche Interpretation des Implikationsverhältnisses zwischen Sollen und Können (begrifflich, normativ, kolloquial) dabei in Anschlag gebracht wird. Vgl. hierzu nochmals oben, Fn. 1.

[29] Vgl. in diesem Sinne nochmals Mason 2003, 323.

naheliegend und auch intuitiv plausibel, weder Alan zu tadeln noch Brian, der im Gegensatz zu Alan über die nötigen Kenntnisse und Fähigkeiten zur Autoreparatur verfügt und dessen Reparaturversuch, obwohl er sein Bestes gegeben hat, ebenso erfolglos bleibt.[30] Zumindest die Beurteilung des Akteurs scheint sich demnach durchaus an den Fähigkeiten des Akteurs und der Güte seiner unternommenen Versuche zu orientieren.

Hätte Brian allerdings Erfolg in seinem Reparaturversuch, müsste man dies als *supererogatorische* Tat ansehen, da er offenbar mehr getan hat, als seine Pflicht war. Denn er hat ja nicht nur gemäß seiner Pflicht versucht, das Auto zu reparieren, sondern er hat es tatsächlich erfolgreich repariert. Zu Letzterem aber war er gemäß der Einschränkung des Inhalts des Sollensanspruchs auf Handlungsversuche gar nicht mehr aufgefordert. Geht man davon aus, dass im Alltag Versuche, Sollensansprüche zu erfüllen, häufig genug erfolgreich sind, so käme es zu einer absurden Inflation supererogatorischen Handelns.

Der Einbezug des Erfolgs von Versuchen im Inhalt konkreter Sollensansprüche erweist sich folglich als unvermeidbar, und zwar nicht nur in der Rolle eines normativen Standards, sondern abgeleitet davon auch mit Blick darauf, wozu genau Brian aufgefordert ist, nämlich nicht nur zu versuchen, das Auto zu reparieren, sondern das Auto erfolgreich zu reparieren. Sein erfolgloser Versuch, sofern er sein anerkannt Bestes gegeben hat, führt lediglich dazu, ihn nicht als tadelnswert zu beurteilen.

## 7. Fazit

Sind die hier angestellten Überlegungen überzeugend, dann sind die anfangs genannten beiden Konklusionen K1 und K2 zurückzuweisen. Die konzeptionellen Schwierigkeiten eines hierfür nötigen eigenständigen Handlungskonzepts des Versuchens haben gezeigt, dass sowohl Handlungsversuche als auch im Anschluss der Inhalt von Sollensansprüchen als begrifflich davon abhängig zu verstehen sind, *was* versucht bzw. gesollt wird. Weder die Explikation von Handlungsversuchen noch diejenige von Sollensansprüchen kommt ohne Erfolgsbezug aus. Sollensansprüche können also weder grundsätzlich auf Handlungsversuche eingeschränkt werden (K1), noch ist unsere Alltagspraxis entsprechend zu ändern oder entspricht dem bereits implizit (K2).

P1 und P2 sind allerdings vergleichsweise überzeugend: Den Erfolg unserer Handlungen haben wir in der Tat nicht vollständig unter unserer Kontrolle (P1), sondern im Sinne des Fähigkeitsansatzes allenfalls unsere Handlungsversuche (P2), wenngleich unter der Voraussetzung, dass die Explikation einen Erfolgsbezug einschließt. In Frage steht deshalb in erster Linie P3, d.h. das Prinzip „Sollen impliziert Können“. Es ist entweder schlicht komplett aufzugeben oder zumindest angemessen zu modifizieren.

Wenn das Sollen auch im Rahmen einer solchen Modifikation jedoch notwendig einen Erfolgsbezug aufweist und inhaltlich nicht auf bloße Handlungsversuche eingeschränkt werden kann, und wenn wir zugleich aber nicht die vollständige Kontrolle über unseren Handlungserfolg haben, so kann das enthaltene Können nicht länger diese vollständige Kontrolle ausdrücken. Denn andernfalls müsste man den alltäglichen Umstand für unmöglich erklären, dass Akteure in ihren Versuchen, einen Sollensanspruch zu erfüllen, hin und wieder eben auch scheitern. Soll an dem Prinzip „Sollen impliziert Können“ also festgehalten werden, so ist das Können deutlich schwächer zu interpretieren, und zwar als die im konkreten Fall *erfolgsunabhängige* Fähigkeit des Akteurs, geeignete Versuche unternehmen zu können. In den zu Beginn genannten Beispielen wäre ich dann sehr wohl aufgefordert, das ertrinkende Kind *zu retten*, allerdings nur wenn ich schwimmen und also einen geeigneten Versuch unternehmen kann. Auch der Stürmer ist aufgefordert, im Spiel

[30] Vgl. in diesem Zusammenhang Halberstam 1979, 123f., und McConnell 1989, 446-449.

Tore *zu schießen*, sofern er ebenso die Fähigkeit zu geeigneten Versuchen hierzu hat. Und in beiden Fällen können die entsprechenden Handlungsversuche dann eben auch scheitern.

Das in dem Prinzip „Sollen impliziert Können" enthaltene Können wird nun zwar üblicherweise ohnehin als Fähigkeit plus Gelegenheit begriffen.[31] Hinzu kommt in den üblichen Formulierungen des Prinzips also noch die situativ vorliegen müssende Gelegenheit des Akteurs, die entsprechende Fähigkeit zu realisieren. Beispielsweise muss der Stürmer, um seine Fähigkeit, Tore zu schießen, aktualisieren zu können, auch am Spiel teilnehmen. Entscheidend ist jedoch, dass im Gegensatz zur Explikation von Handlungsversuchen nach dem Fähigkeitsansatz dabei häufig von *keinerlei* Hinderungen die Rede ist, so dass es *allein* in der Macht des Akteurs zu liegen scheint, ob er den Sollensanspruch erfüllt oder nicht. Gemäß den hier angestellten Überlegungen ist das Können des Akteurs jedoch nochmals in entscheidender Weise abgeschwächt. Denn die Fähigkeit ist ausdrücklich als *erfolgsunabhängig* aufzufassen. Vertreter des Prinzips müssten demnach von der Überzeugung abrücken, dass es *allein* am Akteur liegt, ob er bei vorliegender Gelegenheit durch die willentliche Aktualisierung seiner Fähigkeit den Sollensanspruch – erfolgreich – erfüllt. Damit schließlich ist auch im Rahmen einer Modifikation des Prinzips die wohl wesentliche Intuition für es aufzugeben, nämlich dass von Akteuren nur in ihrer Macht Stehendes gefordert werden kann.[32]

**Michael Kühler**

Westfälische Wilhelms-Universität Münster
Kolleg-Forschergruppe „Normenbegründung in Medizinethik und Biopolitik"
michael.kuehler@uni-muenster.de

## Literatur

Adams, F. 1995: „Trying: You've Got to Believe", *Journal of Philosophical Research* 20, 549-561.

Brand, M. 1995: „Hornsby on Trying", *Journal of Philosophical Research* 20, 541-547.

Frankena, W. 1950: „Obligation and Ability", in M. Black (Hrg.): *Philosophical Analysis*, New York: Arno Press, 1950, 148-165.

Grünbaum, T. 2008: „Trying and the Arguments from Total Failure", *Philosophia* 36, 67-86.

Haji, I. 2002: *Deontic Morality and Control*, New York: Cambridge University Press.

Halberstam, J. 1979: „Trying and Responsibility", *Tulane Studies in Philosophy* 28, 115-124.

Heath, P./Winch, P. 1971: „Trying and Attempting",*Proceedings of the Aristotelian Society* Suppl. Vol. 45, 193-208.

Hornsby, J. 1980: *Actions*, London: Routledge & Kegan Paul.

— 1995: „Reasons for Trying", *Journal of Philosophical Research* 20, 525-539.

Hunter, J. F. M. 1987: „Trying", *The Philosophical Quarterly* 37, 392-401.

[31] Vgl. in diesem Sinne bspw. Haji 2002, 17 und 23f., und Zimmerman 1996, 49f. und 79.

[32] Für eine Einbettung der hier angestellten Überlegungen in eine umfassende Diskussion des Prinzips „Sollen impliziert Können" siehe Kühler 2013. Die Überlegungen hier entsprechen im Wesentlichen Kapitel 4. Die Problematik, dass von Akteuren angeblich nur in ihrer Macht Stehendes gefordert werden kann, zeigt sich zudem in der Diskussion um die Reichweite moralischer Verantwortlichkeit und um die Anerkennung von „Moral Luck". Für eine Übersicht über die entsprechende Debatte siehe Nelkin 2008, für einschlägige Beiträge siehe den Sammelband von Statman 1994. Im Anschluss an die hier angestellten Überlegungen siehe auch Kühler 2012.

Kühler, M. 2012: „'Resultant Moral Luck', 'Sollen impliziert Können' und eine komplexe normative Analyse moralischer Verantwortlichkeit“, *Grazer Philosophische Studien* 86, 181-205.

— 2013: *Sollen ohne Können? Über Sinn und Geltung nicht erfüllbarer Sollensansprüche*, Münster: Mentis, im Erscheinen.

Ludwig, K. A. 1995: „Trying the Impossible: Reply to Adams“, *Journal of Philosophical Research* 20, 563-570.

Mason, E. 2003: „Consequentialism and the Ought Implies Can Principle“, *American Philosophical Quarterly* 40, 319-331.

McConnell, T. C. 1989: „'Ought' Implies 'Can' and the Scope of Moral Requirements“, *Philosophia* 19, 437-454.

Nelkin, D. K. 2008: „Moral Luck“, in E. N. Zalta (Hrg.): *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), URL=http://plato.stanford.edu/archives/fall2008/entries/moral-luck/.

O'Shaughnessy, B. 1973: „Trying (as the Mental 'Pineal Gland')“, *Journal of Philosophy* 70, 365-386.

— 2008: *The Will. A Dual Aspect Theory*, 2 Volumes, 2. Edition, Cambridge: Cambridge University Press.

Schroeder, S. 2001: „The Concept of Trying“, *Philosophical Investigations* 24, 213-227.

Seebaß, G. 1993: *Wollen*, Frankfurt am Main: Klostermann.

Statman, D. 1994 (Hrg.): *Moral Luck*, Albany: State University of New York Press.

Taylor, R. 1966: *Action and Purpose*, Englewood Cliffs: Prentice Hall.

Wilson, G./Shpall, S. 2012: „Action“, in E. N. Zalta (Hrg.): *The Stanford Encyclopedia of Philosophy* (Summer 2012 Edition), URL=http://plato.stanford.edu/archives/sum2012/entries/action/.

Zimmerman, M. J. 1996: *The Concept of Moral Obligation*, Cambridge: Cambridge University Press.

# Drei Arten von Hilfspflichten

Jörg Löschke

Im folgenden Text wird eine Unterscheidung zwischen isomorphistischen und nonisomorphistischen Ansätzen innerhalb der Debatte um positive Pflichten vorgeschlagen. Während isomorphistische Positionen davon ausgehen, dass alle Arten von Hilfspflichten als strukturell gleichartig anzusehen sind, vertreten nonisomorphistische Positionen die Auffassung, dass es strukturelle Unterschiede zwischen verschiedenen Arten von Hilfspflichten gibt, die normativ relevant sind. Nach dieser einleitenden Differenzierung wird eine nonisomorphistische Position in ihren Grundsätzen skizziert. Ausgehend von einer kurzen Darstellung möglicher Differenzierungskriterien zwischen verschiedenen Arten von Hilfspflichten wird das Differenzierungskriterium der prinzipiellen Erfüllbarkeit einer Hilfspflicht vorgeschlagen, das normativ relevante Unterschiede zwischen verschiedenen Arten von Hilfspflichten begründet. Hieraus ergibt sich eine Dreiteilung von positiven Pflichten in situative, projektbezogene sowie konstante Hilfspflichten, deren abnehmende prinzipielle Erfüllbarkeit zu einem abnehmenden deontischen Status führt: Akzeptiert man das Prinzip, dass Sollen Können impliziert, muss man auch das Prinzip akzeptieren, dass ein verringertes Können ein verringertes Sollen impliziert, sodass Hilfspflichten, die prinzipiell leichter zu erfüllen sind, auch einen prinzipiell höheren Verbindlichkeitsgrad aufweisen. Der Text schließt mit einem Ausblick auf Konsequenzen dieser Unterscheidung im Kontext der Debatte um Parteilichkeitspflichten.

## 1. Einleitende Bemerkungen[1]

In der sich immer weiter ausdifferenzierenden Debatte um den normativen Status von positiven bzw. Hilfspflichten haben sich zwei grundsätzliche Positionen herausgebildet, die man als isomorphistische bzw. nonisomorphistische Ansätze bezeichnen kann. Isomorphistische Positionen vertreten die Auffassung, dass alle Formen von Hilfspflichten als strukturell gleichartig zu betrachten sind: Abstufungen im deontischen Status – verstanden als Grad der Verbindlichkeit – können einer solchen Position gemäß zwar auftreten, sie sind aber grundsätzlich abhängig vom Inhalt der jeweiligen Hilfspflicht, und nur von diesem Inhalt. Beispielsweise wird einer Hilfspflicht eine höhere Verbindlichkeit zugesprochen, wenn sie sich auf die Sicherung von zentralen Gütern (wie der Integrität von Leib und Leben) richtet, als wenn sie sich auf vergleichsweise unwichtige Güter richtet. Es werden allerdings keine strukturellen Unterschiede zwischen verschiedenen Arten von Hilfspflichten herangezogen, wenn Hilfspflichten und ihr jeweiliger deontischer Status untersucht werden. Dies ermöglicht weitreichende Analogieschlüsse: Es werden Situationen, in denen Hilfspflichten auftreten, auf die Frage hin untersucht, welche Güter auf dem Spiel stehen, die zu schützen Aufgabe der Hilfspflicht ist; die Einsichten bezüglich des Verbindlichkeitsgrades, die sich aus der Betrachtung einer bestimmten Situation ergeben, in denen eine Hilfspflicht vorliegt, werden dann auf andere Situationen übertragen, in denen Personen ebenfalls auf Hilfe angewiesen sind. Sofern es sich um dieselben Güter handelt, die auf dem Spiel stehen und auf die die jeweilige Hilfsbedürftigkeit bezogen ist, wird ein solcher Analogieschluss als gültig angesehen.

[1] Dieser Aufsatz wurde unterstützt durch den Schweizerischen Nationalfonds zur Förderung der wissenschaftlichen Forschung im Rahmen des SNF-Projekts „Gründe der Parteilichkeit – Zur Ethik der Familienbeziehungen“.

Eine nonisomorphistische Position bestreitet dagegen die Möglichkeit einfacher Ergebnisübertragungen bzw. einfacher Analogieschlüsse von einer Situation auf eine andere. Es werden vielmehr strukturelle Unterschiede zwischen verschiedenen Arten von Hilfspflichten angenommen, die Unterschiede im normativen Status der entsprechenden Pflichten generieren können, und zwar unabhängig von den Gütern, die jeweils auf dem Spiel stehen. Dies bedeutet nicht, dass inhaltliche Erwägungen keine Rolle bei der Beurteilung des deontischen Status einer Hilfspflicht spielen, es bedeutet allerdings, dass solche Erwägungen nicht die einzige Rolle spielen und dass daher einfache Analogieschlüsse von einer Situation, in der eine Hilfspflicht auftritt, auf eine andere Situation, in der eine oder mehrere Personen hilfsbedürftig sind, unzulässig und daher zu vermeiden sind.

Im Folgenden werde ich eine nonisomorphistische Position verteidigen, die strukturelle Unterschiede zwischen verschiedenen Arten von Hilfspflichten postuliert. Dabei werde ich neben den in der Literatur vertretenen und als normativ relevant bewerteten Eigenschaften, die verschiedenen Arten von Hilfspflichten zukommen und Unterschiede im normativen Status generieren, ein Unterscheidungskriterium vorschlagen, das so meines Erachtens noch nicht behandelt wurde, und zwar den Aspekt der prinzipiellen Einlösbarkeit der jeweiligen Hilfspflicht. Basierend auf diesem Kriterium werde ich eine Dreiteilung in verschiedene Formen von Hilfspflichten vorschlagen: Situative Hilfspflichten, projektbezogene Hilfspflichten und konstante Hilfspflichten. Beginnen werde ich mit einer kurzen Darstellung der isomorphistischen Gegenposition zu meinem Ansatz.

## 2. Isomorphistische Konzeptionen positiver Pflichten

Ein prominentes und viel diskutiertes Beispiel für einen isomorphistischen Ansatz stellt die Position von Peter Singer dar. In seinem einflussreichen Aufsatz *Famine, Affluence, and Morality (*Singer 1972) konstruiert Singer ein Beispiel, das mittlerweile klassisch geworden ist, und zwar das Teichbeispiel, in dem die Situation vorgestellt wird, dass ein Passant an einem Teich vorbeikommt, in dem ein Kind ertrinkt und dessen Leben der Passant ohne große Kosten oder große Gefahr für sein eigenes Leben retten kann. In einer solchen Situation hat der Passant laut Singer (und wohl dem überwiegend großen Teil der Leserinnen und Leser, die sich mit Singers Argumentation auseinandersetzen) die Pflicht, in den Teich zu waten und das Kind herauszuziehen, auch wenn dies bedeutet, sich die Kleidung schmutzig zu machen. Das Leben des Kindes ist ein zentraleres Gut als die saubere Kleidung des Passanten, und daher herrscht hier eine starke Hilfspflicht. Singer stellt des Weiteren einen vieldiskutierten Analogieschluss zu der Hilfspflicht auf, Menschen in der Dritten Welt bei der Bekämpfung ihrer lebensbedrohlichen Armut zu helfen. Der Unterschied zwischen diesen beiden Fällen, so Singer, liegt allenfalls in der geographischen Distanz zwischen der Hilfeleistenden und dem Empfänger der Hilfeleistung: Im Teichbeispiel ist diese gering, im Fall der Hilfe angesichts des Welthungerproblems dagegen groß. Da räumliche Entfernung aber nach Singer normativ nicht relevant ist, sind die beiden Fällen gleich zu bewerten: Es ist jeweils Hilfe möglich, durch die zentrale Güter geschützt werden, ohne dass etwas Vergleichbares geopfert wird. Die Hilfspflichten gegenüber Personen in der Dritten Welt haben denselben deontischen Status wie Hilfspflichten gegenüber ertrinkenden Kindern, und das moralische Vergehen bei Nichtbeachtung wiegt in beiden Fällen gleich schwer.

Singer möchte mit seinem Beispiel und dem von ihm angestellten Analogieschluss die Dringlichkeit demonstrieren, die nicht nur mit Blick auf ertrinkende Kinder, sondern auch mit Blick auf das Weltarmutsproblem besteht. Sein Ziel besteht darin, eine revisionistische Auffassung von positiven Pflichten zu begründen: Während für gewöhnlich die Hilfspflichten gegenüber Personen, die von Armut betroffen sind, als weniger dringlich im Vergleich zu Fällen wie dem des ertrinkenden Kindes eingeschätzt werden, möchte er den deontischen Status dieser Pflichten auf Distanz aufwerten. Dagegen macht Onora O'Neill auf ein Problem

aufmerksam, welches ihrer Meinung nach mit Blick auf Hilfspflichten – insbesondere im Kontext der Frage nach den Verpflichtungen bezüglich des Welthungerproblems – besteht: Ihr zufolge stellt sich mit Blick auf positive Pflichten das Unbestimmtheitsproblem. O'Neill diskutiert in ihrem Aufsatz *The dark side of human rights* „rights to goods and services" (O'Neill 2005: 427) und konstatiert, dass es keine fest bestimmbaren Verantwortungsträger gibt, um Güter und Leistungen bereitzustellen, die Gegenstand von positiven Pflichten sein können. Dies ist ihrer Meinung nach ein Problem für die Menschenrechtsdebatte, denn es macht den deontischen Status von positiven Pflichten gewissermaßen obskur.

O'Neills Argument erinnert an Kants Charakterisierung von positiven Pflichten. Auch für Kant sind positive Pflichten dadurch gekennzeichnet, dass sie unbestimmt sind. Die Unbestimmtheit, die Kant positiven Pflichten zuspricht, ist allerdings eine andere als die Unbestimmtheit, wie sie sich im Verständnis von O'Neill zeigt. Kant charakterisiert positive Pflichten – etwa die Pflicht zur Wohltätigkeit – als unvollkommene Pflichten, weil es bei ihnen eine latitudo an Realisierungsmöglichkeiten gibt. So ist es beispielsweise möglich, die Pflicht zur Wohltätigkeit auf verschiedene Weisen zu erfüllen: Man kann gegenüber verschiedenen Personen wohltätig sein, und es mag bis zu einem gewissen Grad unbestimmt sein, was eine wohltätige Handlung inhaltlich auszeichnet. Wohltätigkeit ist eine Pflicht ohne korrelierende Rechte: Grundsätzlich kann Person A keinen Anspruch erheben, dass Person B ihre Pflicht zur Wohltätigkeit durch wohltätige Handlungen ausgerechnet gegenüber A erfüllt.

Hier wird deutlich, inwiefern es zwei Unbestimmtheitsverhältnisse mit Blick auf positive Pflichten gibt, die spiegelbildlich zu verstehen sind. Während O'Neill konstatiert, dass es womöglich klar bestimmbare Anspruchsträger gibt, nicht aber klar bestimmbare Verantwortungsträger, ist das Unbestimmtheitsverhältnis bei Kants Behandlung von positiven Pflichten anders gelagert: Nach ihm gibt es klar bestimmbare Verantwortungsträger (nämlich jede einzelne Person), aber nicht unbedingt klar bestimmbare Anspruchsträger.

Anders als Singer geht es O'Neill nicht darum, den deontischen Status von Hilfspflichten mit Blick auf das Weltarmutsproblem aufzuwerten. Vielmehr möchte sie auf ein grundsätzliches Problem mit Blick auf positive Pflichten hinweisen. Dabei zeigt sich allerdings eine argumentative Tendenz, die der von Singer durchaus ähnelt. Sie spricht von „rights to goods and services", wenn sie das Unbestimmtheitsproblem herausarbeitet, und mit „services" sind Hilfeleistungen gemeint. Da O'Neill hier nicht begrifflich differenziert, liegt die Vermutung nahe, dass auch sie alle Formen von „services" zumindest in formaler Hinsicht als gleichartig auffasst. Auch in ihrem Fall wird also nicht zwischen verschiedenen, strukturell unterschiedlichen Formen von Hilfspflichten differenziert.

## 3. Nonisomorphistische Ansätze

Isomorphistische Ansätze sind in der Diskussion um den normativen Status von Hilfspflichten in die Defensive geraten. Insbesondere Singers Ansatz ist verschiedenen Kritikpunkten ausgesetzt. Während einzelne Autoren Singer darin folgen, dass es keinen normativ relevanten Unterschied zwischen dem Teichbeispiel und Problemen angesichts des Weltarmutsproblems gibt[2], haben viele Autoren darauf hingewiesen, dass es strukturelle Unterschiede zwischen den beiden Fällen gibt, die normativ relevant sind.[3] Im Folgenden werde ich in einem ersten Schritt verschiedene Argumente nennen, die gegen einen Analogieschluss von dem Teichbeispiel auf das Weltarmutsproblem vorgebracht werden, und diese Argumente gegen die Möglichkeit eines solchen Analogieschlusses sind gleichzeitig als Argumente für einen nonisomorphistischen Ansatz anzusehen. In einem zweiten Schritt werde ich ein Differenzierungskriterium vorschlagen, das bislang in der Debatte noch nicht

[2] Besonders prominent ist hierbei Unger 1996.

[3] Vgl. Mieth 2012, insbesondere die Literaturhinweise auf S. 165.

vorgebracht wurde, und zwar das Kriterium der prinzipiellen Erfüllbarkeit der in Frage stehenden Hilfspflichten. Ich schließe mit einem Vorschlag bezüglich einer Dreiteilung von Hilfspflichtarten und zeige auf, inwiefern diese Differenzierung argumentative Vorteile mit sich bringen kann.[4]

Zunächst ist es möglich darauf hinzuweisen, dass die beiden Fälle sich durch den Aspekt der Nähe unterscheiden: Der potenzielle Helfer im Teichbeispiel ist dem Kind geographisch näher, als es im Fall von potenziellen Helfern mit Blick auf das Weltarmutsproblem der Fall ist.[5] Es ist allerdings fraglich, ob physische Nähe wirklich als moralisch relevanter Aspekt aufzufassen ist – dies ist alles andere als intuitiv einleuchtend (und es wurde ja auch von Singer bestritten, weswegen die bloße Postulierung von physischer Distanz als normativ relevant eine *petitio principii* dazustellen scheint). Zu zeigen wäre, dass physische Nähe selbst moralisch relevant ist und nicht etwa bestimmte Aspekte, die sich aus der physischen Nähe ergeben. Dies scheint ein schwieriges Unterfangen zu sein, denn es ist nicht klar, inwiefern physische Nähe *strukturelle* Unterschiede zwischen verschiedenen Arten von Hilfspflichten hervorbringen können soll. Physische Nähe ist ein kontingenter Faktor, der sich durch technische Möglichkeiten zudem verringern kann[6].

Eine weitere Möglichkeit besteht darin, auf einen unterschiedlichen Grad an Bestimmtheit hinzuweisen, der in den jeweiligen Fällen auftritt.[7] Während es im Fall des ertrinkenden Kindes wohlbestimmt ist, wer welche Handlung ausführen muss (der Passant muss in den Teich waten und das Kind herausziehen), gibt es im Fall der Weltarmut Unbestimmtheiten verschiedener Art: Erstens ist nicht ganz klar, wer in erster Linie zuständig sein soll, um das Problem zu lösen (Staaten? Individuen? Individuen ab einem bestimmten Einkommen?), zweitens ist auch nicht klar, wer von den möglichen Pflichtenträgern welche genauen Handlungen ausführen soll.

Eine solche Differenzierung spricht zunächst einmal gegen einen Isomorphismus im Sinne Kants oder O'Neills, sofern diese das Unbestimmtheitsproblem auf alle Arten von positiven Pflichten beziehen. Offenbar können Hilfspflichten mehr oder weniger bestimmt bzw. unbestimmt sein. Es stellt sich allerdings die Frage, ob dieses Kriterium für sich genommen in der Lage ist, strukturelle Unterschiede zwischen verschiedenen Arten von Hilfspflichten zu begründen. Die Unbestimmtheit, von der hier die Rede ist, ist in erster Linie ein epistemisches Problem, das aber prinzipiell lösbar zu sein scheint, und als solches ist es kontingent. Als Vertreter einer isomorphistischen Position könnte man darauf hinweisen, dass das Unbestimmtheitsproblem durch Arbeitsteilungsarrangements gelöst werden kann und dass es eine entsprechende moralische Pflicht für alle auch nur potenziellen Helfenden darstellt, solche Arrangements zu treffen. Strukturelle Unterschiede zwischen dem Teichbeispiel und dem Weltarmutsproblem wären dann nicht mehr aufzeigbar.

Man könnte an dieser Stelle einwenden, dass es die Zumutbarkeit ist, die die beiden angesprochenen Fälle unterscheidet. Im Teichbeispiel ist die Hilfe, die der Passant leisten muss, zumutbar; der Hilfe im Fall des Weltarmutsproblems denselben deontischen Status zuzusprechen, könnte die potenziellen Helfenden dagegen überfordern – sie müssten zu viele ihrer (zeitlichen, emotionalen und materiellen) Ressourcen einsetzen, um ihre moralischen Pflichten erfüllen zu können. Dieser Einwand mag zunächst plausibel klingen, allerdings scheint er mir nicht geeignet, Singers isomorphistische Position zurückzuweisen. Singers

[4] Zu den möglichen Differenzierungskriterien mit Blick auf diese beiden Arten von Hilfspflichten, die im Folgenden angesprochen werden, vgl. die sehr erhellende Darstellung in Mieth 2012.

[5] Vgl. zu diesem Aspekt Kamm 1999. Vgl. auch die Diskussion in Stepanians 2006.

[6] Man denke an das beliebte Gedankenexperiment der Teleskoparme: Wenn eine Person über solch lange Arme verfügen würde, dass sie ein ertrinkendes Kind aus einem Teich retten könnte, das in einem anderen Kontinent zu ertrinken droht, hätte die Person eine entsprechend starke Hilfspflicht, trotz der geographischen Distanz.

[7] Vgl. Igneski 2001 sowie die Diskussion in Mieth 2012.

Beweisziel besteht ja gerade darin zu zeigen, dass die Moral an uns höhere Ansprüche stellt, als wir es gemeinhin akzeptieren und macht einen entsprechenden revisionistischen Vorschlag; diesen mit dem Hinweis zu begegnen, dass seine Position, nach der uns die Moral viel stärker fordert als gewöhnlich angenommen, zu Überforderungseffekten führen kann, scheint mir ein Fall von *question-begging* zu sein.

Im Folgenden möchte ich ein Differenzierungskriterium vorschlagen, das sich von den hier genannten unterscheidet und das normative Unterschiede der beiden angesprochenen Fälle aufzeigen kann. Es ist meiner Auffassung nach in erster Linie die *prinzipielle Erfüllbarkeit*, die normativ relevant ist und die Unterschiede im deontischen Status von verschiedenen Hilfspflichten generiert; dabei stellt dieses Kriterium kein Fall von *question-begging* dar, da es sich durch Erwägungen begründen lässt, die von der Diskussion des deontischen Status von positiven Pflichten unabhängig sind. Das Kriterium der prinzipiellen Erfüllbarkeit sorgt für eine Differenzierung von Hilfspflichten in drei verschiedene Klassen: Es gibt demnach situative Hilfspflichten, projektbezogene Hilfspflichten und konstante Hilfspflichten, und eine Hilfspflicht hat einen abnehmenden deontischen Status, je nachdem, von welcher Klasse sie eine Instanziierung darstellt: Situative Hilfspflichten haben prinzipiell einen höheren Status als projektbezogene Hilfspflichten, und diese sind wiederum prinzipiell gewichtiger als konstante Hilfspflichten. Um diese These deutlicher zu machen, stelle ich im Folgenden zunächst die Differenzierung in drei verschiedene Arten von Hilfspflichten vor, um anschließend auf das Differenzierungskriterium einzugehen und aufzuzeigen, inwiefern es sich von bisherigen Vorschlägen in der Debatte unterscheidet.

## 4. Drei Arten von Hilfspflichten

Ich komme zunächst zur ersten Art von Hilfspflichten, den situativen Hilfspflichten. *Situative Hilfspflichten* sind dadurch gekennzeichnet, dass sie sich in einer singulären, klar bestimmbaren Situation ergeben und auch situativ behoben werden können. Das Beispiel des ertrinkenden Kindes stellt einen paradigmatischen Fall für diese Art von Hilfspflicht dar. Es gibt eine klar bestimmbare Situation, in der sich die Hilfspflicht ergibt – das Kind ist in den Teich gefallen und droht zu ertrinken. Es ist außerdem möglich, diese Hilfspflicht situativ zu erfüllen, nämlich indem der Passant in den Teich watet und das Kind heraus zieht.

Man könnte an dieser Stelle vermuten, dass es bei dieser Charakterisierung um den Aspekt der Bestimmtheit geht – die situative Einlösbarkeit der Hilfspflicht liegt ja unter anderen darin begründet, dass es eine klar bestimmbare Handlung gibt, mit der eine mögliche Retterin ihre Hilfspflicht erfüllt. Dies ist allerdings nicht das entscheidende Kriterium. Um sich dies zu verdeutlichen, muss man sich eine Situation vorstellen, in der mehrere Personen auf eine Unfallstelle mit mehreren Schwerverletzten treffen, die dringend erste Hilfe und weitere ärztliche Versorgung benötigen. In einem solchen Fall ist klar, dass verschiedene Maßnahmen nötig sind, um die benötigte Hilfe zu leisten: Die Unfallstelle muss gesichert werden, es muss Erste Hilfe geleistet werden, ein Krankenwagen muss gerufen werden etc. In diesem Fall liegt Unbestimmtheit der Aufgabenzuschreibung vor, denn es ist unbestimmt, wer von den potenziellen Helfern welche Aufgabe erledigen soll. Dies ändert allerdings nichts daran, dass es sich um eine situative Hilfspflicht handelt, denn es gibt eine zeitlich eingrenzbare Situation, die die entsprechende Hilfspflicht generiert.

*Projektbezogene Hilfspflichten* ähneln situativen Hilfspflichten darin, dass sie grundsätzlich erfüllt werden können. Sie unterscheiden sich von diesen allerdings mit Blick auf die Art und Weise, *wie* sie erfüllt werden können. Dies wird durch die Bezeichnung „projektbezogen“ zum Ausdruck gebracht. Allerdings muss schon an dieser Stelle ein Missverständnis vermieden werden: Der Ausdruck von „projektbezogenen Hilfspflichten“ ist nicht so zu verstehen, dass ihr Inhalt sich darauf richtet, anderen Personen bei der Beförderung ihrer Projekte zu helfen.

Vielmehr ist gemeint, dass sich die Hilfspflichten ihrerseits auf moralische Projekte beziehen. So wie die Projekte einer Einzelperson die Eigenschaft haben, bestimmte Einzelhandlungen dieser Person über die Zeit hinweg zu koordinieren und so in einen größeren Sinnzusammenhang zu überführen, strukturieren moralische Projekte die Einzelhandlungen verschiedener Akteure über die Zeit hinweg und überführen sie in einen größeren Sinnzusammenhang. Moralische Projekte fordern in diesem Verständnis Modelle moralischer Arbeitsteilung, um die Einzelhandlungen von Personen angemessen zu koordinieren. Dabei ist offen, ob diese Modelle eine Arbeitsteilung zwischen Personen und Personen, zwischen Personen und Institutionen oder auch zwischen Institutionen und Institutionen vorsehen – hier gibt es je nach moralischem Projekt unterschiedliche Möglichkeiten. Entscheidend ist aber, dass eine koordinierte Arbeitsteilung zwischen Akteuren stattfindet, um das entsprechende Projekt „abzuarbeiten". Dabei ist wichtig, dass die Koordination der Einzelhandlungen über einen relevanten Zeitraum hinweg geschehen muss: Das angesprochene Beispiel des Verkehrsunfall fordert ebenfalls eine Form der Arbeitsteilung zwischen Personen, allerdings ist diese Arbeitsteilung zeitlich hinreichend eingeschränkt, um eher als situative Hilfspflicht zu gelten.

Als Beispiel für eine projektbezogene Hilfspflicht kann die Bekämpfung des Welthungerproblems genannt werden. Wenn man um des Arguments willen annimmt, dass das Welthungerproblem grundsätzlich gelöst werden kann (was natürlich eine empirische Frage darstellt und an dieser Stelle nicht untersucht werden kann), dann ist es als moralisches Projekt anzusehen, weil zu erwarten ist, dass die Handlungen vieler einzelner Akteure über einen signifikanten Zeitraum hinweg koordiniert werden müssen, um es zu lösen. Die individuellen Beiträge der einzelnen Akteure (die Personen oder Institutionen sein können) müssen in diesem Sinne als projektbezogene Hilfspflichten interpretiert werden.

Hilfspflichten, die sich auf die Beseitigung des Welthungerproblems richten, sind nicht das einzige Beispiel für projektbezogene Hilfspflichten. Diverse Formen von Solidaritätspflichten lassen sich ebenfalls als projektbezogene Hilfspflichten charakterisieren. Wenn etwa diskriminierte Frauen zusammenstehen, um sexistische Praktiken zu beenden, so handelt es sich um eine konzertierte Aktion, um ein moralisches Projekt zum Erfolg zu bringen, nämlich die Beendigung der sexistischen Praxis. Ist dieses Ziel erreicht, ist das Projekt „abgearbeitet", und die entsprechende Hilfspflicht hört auf, eine zu sein – die Voraussetzung für ihr Entstehen ist beseitigt, und somit liegt das entsprechende moralische Projekt nicht mehr vor.

Die dritte Form von Hilfspflichten sind *konstante Hilfspflichten*, auch wenn hier die Bezeichnung einer Hilfspflicht unter Umständen fehl am Platz ist. Situative und projektbezogene Hilfspflichten sind dadurch gekennzeichnet, dass sie grundsätzlich erfüllt werden können, oder mit anderen Worten: Der Weltzustand, der die entsprechenden Hilfspflichten generiert, kann prinzipiell so verändert werden, dass es keinen Anlass zur entsprechenden Hilfspflicht mehr gibt – das ertrinkende Kind kann aus dem Teich gezogen werden, und der Welthunger kann beseitigt werden. Konstante Hilfspflichten unterscheiden sich in diesem Punkt sowohl von situativen als auch von projektbezogenen Hilfspflichten, denn es lässt sich nicht sinnvoll denken, dass der Weltzustand, der Anlass einer konstanten Hilfspflicht ist, so verändert wird, dass die entsprechende Hilfspflicht sich nicht mehr stellt.

Ein paradigmatisches Beispiel für diese Klasse von positiven Pflichten stellt die Förderung fremden Wohlergehens dar, also den klassischen Fall einer unvollkommenen Pflicht in der Kantischen Konzeption. Menschen sind Wesen, die auf die Hilfe anderer Personen angewiesen sind, um das eigene Wohlergehen zu erreichen – teilweise, weil andere Personen direkt zu ihrem Wohlergehen beitragen (so sind gesunde persönliche Nahbeziehungen wohl integraler Bestandteil des gelingenden Lebens von nahezu jeder Person), teilweise weil die Güter, die Personen zum Wohlergehen brauchen, nur durch die Zusammenarbeit verschiedener Personen erzeugt werden können – man denke an die Aufzählung verschiedener Güter im *Leviathan*, die allein durch Kooperation erzeugt werden können und

ohne die das menschliche Leben „einsam, armselig, widerwärtig, tierisch und kurz“ (Hobbes 1984: 96) wäre. Es lässt sich nicht sinnvoll denken, dass der Weltzustand, in dem Menschen zum Erreichen von Wohlergehen auf die Hilfe anderer Menschen angewiesen sind, umgewandelt wird in einen Weltzustand, in dem Menschen nicht mehr die Hilfe anderer Personen benötigen, um ihr eigenes Wohlergehen zu erreichen – selbst wenn ein solcher Weltzustand realisiert werden könnte, wäre er wohl kaum erstrebenswert. Da kein Weltzustand realisiert werden kann, in dem es keinen Anlass zu der Hilfspflicht mehr gibt, anderen Menschen bei der Förderung ihres Wohlergehens zu helfen, handelt es sich bei dieser Pflicht um eine konstante Hilfspflicht – sie ist konstant in dem Sinne, dass sie prinzipiell nicht ein für alle Mal erfüllt werden kann.

Das maßgebliche Differenzierungskriterium zwischen diesen drei Arten von Hilfspflichten ist, wie schon angemerkt, das Kriterium der prinzipiellen Erfüllbarkeit. Während situative und projektbezogene Hilfspflichten prinzipiell erfüllbar sind, sind konstante Hilfspflichten prinzipiell nicht erfüllbar. Gleichzeitig ist die prinzipielle Erfüllbarkeit bei situativen Hilfspflichten in einem höheren Grad gegeben als bei projektbezogenen Hilfspflichten. Dies liegt daran, dass die zeitliche Ausdehnung, die den moralischen Projekten zu eigen ist, dazu führt, dass der Erfolg bei der Einlösung der Hilfspflicht von mehr Faktoren abhängig ist als es bei situativen Hilfspflichten der Fall ist. Mehr Akteure müssen ihre Handlungen koordinieren, die Arbeitsteilungsarrangements werden komplexer, und der Handlungserfolg einzelner Akteure ist in diesem Sinne in weitaus stärkerem Maße von den Handlungen anderer Personen abhängig, als es bei situativen Hilfspflichten der Fall ist.[8] Während konstante Hilfspflichten prinzipiell nicht erfüllbar sind, sind situative Hilfspflichten prinzipiell leichter zu erfüllen als projektbezogene Hilfspflichten.

Im folgenden Abschnitt werde ich das Differenzierungskriterium der prinzipiellen Einlösbarkeit ein wenig genauer erläutern. Hierzu werde ich zunächst darstellen, inwiefern es sich von anderen Differenzierungskriterien unterscheidet, um anschließend auf einige Konsequenzen der dargestellten Unterscheidung zu sprechen zu kommen.

## 5. Das Kriterium der prinzipiellen Erfüllbarkeit

Um das Kriterium der prinzipiellen Erfüllbarkeit genauer darzustellen, ist es zunächst nötig, es von anderen Differenzierungskriterien mit Blick auf positive Pflichten abzugrenzen. Es ist hoffentlich deutlich geworden, dass es sich bei der Frage nach der prinzipiellen Erfüllbarkeit um ein strukturelles Merkmal der entsprechenden Hilfspflicht handelt und nicht um ein inhaltliches Merkmal. Es geht also nicht um die Güter, die auf dem Spiel stehen. Dies sollte nicht verwundern, denn es war ja gerade die Pointe der hier getroffenen Unterscheidung, dass es einen relevanten Unterschied der Fälle des ertrinkenden Kindes und des Weltarmutsproblems gibt, auch wenn in beiden Fällen mit dem Leben der betroffenen Personen dasselbe Gut geschützt werden soll.

Man könnte nun annehmen, dass es eigentlich um den Grad der Komplexität der jeweiligen Hilfspflicht geht. Dies ist allerdings nicht das entscheidende Kriterium. Innerhalb der einzelnen Klassen von Hilfspflichten können komplexe und weniger komplexe Instanziierungen auftreten – so ist etwa die erwähnte Hilfe beim Autounfall komplexer als die Hilfe, die das ertrinkende Kind braucht. Moralische Projekte können ebenfalls mehr oder

[8] Gleiches gilt für den Fall, dass es sich um ein moralisches Projekt handelt, das eine Person alleine erfüllen möchte und das daher keine Arbeitsteilung zwischen verschiedenen Personen verlangt. Auch in einem solchen Fall ist der Handlungserfolg abhängig von mehr Faktoren als bei situativen Hilfspflichten, die sich nur an eine Person richten: Damit das Projekt Erfolg hat, müssen die einzelnen Schritte, die zu dem Ziel führen, Erfolg haben, und hier besteht eher die Gefahr des Scheiterns, als wenn es sich nur um eine einzelne Handlung handelt.

weniger komplex sein, und gleiches gilt für konstante Hilfspflichten. Wenn eine Person zum Glücklichsein nicht mehr braucht als den freundlichen Zuspruch einer beliebigen anderen Person, allerdings immer nur für einen sehr kurzen Zeitpunkt glücklich ist, gibt es einen Zustand, der eine entsprechende Hilfspflicht generiert, wobei diese ihrerseits prinzipiell nicht erfüllbar ist – die Person braucht immer wieder neuen Zuspruch. Die Hilfspflicht ist aber nicht sonderlich komplex – alles, was nötig ist, um sie zu erfüllen, ist ein freundliches Wort. Es dürfte oft der Fall sein, dass konstante Hilfspflichten gleichzeitig auch komplex sind, aber dies ist nicht notwendigerweise der Fall, und insofern ist die Komplexität nicht das entscheidende Differenzierungskriterium.

Es ist auch nicht die Aussicht auf Erfolg, die das entscheidende Differenzierungskriterium darstellt.[9] Ob eine Hilfspflicht Aussicht auf Erfolg hat oder nicht, ist eine inhaltliche Frage. Das Kriterium der prinzipiellen Erfüllbarkeit nimmt allerdings nicht direkten Bezug auf den Inhalt der jeweiligen Hilfspflichten. Vielmehr geht es darum, ob eine konkrete Hilfspflicht als Instanziierung einer Klasse darstellt, die durch eine allgemeine bzw. prinzipielle Unterscheidung bestimmt wird. Auch hier gilt, dass es innerhalb der einzelnen Klassen von Hilfspflichten Abstufungen geben kann, was die Aussicht auf Erfolg angeht. So kann eine situative Hilfspflicht eine sehr geringe Aussicht auf Erfolg haben, eine projektbezogene Hilfspflicht dagegen eine an Sicherheit grenzende Aussicht auf Erfolg. Die Aussicht auf Erfolg ist demzufolge nicht gleichzusetzen mit dem Kriterium der prinzipiellen Erfüllbarkeit.

Betrachtet man nun diese drei Arten von Hilfspflichten, die sich aus dem Kriterium der prinzipiellen Erfüllbarkeit ergeben, lässt sich feststellen, dass die von O'Neill konstatierte Unbestimmtheit nicht auf alle Arten gleichermaßen zutrifft. Vielmehr nimmt die Unbestimmtheit mit der hier vorgestellten Reihenfolge der Klassen von Hilfspflichten ab. Situative Hilfspflichten sind in doppelter Weise wohlbestimmt, sowohl was den Träger, als auch was den Inhalt der Pflicht betrifft. Projektbezogene Hilfspflichten weisen ein geringeres Maß an Wohlbestimmtheit auf: Da projektbezogene Hilfspflichten typischerweise nach Arrangements von moralischer Arbeitsteilung verlangen, ist der Inhalt der Pflicht für einzelne Personen nicht in dem Maße kontextfrei bestimmbar, wie es bei situativen Hilfspflichten der Fall ist. Der genaue Inhalt der Pflicht, die sich an eine einzelne Person richtet, ist abhängig vom Arrangement der moralischen Arbeitsteilung und insofern unterbestimmt, und gleiches gilt für die Träger der Pflicht. Je nach Arrangement ist es denkbar, dass die eigentliche Hilfspflicht auf einige wenige Akteure „ausgelagert" wird und die übrigen Akteure nur die Pflicht haben, entsprechende Kompensationsleistungen zu treffen.[10] Der höhere Grad an Unbestimmtheit, der projektbezogenen Hilfspflichten zukommt, liegt auch dann vor, wenn die situative Hilfspflicht ebenfalls moralische Arbeitsteilungsarrangements zu ihrer Erfüllung erfordert. Situative Hilfspflichten mögen nicht in allen Fällen durch und durch wohlbestimmt sein, aber dies ändert nichts daran, dass sie grundsätzlich bestimmter sind als projektbezogene Hilfspflichten; dies liegt schon an der zeitlichen Dauer, die projektbezogene Hilfspflichten für die Erfüllung benötigen und die daher ein stärkeres Maß an Koordinierung erfordert. Konstante Hilfspflichten wiederum weisen ein hohes Maß an Unbestimmtheit auf. Dies gilt insbesondere für die Pflicht zur Förderung von fremden Wohlergehen: Da das Wohlergehen einzelner Personen immer auch eine subjektrelative Komponente hat, sind auch die Erfüllungsbedingungen der entsprechenden Hilfspflichten an die Subjektivität der einzelnen Personen gebunden und in diesem Sinne unterbestimmt.

Mit dem Grad an Unbestimmtheit nimmt auch die normative Verbindlichkeit ab. Situative Hilfspflichten haben einen hohen Verbindlichkeitsgrad, während dieser bei projektbezogenen Hilfspflichten als geringer eingestuft werden muss. Die Erfüllung von projektbezogenen Hilfspflichten ist keineswegs optional, schon alleine aus dem Grund, dass die Personen

[9] Zu diesem Kriterium vgl. Mieth 2012: 219ff.

[10] Vgl. hierzu auch Schlothfeldt 2009.

innerhalb des entsprechenden Arbeitsteilungsarrangements zum Erfüllen ihrer jeweiligen Hilfspflicht darauf angewiesen sind, dass auch die anderen Akteure ihren Teil erledigen. Allerdings kann man hier sagen, dass es „Notfalloptionen" gibt, wenn ein Akteur seinen Teil der moralischen Arbeit nicht erfüllt. Andere Akteure können einspringen und den entsprechenden Teil der Arbeit entweder ganz übernehmen oder ihn unter sich aufteilen. Konstante Hilfspflichten schliesslich sind dadurch gekennzeichnet, dass sie ein sehr geringes Maß an normativer Verbindlichkeit aufweisen. Wenn die Akteure in keinem besonderen Verhältnis zueinander stehen, wie es Freunde oder Eltern und Kinder tun, scheinen diese Formen von Hilfspflichten sogar ihren Charakter als Hilfspflichten zu verlieren. Einer völlig fremden Person beim Erreichen ihres Wohlergehens zu helfen, ist wohl in den meisten Fällen als eine supererogatorische Leistung statt als Erfüllung einer Hilfspflicht zu bewerten.

Diese abnehmende Verbindlichkeit lässt sich erklären, wenn man sich noch einmal vergegenwärtigt, was mit dem Differenzierungskriterium der prinzipiellen Erfüllbarkeit impliziert ist. Bei situativen Hilfspflichten ist die jeweilige Akteurin bzw. sind die beteiligten Akteure direkt in der Lage, den Weltzustand, der Anlass zur Hilfspflicht gibt, so umzuwandeln, dass die Hilfspflicht erfüllt ist. Dies verleiht der entsprechenden Hilfspflicht einen entsprechend hohen Grad an Verbindlichkeit. Bei projektbezogenen Hilfspflichten ist es den einzelnen Akteuren nicht im selben Maße möglich, den defizitären Weltzustand direkt zu ändern, und dies hat Einfluss auf den deontischen Status der Hilfspflicht.

Konstante Hilfspflichten lassen sich wiederum prinzipiell nicht erfüllen, weil der entsprechende Weltzustand nicht ein für alle Mal auf die benötigte Weise verändert werden kann. Zudem liegt bei konstanten Hilfspflichten ein hoher Grad an Unbestimmtheit vor, und dies kann dazu führen, dass den Akteuren relevante Informationen fehlen. Dies wird insbesondere deutlich, wenn es darum geht, das Wohlergehen einer fremden Person zu befördern: Die Akteurin, die das Wohlergehen einer ihr fremden Person steigern möchte, weiß womöglich gar nicht, worin das Wohlergehen der anderen Person besteht, was ihre Präferenzen sind, wie ihr Lebensplan aussieht etc. Gerade bei einander fremden Personen ist es also nicht nur so, dass kein Weltzustand denkbar ist, in dem Menschen nicht mehr die Hilfe anderer Menschen benötigen, um Wohlergehen zu erreichen und es daher konstante Hilfspflichten gibt, die auf dieser prinzipiellen Ebene nicht eingelöst werden können. Es lassen sich darüber hinaus Fälle denken, bei denen in praktischer Hinsicht keine konkreten Handlungen vollzogen werden können, die dem allgemeinen Ziel der Förderung fremden Wohlergehens dienen, weil hierzu relevante Informationen fehlen.

Die Korrelation von abnehmender prinzipieller Erfüllbarkeit, abnehmender prinzipieller Wohlbestimmtheit und abnehmendem deontischen Status lässt sich erklären, wenn man das Prinzip akzeptiert, dass Sollen Können impliziert. Mit abnehmender prinzipieller Erfüllbarkeit nimmt auch die prinzipielle Möglichkeit ab, die entsprechende Hilfspflicht zu erfüllen. Da also das Erfüllen-Können geringer wird, wird auch das Sollen geringer. Anders ausgedrückt: Sollen impliziert Können; Abstufungen im Können implizieren Abstufungen im Sollen. Das Prinzip „Sollen impliziert Können" stellt seinerseits ein Prinzip dar, dass in vielen Kontexten der praktischen Philosophie eine wichtige Rolle spielt, und daher ist die hier vorgestellte Differenzierung nicht *ad hoc*. Vielmehr stützt sie sich in letzter Konsequenz auf ein Prinzip, das nicht nur im Kontext der Debatte um positive Pflichten akzeptiert wird.

## 6. Mögliche Einwände

An dieser Stelle möchte ich auf mögliche Einwände eingehen, die gegen die von mir entwickelte Differenzierung erhoben werden könnten. Erstens könnte man bemängeln, dass sich die verschiedenen Formen von Hilfspflichten schon deswegen nicht unterscheiden lassen, weil sich im Kontext einer projektbezogenen Hilfspflicht oder einer konstanten

Hilfspflicht situative Hilfspflichten ergeben können. Zweitens ließe sich einwenden, dass der Grad an Verbindlichkeit von anderen Dingen abhängt als von der prinzipiellen Erfüllbarkeit. Drittens ist fraglich, ob der Begriff des Sollens überhaupt Abstufungen zulässt, wie ich es in meiner Darstellung angenommen habe, und viertens stellt sich die Frage, welchen theoretischen Nutzen die in diesem Aufsatz getroffene Unterscheidung hat.

Zunächst zum ersten Kritikpunkt. Projektbezogene wie auch konstante Hilfspflichten sind dadurch gekennzeichnet, dass sie einzelne Handlungen in einen größeren Sinnzusammenhang überführen. Dabei sind die einzelnen Handlungen aber nicht so zu verstehen, dass sie allein mit Blick auf den größeren Sinnzusammenhang als Hilfehandlungen identifizierbar sind. Dass eine Person ihre Steuern zahlt, ist beispielsweise nur dann als Hilfehandlung interpretierbar, wenn man weiß, dass die Steuern zur Bekämpfung des Welthungerproblems eingesetzt werden – hier handelt es sich um eine Handlung, die erst mit Blick auf den größeren Sinnzusammenhang in ihrem Charakter als Hilfehandlung verständlich wird. Allerdings ließe sich auch der Fall denken, dass eine Akteurin in der Lage ist, ein verhungerndes Kind direkt vor dem Verhungern zu retten, indem sie ihr Nahrung gibt, und es stellt sich die Frage, ob es sich hierbei um das Einlösen einer situativen Hilfspflicht handelt oder um eine Handlung, die sich innerhalb des moralischen Projekts der Bekämpfung des Welthungerproblems stellt. Dies scheint die hier angeführte prinzipielle Differenzierung in Frage zu stellen.

Zum diesem ersten Kritikpunkt lässt sich sagen, dass er ein allgemeineres handlungstheoretisches Problem anspricht. Handlungen können auf verschiedene Weisen beschrieben werden, und es ist nicht direkt klar, welche Handlungsbeschreibung die angemessenste ist.[11] Allerdings spricht meines Erachtens nichts von dem, was ich ausgeführt habe, gegen die Möglichkeit von situativen Hilfspflichten innerhalb von projektbezogenen oder konstanten Hilfspflichten. Es ist möglich, dass unter den Handlungen, die zur Erfüllung etwa einer konstanten Hilfspflicht ausgeübt werden müssen, auch solche sind, die sich als situative Hilfspflichten beschreiben lassen. Wenn es solche „Pflichten innerhalb von Pflichten“ gibt, gilt für sie, was für andere situative Hilfspflichten gilt: Sie haben prinzipiell einen höheren Verbindlichkeitsgrad im Vergleich zu Hilfspflichten, denen dieser situative Charakter abgeht.

Der zweite mögliche Kritikpunkt lässt sich folgendermassen darstellen: Jemandem bei einer Reifenpanne zu helfen ist eine situative Hilfspflicht, da die Hilfeleistung zeitlich eingrenzbar ist. Es wäre allerdings unplausibel, der Hilfe bei einer Reifenpanne eine höhere normative Verbindlichkeit zuzusprechen als geforderten Handlungen im Kontext der Bekämpfung von Weltarmut, nur weil die entsprechende Handlung in die Klasse der situativen Hilfspflichten fällt. Die Bekämpfung des Welthungers ist moralisch dringlicher, weil hier wichtigere Güter als im Fall einer Reifenpanne auf dem Spiel stehen, und diese gütertheoretische Betrachtung zeichnet somit bestimmte projektbezogene Hilfspflichten gegenüber situativen Hilfspflichten aus, die sich auf weniger wichtige Güter richten. Die Zugehörigkeit einer Hilfspflicht zu einer der vorgestellten Klassen sagt in dieser Perspektive nichts über ihren deontischen Status aus. Vielmehr muss der deontische Status einer Hilfspflicht mit Blick auf die in Frage stehenden Güter bestimmt werden.

Dieser Einwand ist teilweise korrekt. Natürlich wäre es absurd anzunehmen, dass nur die Zugehörigkeit zu einer strukturell bestimmten Klasse Aufschluss über den deontischen Status einer Hilfspflicht gibt und inhaltliche Erwägungen vollkommen irrelevant sind. Meine Differenzierung und die Aussagen über die entsprechende normative Verbindlichkeit sind mit einer *ceteris paribus*-Klausel versehen: Wenn es zwei Hilfspflichten gibt, die sich auf die gleichen moralisch relevanten Güter beziehen, und wenn eine dieser beiden Hilfspflichten der Klasse der situativen Hilfspflichten angehört, während die andere in die Klasse der

[11] Vgl. die Beispiele in Anscombe 2011.

projektbezogenen Hilfspflichten fällt, dann hat die situative Hilfspflicht einen höheren Verbindlichkeitsgrad, und dies alleine ist hinreichend, um isomorphistische Positionen, wie sie zu Beginn dieses Aufsatzes gekennzeichnet wurden, zurückzuweisen. Anders ausgedrückt: Die Tatsache, dass eine Hilfspflicht in die Klasse der situativen Hilfspflichten fällt, liefert einen moralisch relevanten Grund, der für die Ausführung der Handlung spricht. Damit ist allerdings nicht gesagt, dass die Zuordnung zu der entsprechenden Klasse der einzige moralisch relevante Grund ist, der herangezogen werden muss, wenn eine Akteurin überlegt, was sie tun soll. Dass durch eine Hilfehandlung bestimmte Güter geschützt werden (oder langfristig geschützt werden können), stellt ebenfalls einen moralisch relevanten Grund dar. Der Grund, der sich aus der Zugehörigkeit zu einer bestimmten Klasse von Hilfspflichten ergibt, kann durch andere moralisch relevante Gründe ausgestochen werden, sodass eine andere Handlung alles in allem ausgeführt werden soll. Sind die übrigen relevanten Gründe bei zwei Hilfspflichten allerdings gleich stark, kann die Zugehörigkeit zu einer bestimmten Klasse von Hilfspflichten den Ausschlag bei der Frage geben, was eine Person tun soll.

Ich komme damit zum dritten möglichen Kritikpunkt: Kann man überhaupt von Abstufungen im Sollen sprechen? Ein moralisches Sollen scheint einen Absolutheitsanspruch mit sich zu bringen: Entweder soll ich eine Handlung ausüben oder nicht. Von einem stärkeren oder schwächeren Sollen zu sprechen, scheint zunächst eine begriffliche Fehlverwendung darzustellen. Ist daher meine These des abnehmenden deontischen Status zum Scheitern verurteilt?

Um diese Frage zu beantworten, wäre eine ausführliche Untersuchung bezüglich des Feldes der Moral nötig, die an dieser Stelle nicht geleistet werden kann. Fest steht, dass der angesprochene Einwand mit der ethischen Hintergrundtheorie steht und fällt, die man vertritt. In einer orthodoxen Kantischen Position scheint es in der Tat problematisch, innerhalb des Bereichs des Gesollten noch Abstufungen vorzunehmen, denn entweder ist eine Handlungsmaxime widerspruchsfrei universalisierbar, oder sie ist es nicht. Vertritt man allerdings einen schon angedeuteten Ansatz, der sich bei der Bestimmung des Gesollten auf die vorliegenden moralischen Gründe stützt, scheint es mir nicht mehr unplausibel, von Abstufungen im Bereich des Gesollten zu sprechen. Gründe können stärker oder schwächer sein, und es können gleichzeitig Gründe vorliegen, die für eine Handlung sprechen und solche, die gegen dieselbe Handlung sprechen. Eine solche Konzeption würde also die Möglichkeit eröffnen, Abstufungen im Sollen vorzunehmen, und insofern gibt es eine Möglichkeit, diesen dritten Kritikpunkt zurückzuweisen.

Ich komme zum letzten antizipierten Kritikpunkt – liefert die von mir entwickelte Differenzierung einen theoretischen Vorteil? Grundsätzlich ging es mir darum zu zeigen, dass isomorphistische Positionen unplausibel sind. Dies ließe sich womöglich allerdings schon dadurch erreichen, dass man sich auf Differenzierungskriterien beruft, die bereits in die Debatte eingebracht wurden – warum sollte man ein zusätzliches Differenzierungskriterium einführen?

Hierzu ist zunächst zu sagen, dass die bisherigen Differenzierungskriterien meines Erachtens allesamt defizitär sind, um prinzipielle Unterschiede zwischen verschiedenen Klassen von positiven Pflichten zu begründen. Das Kriterium der prinzipiellen Erfüllbarkeit stützt sich in letzter Konsequenz auf das Prinzip „Sollen impliziert Können“ und ist somit unabhängig von der Debatte um positive Pflichten rechtfertigbar, da dieses Prinzip auch in anderen Kontexten der praktischen Philosophie Anwendung findet. Der Vertreter eines isomorphistischen Ansatzes kann das Differenzierungskriterium der prinzipiellen Erfüllbarkeit mithin nicht als eine *petitio principii* oder als einen *ad hoc*-Einwand zurückweisen, und dies scheint mir einen Vorteil gegenüber den anderen möglichen Differenzierungskriterien darzustellen, die in diesem Aufsatz diskutiert wurden.

Ein weiterer Vorteil der hier vorgestellten Differenzierung besteht darin, dass sie auch in anderen Kontexten der praktischen Philosophie angewendet werden kann, nicht nur mit

Blick auf die Frage nach dem deontischen Status von Pflichten mit Blick auf das Welthungerproblem und der Frage nach der Plausibilität von Singers Analogiethese. Dies wird im abschließenden Abschnitt gezeigt, wenn auch nur tentativ.

## 7. Anwendungsmöglichkeiten

Die bisherigen Überlegungen bezüglich der Arten von Hilfspflichten und ihrem jeweiligen deontischen Status, insbesondere der supererogatorische Status von konstanten Hilfspflichten, betreffen in erster Linie Fälle, in denen die beteiligten Personen in keinem besonderen Verhältnis zueinander stehen. Insbesondere die Beurteilung konstanter Hilfspflichten ändert sich, wenn die beteiligten Personen eine persönliche Nahbeziehung unterhalten. Persönliche Nahbeziehungen können den normativen Charakter von positiven Pflichten verändern: In ihrem Kontext können ansonsten supererogatorische Handlungen ihren supererogatorischen Charakter verlieren und zu geschuldeten Handlungen werden. Die Pflicht von Eltern, sich um ihre eigenen Kinder zu kümmern und ihnen beim Erreichen von Wohlergehen zu helfen, ist hier ein paradigmatisches Beispiel. Unternehmen Eltern große Anstrengungen, um fremden Kindern bzw. anderen Kindern als den eigenen beim Erreichen von Wohlergehen zu helfen, ist dies eine supererogatorische Handlung. Helfen sie dagegen ihren eigenen Kindern beim Erreichen von Wohlergehen, erfüllen sie die spezielle Pflicht, die sie als Eltern diesen speziellen Personen gegenüber haben.[12]

Dieser Wechsel des deontischen Status von supererogatorisch zu obligatorisch lässt sich meines Erachtens ebenfalls mit dem oben angesprochenen Prinzip des „Sollen impliziert Können“ erklären. Für Personen, die miteinander in persönlichen Nahbeziehungen stehen, stellt sich das Unbestimmtheitsproblem nicht in dem Maße, wie es Personen betrifft, die einander fremd sind. So kennen Eltern die besonderen Bedürfnisse ihrer Kinder besser als die besonderen Bedürfnisse von fremden Personen, und diese subjektrelativen Erfolgsbedingungen waren es ja, die zu dem besonderen Unbestimmtheitsstatus von konstanten Hilfspflichten geführt haben. Wenn es so ist, dass die reale Möglichkeit, eine Hilfspflicht zu erfüllen, Einfluss auf den Grad der Verbindlichkeit hat, dann scheint es nicht unplausibel zu sein, dass in persönlichen Nahverhältnissen, in denen die Mitglieder konstante Hilfspflichten besser erfüllen können, auch eine höhere normative Verbindlichkeit herrscht, die entsprechenden Handlungen auszuführen. Das Kriterium der prinzipiellen Erfüllbarkeit kann also erklären, warum Beziehungen deontische Statusveränderer sein können.

Dies soll nicht heissen, dass die bessere Möglichkeit, die entsprechende Hilfe zu leisten, die einzige Begründung dafür ist, dass in Nahbeziehungen Handlungen einen verpflichtenden Charakter annehmen, die ansonsten einen eher supererogatorischen Charakter haben. Ich möchte also nicht behaupten, dass spezielle Pflichten, beispielsweise parentale Pflichten, nur mit Blick auf funktionale Erwägungen zu begründen sind. Auch an dieser Stelle scheint es mir sinnvoll, die Sachlage mit Blick auf vorliegende moralische Gründe zu beschreiben. Funktionale Erwägungen können somit einen relevanten Grund darstellen, der bei der Begründung von speziellen Pflichten relevant sein kann, aber hieraus folgt noch nicht, dass solche funktionalen Erwägungen der einzige relevante Grund sind, der bei der Begründung spezieller Pflichten eine Rolle spielt.

[12] Natürlich ist es möglich, dass die Eltern sich selbst in einer Weise aufopfern, um ihren Kindern beim Erreichen von Wohlergehen zu helfen, die über das gebotene Maß hinausgeht. In einem solchen Fall könne man immer noch von supererogatorischem Verhalten mit Blick auf die eigenen Kinder sprechen. Dies spricht meines Erachtens allerdings nicht gegen die These, dass sich in persönlichen Beziehungen der deontische Status einzelner Handlungen von „supererogatorisch“ hin zu „geboten“ verändern kann.

Neben der Debatte um Hilfspflichten mit Blick auf das Welthungerproblem stellt der Bereich der speziellen Pflichten somit den zweiten Bereich dar, für den die hier entwickelte Differenzierung und die mit ihr einhergehenden Erwägungen hilfreich sein können. Das Kriterium der prinzipiellen Erfüllbarkeit ist in der Lage zu erklären, warum bestimmte Hilfspflichten ihren deontischen Status verändern können, je nachdem, ob sie im Kontext persönlicher Beziehungen auftreten oder nicht. Es lassen sich allerdings auch weitere allgemeine begründungstheoretische Einsichten mit Blick auf spezielle Pflichten aus der hier entwickelten Differenzierung gewinnen.

Um dies darzustellen, sei eine kurze Bemerkung über die generelle Begründungslast von speziellen Pflichten vorangestellt. Allgemein gesprochen handelt es sich bei speziellen Pflichten um Pflichten, die ein Akteur gegenüber einem eingeschränkten Kreis von Personen hat. Legt man eine Unterscheidung zwischen positiven und negativen Pflichten an und bestimmt die positiven Pflichten als Hilfspflichten, dann ist klar, dass spezielle Pflichten als Hilfspflichten zu bestimmen sind: Negative Pflichten – beispielsweise die Pflicht, die körperliche Unversehrtheit anderer Personen zu respektieren – haben keinen eingeschränkten Kreis von Bezugspersonen. Vielmehr hat eine Person die Pflicht, die körperliche Unversehrtheit *aller* Personen zu respektieren. Positive Pflichten können dagegen nur gegenüber einem eingeschränkten Personenkreis existieren. Um die Existenz positiver Pflichten mit eingeschränktem Personenkreis zu begründen, muss man allerdings argumentativen Aufwand betreiben, denn *prima facie* scheint es der Charakterisierung des moralischen Standpunktes als dem Standpunkt der Unparteilichkeit zu widersprechen, wenn es Pflichten gibt, die man gegenüber einigen, nicht aber gegenüber allen Personen hat.[13] Bezogen auf die Begründung von speziellen Pflichten möchte ich nun auf zwei Punkte hinweisen, die sich aus meinen Überlegungen ergeben.

Der erste Punkt besteht in der Zurückweisung der These, dass es bei speziellen Pflichten gar keine besondere Begründungslast gibt. Hier gibt es in der Literatur folgendes Argument: Wenn Eltern sich um ihre Kinder sorgen, ist dies ein Fall von wohltätigem Verhalten. Folgt man Kant, gibt es bei der Einlösung von Wohltätigkeitspflichten den schon angesprochen Spielraum an Realisierungsmöglichkeiten, und insofern ist es auch aus unparteilicher Perspektive unproblematisch, wenn die Einlösung der Wohltätigkeitspflicht an den eigenen Kindern vorgenommen wird.[14] An diesem Argument ist allerdings unplausibel, dass hiermit allenfalls die Erlaubnis nachgewiesen werden kann, sich um die eigenen Kinder zu sorgen, nicht aber eine spezielle Pflicht begründet werden kann. Akzeptiert man hingegen die These, dass Nahbeziehungen den deontischen Status von Hilfehandlungen – insbesondere von solchen, die sich auf die Förderung von Wohlergehen beziehen – von supererogatorisch hin zu geboten verändern können, ist es unplausibel, spezielle Pflichten auf die skizzierte Weise mit einem Unparteilichkeitsansatz zu versöhnen.

Der zweite wichtige Punkt besteht darin, dass sich auch in der Debatte um spezielle Pflichten eine isomorphistische Tendenz zeigt, nach der es keine relevanten Unterschiede zwischen den einzelnen Arten von speziellen Pflichten gibt und alle Formen von speziellen Pflichten auf gleiche Weise untersucht, begründet oder gerechtfertigt werden können. Wenn die von mir vorgeschlagene Charakterisierung von verschiedenen Arten von Hilfspflichten zutreffend ist, scheint mir dies allerdings ein starkes Argument darzustellen, diesbezüglich vorsichtig zu sein: Nicht alle Formen von speziellen Pflichten sind eindeutig einer (bzw. der dritten) Klasse der von mir vorgeschlagenen Unterscheidung zuzuordnen.

Spezielle Pflichten gegenüber Freunden oder Familienangehörigen fallen für gewöhnlich in die Kategorie der konstanten Hilfspflichten. Sie mögen zwar auf einzelne, klar bestimmbare

[13] Dies ist natürlich nur eine erste Charakterisierung des Problems spezieller Pflichten. Für einen ersten Überblick der Debatte vgl. Graham und LaFollette 1989 sowie Feltham und Cottingham 2010.

[14] So argumentiert etwa Baron 2008.

Handlungen bezogen sein (etwa der Hilfe beim Umzug, die sich Freunde stärker schulden als komplett Fremde), aber sie lassen sich nicht auf diese einzelnen Handlungen reduzieren, sondern stehen im größeren Zusammenhang, der anderen Person bei der Erlangung von Wohlergehen zu helfen. Auch wenn die These nicht überzeugt, dass jeder Freundschaftsdienst oder jede Handlung zugunsten von Familienangehörigen auf das Wohlergehen dieser Personen gerichtet ist, lässt sich konstatieren, dass in diesen Beziehungen kein Zustand erreicht werden kann, in dem sämtliche Formen von Hilfspflichten ein für alle Mal „abgearbeitet" sind, sodass sich die Beteiligten grundsätzlich keine besonderen Dinge mehr schulden. Insbesondere bei persönlichen Nahbeziehungen liegt es also nahe, die speziellen Pflichten, die sich innerhalb dieser Kontexte ergeben, als konstante Hilfspflichten zu bestimmen. Allerdings gibt es auch Formen von speziellen Pflichten, die nicht im angesprochenen Sinne als konstante Hilfspflichten zu bezeichnen sind. So lassen sich beispielsweise Solidaritätspflichten als Formen von speziellen Pflichten bezeichnen: Sie sind auf Solidaritätsgruppen bezogen, gelten also nur gegenüber einem bestimmten Personenkreis und sind darauf ausgerichtet, bestimmte Missstände, die Gruppenmitglieder betreffen, zu beseitigen – die schon angesprochene Solidaritätspflicht, sexistische Zustände zu beheben, sei an dieser Stelle noch einmal als Beispiel angeführt. Diese Form von speziellen (weil gruppenbezogenen) Hilfspflichten ist allerdings keineswegs als Instanziierung konstanter Hilfspflichten zu bestimmen. Sie sind darauf gerichtet, den entsprechenden negativ bewerteten Weltzustand zu beheben. Ist dieses Ziel erreicht, haben die Gruppenmitglieder keine weitergehenden speziellen Pflichten mehr gegeneinander, und insofern handelt es sich bei der Beseitigung des Weltzustandes, der Solidaritätspflichten generiert, um ein moralisches Projekt und bei den entsprechenden Solidaritätspflichten um projektbezogene Hilfspflichten.

Beachtet man, dass spezielle Pflichten unterschiedlichen Klassen angehören können, ergibt sich, dass man einem Begründungsmonismus mit Bezug auf spezielle Pflichten skeptisch gegenüber stehen sollte. Die Begründungsstrategie, die für spezielle Pflichten der dritten Klasse angemessen ist, muss nicht unbedingt diejenige sein, die für spezielle Pflichten der zweiten Klasse, also projektbezogene Hilfspflichten, einschlägig ist. Sie mag erfolgreich sein, aber dies erfordert eine substanzielle Argumentation und darf nicht einfach durch isomorphistische Analogieschlüsse unterstellt werden.

**Jörg Löschke**
Universität Bern
joerg.loeschke@philo.unibe.ch

## Literatur

Anscombe, G.E.M.: *Absicht*, Berlin: Suhrkamp.

Baron, M. 2008: „Virtue Ethics, Kantian Ethics, and the ‚One Thought Too Many' Objection", in M. Betzler (Hrg.): *Kant's Ethics of Virtue*, Berlin: Walter de Gruyter, 245–277.

Feltham, B. und J. Cottingham (Hrg.) 2010: *Partiality and Impartiality. Morality, Special Relationships and the Wider World*, Oxford: Oxford University Press.

Graham, G. und H. LaFollette (Hrg.) 1989: *Person to Person*, Philadelphia: Temple University Press.

Hobbes, T. 1984: *Leviathan oder Stoff, Form und Gewalt eines kirchlichen und bürgerlichen Staates*, Frankfurt am Main: Suhrkamp.

Igneski, V. 2001: „Distance, Determinacy and the Duty to Aid: A Reply to Kamm", *Law and Philosophy* 20, 605–621.

Kamm, F. 1999: „Famine Ethics. The Problem of Distance in Morality and Singer's Ethical Theory", in D. Jamieson (Hrg.), *Singer and His Critics*, Blackwell: Wiley, 181–186.

Mieth, C. 2012: *Positive Pflichten. Über das Verhältnis von Hilfe und Gerechtigkeit in Bezug auf das Welthungerproblem*, Berlin: Walter de Gruyter.

O'Neill, O. 2005: „The dark side of human rights", *International Affairs* 81, 427–439.

Schlothfeldt, S. 2009: *Individuelle oder gemeinsame Verpflichtung? Das Problem der Zuständigkeit bei der Behebung gravierender Übel*, Paderborn: mentis.

Singer, P. 1972: „Famine, Affluence, and Morality", *Philosophy and Public Affairs* 1, 229–243.

Stepanians, M. 2006: „Hilfspflichten und die unvollkommenen Rechte Fremder", in J. Wallacher und M. Kiefer (Hrg.): *Globalisierung und Armut – wie realistisch sind die Milleniums-Entwicklungsziele der Vereinten Nationen?* München: Kohlhammer.

Unger, P. 1996: *Living High and Letting Die: Our Illusion of Innocence*, Oxford: Oxford University Press.

# Willensschwäche – Eine Systematisierung und eine Erklärung

Christoph Lumer

Der Artikel diskutiert die klassische, von Platon definierte Form der Willensschwäche, hier „Akrasie“ genannt: 1. Man tut absichtlich *a*, 2. obwohl man glaubt, *b* zu tun sei optimal, 3. und *b* auch tun kann, 4. und der Glaube 2 ist besser begründet ist als die Absicht zu *a*. Akrasie ist eigentlich ein Problem i.w.S. kognitivistischer Handlungstheorien, deren aktuellste Version die Optimalitätsurteilstheorie ist, nach der Absichten Urteile der Art sind: ‘Die Handlung *a* ist für mich optimal’. Akrasie ist danach prima facie nicht möglich: Der Optimalitätsglaube 2 hätte zur Ausführung von *b* führen müssen. Der Artikel entwickelt eine Lösung dieses Problems für die Optimalitätsurteilstheorie, erklärt, wie Akrasie auch nach dieser Theorie möglich ist, und zeigt, daß nonkognitivistische Theorien größere Probleme durch das Phänomen der Akrasie haben. Genauer werden fünf durch Akrasie aufgeworfene Erklärungsprobleme herausgearbeitet: 1. das Versuchungsproblem, 2. das Nichtdurchsetzungsproblem, 3. das Widerspruchsproblem, 4. das Irrationalitätsproblem, 5. das Willensstärkeproblem. Anschließend werden Erklärungen zu diesen Problemen insbesondere aus der Sicht der Optimalitätsurteilstheorie entwickelt. Zum Versuchungsproblem werden beispielsweise zwei Erklärungen angeboten: zeitliche Kurzsichtigkeit (entstanden aus der größeren Komplexität langfristiger Folgen, gepaart mit Zeitdruck und unmittelbarer Präsenz kurzfristiger Folgen) und gefühlsinduzierte Veränderungen intrinsischer Bewertungen.

## 1. Einleitung: Akrasie und das Thema dieses Beitrags

Willensschwäche ist, vereinfacht gesagt, intentionales Handeln gegen das eigene bessere Urteil, d.h.: Jemand tut absichtlich *a* – d.h. er führt die Handlung *a* aus, oder er handelt absichtlich nicht –, obwohl er gleichzeitig gut begründet glaubt, sich in einer bestimmten Weise anders zu verhalten, *b* zu tun, sei besser oder gar optimal, und obwohl er *b* auch tun kann. (*a* ist im folgenden immer die willensschwache, *b* die bessere, aber nicht gewählte Handlung.) Dies ist zumindest die klassische, von Platon (Protagoras 352d, 355a-b) eingeführte Konzeption der Willensschwäche, die die Diskussion in der Tradition bestimmt hat. Wegen dieser Herkunft werde ich das so Definierte mit dem griechischen Ausdruck „Akrasie“ bezeichnen und es damit von anderem, was heute alles „Willensschwäche“ genannt wird und was auf andere Probleme verweist, klar unterscheiden;[1] ‘Willensschwäche’ sei hier also ein Oberbegriff, der Akrasie und verwandte Phänomene umfaßt. Im folgenden geht es primär um Akrasie.[2]

---

[1] Holton (1999, 2009) beispielsweise orientiert sich mit seiner Definition der „weakness of will“ als irrationale Revision einer Absicht eher an psychischen Problemen des Alltags als an den theoretischen Problemen der Handlungsphilosophie. Das in der Tradition diskutierte Phänomen nennt er wie hier „akrasia“ (Holton 2009: 72).

[2] „Willensschwäche“ bezeichnet hier also ein ganzes Feld von Verhaltensweisen, bei denen das Subjekt jeweils bessere Gründe für ein anderes als das tatsächliche Handeln oder Unterlassen hat und bei denen die Abweichung von diesen guten Gründen motiviert ist. Abgesehen von der Akrasie gibt es u.a. noch folgende Formen der Willensschwäche: *Launenhaftigkeit* (oder Wankelmut oder Unstetigkeit), bei der jemand laufend die Absicht wechselt; *Apathie*, bei das Subjekt antriebsschwach und den Handlungsfolgen gegenüber gleichgültig ist und gar keine an längerfristigen Interessen orientierte Absicht bildet; oder *Sorglosigkeit*, bei das Subjekt durchaus gesehene problematische Folgen seines Tuns oder Unterlassens abtut; oder Willensschwäche im Sinne von *Willensschwund* oder *Weichlichkeit*:

Die eingangs verwendete Charakterisierung kann man wie folgt in Form einer Definition präzisieren.

**Definition der 'Akrasie'**

Ein Subjekt *s* handelt zur Zeit *t akratisch* :=

*A1:* *absichtliches Tun: s* tut zu *t* absichtlich *a* (wobei entweder *a* oder *b* (s. A2) auch die Nullhandlung (nichts zu tun) sein kann), obwohl

*A2:* *Antagonist, kontrastierendes Optimalitätsurteil: s* unmittelbar vor *t* (zu *t*-$\varepsilon$) glaubt und dieser Glaube bewußt ist, daß jetzt sofort (zu *t*) *b* zu tun (wobei *b* mit *a* unvereinbar ist) unter den aktuell bekannten aktuell möglichen Handlungsalternativen für *s* optimal wäre, also auch besser wäre, als *a* zu tun; und

*A3:* *Möglichkeit zur besseren Handlung: s* kann zu *t b* tun, d.h., wenn *s* zu *t*-$\varepsilon$ beabsichtigt hätte, *b* zu tun, und eine angemessene synchrone Selbststeuerungsstrategie (s.u., Abschn. 8) (aus der Menge der allgemein bekannten Selbststeuerungsstrategien) eingesetzt hätte, dann hätte *s b* getan; und

*A4:* *Antagonist besser begründet:* der Optimalitätsglaube aus A2 ist besser begründet als das in A1 beschriebene Verhalten.

*Erläuterungen: Zu A2:* Bei Akrasie ist der Antagonist ein persönliches Optimalitätsurteil, nach dem also die Handlung *b* für den Handelnden selbst besser ist. Selbstverständlich gibt es auch moralische Schwäche, bei der wir gegen unsere moralischen Überzeugungen handeln; aber diese ist nicht sonderlich schwer zu erklären: Die moralischen Motive machen nur einen Teil unserer Motive aus und haben sich dann nicht gegen die anderen Motive durchsetzen können. Bei Akrasie geht es um den viel schwieriger zu erklärenden Fall, daß wir gegen ein persönliches Optimalitätsurteil handeln, das alle eigenen (rationalen) Motive aufgreift, übrigens auch die moralischen. – Das bessere Urteil muß zudem bewußt sein. Sonst liegt nicht Handeln *wider* besseres Urteil vor, sondern Handeln, das *unvereinbar* ist mit einem Urteil (Audi 1979: 176). *Zu A3:* Wenn der Handelnde die bessere Handlung *b* nicht ausführen kann, liegt keine Willensschwäche vor, sondern praktische Unfähigkeit oder Zwanghaftigkeit. Das ist unstrittig. Das „können" in der Bedingung A3 darf allerdings nicht im einfachen Mooreschen konditionalen Sinne verstanden werden (= wenn *s* die Absicht zu *b* bildet, tut *s b*). Denn nach der Optimalitätsurteilstheorie ist das Optimalitätsurteil A2 ja meist schon eine Absicht, so daß die Optimalitätsurteilstheorie mit der Mooreschen Könnensbedingung Akrasie per definitionem ausschließen würde: Wenn *s* die Handlung *b* trotz des besseren Urteils nicht getan hat, dann konnte *s b* eben nicht ausführen. Um diesen Kurzschluß zu vermeiden, ist das 'Können' hier weiter definiert: Neben der Mooreschen Antezedensbedingung, daß *s* die Absicht zu *b* hat, wird zusätzlich gefordert, daß *s* in der hypothetischen Situation auch eine angemessene Selbststeuerungsstrategie eingesetzt hätte.[3] Wenn *s* trotz

---

eine gute Absicht wird revidiert, sobald es schwierig und ernst wird; schließlich *Indolenz*, bei der eine einmal gefaßte Absicht in ein unverbindliches Urteil uminterpretiert und praktisch ignoriert wird.

[3] Die Selbststeuerungsbedingung als Spezifizierung des 'Könnens' geht auf Watson (1977: 330 f.) zurück und ist auch von Mele (1987: 93-95), Kennett (2001: 159-163) und Michael Smith (2003: 114 f.) aufgegriffen worden. Kennett kritisiert allerdings die einfache, oben verwendete Ausgangsversion und führt eine Zusatzbedingung ein: *s* hätte die erfolgreiche Handlungssteuerungsstrategie auch ausführen können (2001: 162 f.). Ihre Begründung ist: Wenn *s* die Handlungssteuerungsstrategie nicht kannte oder sie aus anderen Gründen nicht ausführen konnte, dann konnte *s* in der Situation eben doch nicht anders handeln (ibid.: 161-163). Zum einen ist Kennetts Zusatzbedingung jedoch völlig unklar. Man könnte sie mit einer weiteren konditionalen Analyse präzisieren. Zum anderen aber kann man bei allen konditionalen Analysen des 'Könnens', auch der von Kennetts Zusatzbedingung, immer einwenden, *s* hat aber das Antezedens dieser Bedingung nicht erfüllt und konnte es deshalb nicht erfüllen. Die generelle Replik der Vertreter der konditionalen Analyse auf diesen Einwand im Umfeld der Verantwortungsproblematik ist: Es gibt eine Erziehung zur Verantwortung; Menschen wissen, daß sie

Einsatzes einer solchen Selbststeuerungsstrategie immer noch nicht *b* getan hätte, dann war *s* nicht willensschwach, sondern er konnte *b* nicht tun – aus Unfähigkeit oder Zwanghaftigkeit.[4] *Zu A4:* Der Optimalitätsglaube muß besser begründet sein,[5] weil die Ausführung von *a* anderenfalls auf einem epistemischen Fortschritt des Subjekts *s* beruhen könnte (statt auf Akrasie). Manchmal ist es in der Tat besser, sich z.B. durch emotionale Widerstände gegen einen einmal gefaßten Entschluß zu einer erneuten Deliberation bewegen zu lassen und diesen Entschluß dann zu revidieren; aber dies ist eben nicht der Fall der Akrasie.

Akrasie ist ein Problem für kognitivistische empirische Handlungstheorien, wie sie insbesondere auch Platon vertreten hat und deren aktuell entwickeltste Form die Optimalitätsurteilstheorie ist. Diese besagt, daß eine Absicht zu einer Handlung *a* ein Optimalitätsurteil ist, also ein Urteil, daß diese Handlung *a* die für den Handelnden beste unter den betrachteten Handlungen ist. Nach solchen kognitivistischen Handlungstheorien ist Akrasie zumindest prima facie gar nicht möglich, denn das Urteil, daß die andere Handlung *b* besser ist, ist danach ja schon eine Absicht, die dann eigentlich automatisch ausgeführt werden sollte. Deswegen hat Platon ja die Möglichkeit der Akrasie geleugnet. Kognitivistische Handlungstheorien sind u.a. wegen des Akrasieproblems heute weitgehend in Mißkredit geraten;[6] und nonkognitivistische Theorien, insbesondere die Sui-generis-Theorie der Absicht, nach der Absichten propositionale Einstellungen eigenen Typs sind, die sich nicht z.B. auf Urteile oder Wünsche zurückführen lassen (z.B. Bratman 1987, Mele 1992), dominieren.

Im folgenden möchte ich demgegenüber zeigen, daß eine gut ausgearbeitete aktuelle Version der kognitivistischen Handlungstheorie, nämlich eine Optimalitätsurteilstheorie der Absicht, die durch die Akrasie aufgeworfenen Probleme lösen kann und daß die konkurrierenden nonkognitivistischen Theorien ebenfalls Probleme mit der Akrasie und zusätzlich noch viel grundlegendere Probleme haben.

---

unter den und den Bedingungen zur Verantwortung gezogen werden; sie haben bis zur vollen Verantwortlichkeit im Erwachsenenalter eine Menge Zeit, zu lernen, sich die Einhaltung der Bedingungen anzueignen, die nachher von ihnen eingefordert werden – in diesem Fall also: den Einsatz von Selbststeuerungsstrategien im Falle von akratischen Versuchungen –; und die Gesellschaft unterstützt sie normalerweise bei diesem Lernprozeß. Eine Ursache für Kennetts m.E. falsche Bedingung ist, daß sie sich zu sehr vom Problem der moralischen Verantwortlichkeit leiten läßt und voraussetzt, daß bei Willensschwäche auf jeden Fall moralische Verantwortlichkeit vorliegt. Tatsächlich erkennen wir aber z.T. auch nach der Zuschreibung von Willensschwäche entlastende Umstände an, die die Verantwortlichkeit einschränkten, etwa bestimmte Formen besonders schwieriger Sozialisation. Nach der hier vertretenen einfacheren Bedingung A3 ist für Akrasie nicht einmal erforderlich, daß der Akratiker die angemessene Selbststeuerungsstrategie kennt, zumindest ist nicht erforderlich, daß er deren Einsatz beherrscht, und schon gar nicht, daß auch alle sonstigen Bedingungen für den Einsatz dieser Selbststeuerungsstrategie erfüllt waren. Willensstärke ist z.T. einfach das Beherrschen einer bestimmten Technik; wer nicht in der Lage ist, in Versuchungssituationen bestimmte Selbststeuerungstechniken einzusetzen, ist eben in der fraglichen Situation akratisch – egal ob er besonders verweichlicht erzogen wurde oder gerade depressiv ist. Moralische und Verantwortungsfragen stellen sich später.

[4] Bei Zwanghaftigkeit hilft nicht einmal der Einsatz von synchronen Selbststeuerungsstrategien; dies scheint eine klare Abgrenzung zur Akrasie zu sein. De facto ist diese Grenze aber ziemlich fließend; sie ist nicht klar wie bei physischem Zwang. Vielfach können Menschen, die wir als zwanghaft einschätzen, bei geeigneten Belohnungsversprechen oder Strafandrohungen diesen Zwang ohne weiteres überwinden (Elster 1999: 166-168, Watson 1999: 63; s.a. Kant, KpV A 54). Zwanghafte können, müssen aber übrigens nicht glauben, daß eine alternative Handlung *b* besser wäre; die Bedingung A2 gehört also nicht zum Definiens von ‘Zwanghaftigkeit’.

[5] Die Kriterien für gut begründete persönliche Optimalitätsurteile können hier nicht thematisiert werden; s. aber: Lumer 2009, 241-427; 521-548.

[6] Kritiken der Optimalitätsurteilstheorie der Absicht, weil sie das Akrasieproblem nicht lösen könne, z.B.: Bratman 1979, 160; McCann 1998, 223-225; Mele 1992, 122 f.; 228 f.; 233 f.; Moore <1993> 2010, 149.

## 2. Eine subjektivistische kognitivistische empirische Handlungstheorie – die Optimalitätsurteilstheorie

Vorausgesetzt sei im folgenden die *analytische* Festlegung, daß Absichten u.a. exekutive und volitive Zustände sind, die die Einstellungen des Subjekts zu der beabsichtigten Handlung zusammenfassen und unter Standardbedingung die Ausführung der Absicht verursachen. Die Optimalitätsurteilsthese ist dann eine *empirische* Hypothese über Menschen, die ungefähr besagt: Ein Mensch hat eine Absicht zu einer Handlung *a*, gdw. er glaubt, *a* sei unter den in Betracht gezogenen ihm möglichen Handlungen die für ihn beste (Details s.: Lumer 2005, 2009: 128-240). Diese empirische handlungspsychologische These versucht zu beschreiben, worin menschliche Handlungsentscheidungen und Absichten de facto bestehen; *analytisch* spricht natürlich nichts dagegen, daß Absichten aus ganz anderen Arten von Einstellungen bestehen, z.B. Bratmanschen Sui-generis-Absichten oder auch aus normativen Urteilen über die Handlung. Die Optimalitätsurteilsthese ist nur der Kern einer umfassenderen empirischen Handlungstheorie, deren weitere Teile u.a. klären, was die genaue Bedeutung solcher Optimalitätsurteile ist, wie sie aus intrinsischen Werturteilen gebildet werden, was der Inhalt intrinsischer Werturteile ist usw. (s. Lumer 2009: 128-240, 428-521; 2007). Für die Diskussion der Akrasie ist davon in der von mir entwickelten Version der Optimalitätsurteilstheorie insbesondere die Theorie der intrinsischen Wünsche interessant (Lumer 1997, 2007: 167-173, 2009: 428-521; 2012). Danach gibt es zwei grundverschiedene Arten intrinsischer Wünsche – oder genauer: mehr oder weniger angeborene Kriterien für unsere intrinsischen Bewertungen. Zum einen gibt es Kriterien für zeitlich stabile Wünsche mit im wesentlichen hedonischem Inhalt. Zeitliche Stabilität bedeutet dabei: Dasselbe Objekt, also dasselbe angenehme oder unangenehme Gefühl, wird (bei unveränderter Informationslage) zu unterschiedlichen Zeiten gleich bewertet. Daß ich morgen Hunger habe, bewerte ich heute und morgen gleich. Weil dies so ist, kann ich schon heute gegen den morgigen Hunger vorsorgen, indem ich heute z.B. Lebensmittel einkaufe. Zeitlich stabile Bewertungen sind deshalb die Basis rationaler Planung und damit auch rationaler Bewertungen überhaupt. Wegen dieser Rationalität liefern sie auch die höhere Begründungsbasis für Gesamtbewertungen von Handlungen. Zum anderen gibt es daneben aber Kriterien für gefühlsabhängige intrinsische Bewertungen: Gefühle, vor allem Emotionen und Körpergefühle induzieren von diesen Gefühlen abhängige und für diese Gefühle spezifische intrinsische Bewertungen. In Wut entsteht der spezifische intrinsische Wunsch, das Objekt unserer Wut zu bestrafen; bei Furcht entsteht der spezifische intrinsische Wunsch, der gefürchteten Gefahr zu entfliehen; durch Schmerzen wird ein zusätzlicher intrinsischer Wunsch nach Beendigung der Schmerzen induziert; durch angenehme Geschmackserlebnisse oder sexuelle Erregung wird ein zusätzlicher Wunsch nach verstärktem und verlängertem Erleben dieses Geschmacks bzw. nach verstärkter Erregung und Entspannung induziert. Die ersten Beispiele zeigen, daß emotionsinduzierte intrinsische Wünsche keinen hedonischen Inhalt haben. Wie an den letzten Beispielen zu sehen ist, induzieren die körperlichen Gefühle meist nur Verstärkungen ohnehin schon vorhandener hedonischer Wünsche; diese werden gegenüber der normalen Bewertung übertrieben. Ein Charakteristikum gefühlsinduzierter Wünsche ist, daß sie mit dem Gefühl entstehen und vergehen. Sie eignen sich deshalb nicht als Basis langfristiger Planung, und die auf ihnen beruhenden Handlungen werden oft als irrational, eben z.B. als Affekthandlung angesehen; Handelnde bereuen sie oft nachher – eben weil sie sie nun, auf einer anderen intrinsischen Wertbasis, niedriger bewerten.

Die gerade skizzierte empirische Handlungstheorie ist *kognitivistisch* in dem Sinne, daß sie besagt, daß die Absichten und die sie begründenden Einstellungen kognitiv sind, also Meinungen, in denen Propositionen als wahr oder falsch beurteilt werden. Die Absichten sind Optimalitätsurteile; ihnen liegen Meinungen über den Wert der Handlungen, Meinungen über die Folgen solcher Handlungen und über den intrinsischen Wert solcher Folgen zugrunde. Dieser kognitive Inhalt schließt aber überhaupt nicht aus, daß die

Optimalitätsurteile Absichten sind, daß sie also u.a. eine exekutive Funktion haben, daß sie folglich unter Standardbedingungen die beabsichtigte Handlung verursachen. Andererseits ist die hier vorgestellte Form von Kognitivismus *subjektivistisch*, d.h. an dieser Stelle, die Kriterien für die Bewertung von Gegenständen liegen nicht in den Gegenständen selbst (Wertrealismus), sondern entstammen dem Subjekt; sie sind Teil seines Motivationssystems. Die Motive stellen sich lediglich in der Form von Meinungen dar. Die gerade skizzierte Form von Subjektivismus ist antirealistisch, sie ist aber nicht antikognitivistisch: Die Optimalitätsurteile und Wertmeinungen sind nicht per se korrekt oder unkritisierbar, sondern können wahr oder falsch und gut oder schlecht begründet sein: Unsere Entscheidungspsychologie gibt nur die *Kriterien* für die Bewertung und Entscheidung vor – und auch dies nur z.T., aber diese Komplikation kann hier ignoriert werden –; ob diese Kriterien erfüllt sind, müssen wir erst erkennen; und dabei können wir uns irren.

## 3. Die sich aus der Akrasie ergebenden Erklärungsprobleme für Handlungstheorien

Gegen die Optimalitätsurteilstheorie sind viele Einwände erhoben worden, insbesondere auch Einwände aus der Möglichkeit der Akrasie (s. Anm. 6). Eine spezifizierte Form des Akrasie-Einwandes gegen die Optimalitätsurteilsthese ist: Wenn Optimalitätsurteile Absichten sind und der Verstoß dagegen intentional ist, dann gibt es zwei sich diametral widersprechende Absichten; das wäre Geistesgestörtheit, aber nicht Willensschwäche (McCann 1998: 224 f.). Es gibt aber noch ein paar andere Erklärungsprobleme durch die Akrasie, nicht nur für die Optimalitätsurteilstheorie. Ich möchte diese Fragen hier systematischer angehen und diese Erklärungsprobleme sammeln und untersuchen, ob und wie die diversen Theorien, vor allem die Optimalitätsurteilstheorie, sie lösen können. McCanns Einwand wird dabei in drei Teilprobleme (die ersten drei der folgenden Liste) ausdifferenziert.

Ich sehe für Handlungstheorien mindestens fünf Erklärungsaufgaben, die sich aus der Akrasie ergeben:

*Erklärungsaufgabe 1: das Versuchungsproblem: Wieso Absicht zu a?:* Wie kommt der Handelnde überhaupt auf die Idee, *a* (die ausgeführte Alternative) tun zu wollen? Und wie kann diese Idee zu einer Absicht werden? Der Handelnde hatte ja immerhin schon das besser begründete Urteil gebildet, daß die alternative Handlung *b* besser sei. – Dieses Problem betrifft alle Handlungstheorien, die kognitivistische aber besonders: Warum werden nicht einfach kognitiv Einstellungen gebildet, ohne inneren Konflikt – wenn das Entstehen von Absichten und Optimalitätsurteilen doch eine kognitive Angelegenheit ist?

*Erklärungsaufgabe 2: das Nichtdurchsetzungsproblem: Wieso setzt sich das Optimalitätsurteil nicht durch?:* Wenn ein Optimalitätsurteil eine Absicht ist (oder, in Meles (1992: 228 f., 230-232) Theorie, normalerweise automatisch zu einer Absicht führt), wieso ist das Optimalitätsurteil bei Akrasie nicht handlungswirksam? Wenn Optimalitätsurteile schon Absichten sind, dann bestehen bei Akrasie zwei konkurrierende Absichten. Eine der Absichten müssen wir aufgeben, oder sie wird irgendwie übertrumpft. Wieso wird bei Akrasie die besser begründete Absicht besiegt? – Wenn Optimalitätsurteile keine Absichten sind – wie (hinsichtlich der Absichten) nonkognitivistische Theorien behaupten –, dann besteht an dieser Stelle kein so großes Problem: Ein Optimalitätsurteil ist per se praktisch weniger effektiv als eine Absicht, es hat keine unmittelbaren Handlungswirkungen. Hier besteht also für die konkurrierenden Theorien eine Asymmetrie: Für nonkognitivistische Theorien, die also annehmen, daß Absichten etwas anderes als Urteile, insbesondere als Optimalitäts-

urteile sind, ist die Erklärung der Nichtdurchsetzung, in sportlicher Terminologie, sozusagen ein Heimspiel, für kognitivistische hingegen ist sie ein schwieriges Auswärtsspiel.

*Erklärungsaufgabe 3: das Widerspruchsproblem: Wie können wir prima facie inkohärente Einstellungen haben?:* Bei Akrasie steht das faktische Tun prima facie in einem noch näher zu klärenden Widerspruch zum Optimalitätsurteil: Wie können wir überhaupt solche inkohärenten Einstellungen haben? – Für *non*kognitivistische Theorien liegt bei Akrasie eine relativ mittelbare Inkohärenz vor: Man hat einerseits rationale Einsichten, andererseits irrationale Absichten; es liegt eine Inkohärenz zwischen einer Art kognitivem, etwas ideologischem Überbau und der volitiven Basis vor, die sich im Zweifelsfall wenig um diesen Überbau schert. Nach der Optimalitätsurteilsurteilstheorie hingegen impliziert auch das absichtliche *a*-Tun ein Optimalitätsurteil, daß nämlich *a* optimal ist und damit besser als *b* – im Gegensatz zum antagonistischen Optimalitätsurteil, nach dem *b* optimal ist und damit besser als *a*. Dies ist ein ziemlich unmittelbarer doxastischer Widerspruch.

*Erklärungsaufgabe 4: das Irrationalitätsproblem: Warum dominiert die schlechter begründete Einstellung?:* Menschen tendieren weitgehend zu rationalen Einstellungen und zur rationalen Lösung von Konflikten zwischen Einstellungen. Warum dominiert bei Willensschwäche irrationalerweise die schlechter begründete Einstellung?

*Erklärungsaufgabe 5: das Willensstärkeproblem: Wie hätte der Akratiker die bessere Handlung b doch ausführen können; und wieso handeln wir meistens enkratisch?:* Der Akratiker hat die besser begründet für optimal gehaltene Handlung *b* nicht ausgeführt. Aber akratische Handlungen sind nicht im strengen Sinne zwanghaft; dies besagt auch die Bedingung A3 der Definition der ‘Akrasie’: *s* hätte die bessere Handlung ausführen können. Zu erklären bleibt dann, wie dies dem Akratiker möglich gewesen wäre. Dieses Problem kann man auch ausdehnen: Wieso handeln wir überhaupt meistens doch willensstark, also enkratisch? – Zumindest die letzte Frage kann die Optimalitätsurteilstheorie sehr einfach beantworten: Das gut begründete Optimalitätsurteil zugunsten von *b* ist ja schon eine Absicht; eventuelle Zweifel daran zugunsten von *a* werden kognitiv ausgeräumt; und dann führt die Absicht mehr oder weniger automatisch zur Ausführung der Handlung *b*. Der Nonkognitivismus hat hier viel größere Schwierigkeiten, die Willensstärke und Rationalität unserer Handlungen zu erklären; die Absicht ist nach diesen Theorien nämlich völlig vom eventuellen Optimalitätsurteil abgekoppelt. Im Sportjargon: Für die Optimalitätsurteilstheorie ist die Erklärung des enkratischen Normalfalls ein Heimspiel, für den Nonkognitivismus hingegen ein Auswärtsspiel. Und damit sei gleich eine theoriestrategische Einschätzung verbunden: Mir ist es lieber, wie bei der Optimalitätsurteilstheorie, für den *Normal*fall keine Erklärungsprobleme zu haben (also ein Heimspiel) und beim Ausnahmefall größere Erklärungsschwierigkeiten (also ein Auswärtsspiel), als umgekehrt, wie beim Nonkognitivismus.

Für diese fünf Erklärungsaufgaben werden im Rest des Artikels Lösungsvorschläge entwickelt.

## 4. Lösungsansätze zum Versuchungsproblem

Die Frage des Versuchungsproblems ist: Warum gibt es überhaupt eine Art Drang weg vom besser begründeten Urteil? Offensichtlich gibt es hier Konflikte, die nicht rein kognitiv sind. Die Erklärungen des Versuchungsproblems, die ich hier anbieten möchte, beruhen auf zwei Mechanismen, die z.T. alleine wirken, häufig aber in Kombination.

1. Der erste Mechanismus ist die *zeitliche Kurzsichtigkeit*: Der Willensschwache achtet nur oder vorwiegend auf die kurzfristigen Folgen; die langfristigen werden ignoriert, abgetan oder verdrängt; und die Konfliktlage bei Willensschwäche ist dann meist so, daß die kurzfristigen Folgen der willensschwachen Handlung *a* besser sind als die der willensstarken Handlung *b*, während sich dieses Verhältnis bei Berücksichtigung der langfristigen Folgen umkehrt. Der Akratiker hat z.B. keine Lust, morgens aufzustehen, sich an die Arbeit zu begeben, hat Angst, vom Fünf-Meter-Brett zu springen, weil all dies kurzfristig unangenehm ist, obwohl das frühere Aufstehen, die Arbeit, der Sprung langfristig besser sind.[7] Zeitliche Kurzsichtigkeit erklärt Akrasie zwar zum Teil, aber nicht vollständig. Denn der Akratiker *weiß* ja eigentlich, daß *b* besser ist; er hat *b* ja schon vorher bewertet. Warum behält er dieses besser begründete Urteil nicht bei?

Der Mechanismus der zeitlichen Kurzsichtigkeit ist also etwas komplexer, kann aber durch das Zusammenwirken zweier Faktoren erklärt werden: 1.1. Zum einen sind die langfristigen Folgen kognitiv nur sehr viel schwieriger zu berücksichtigen; wenn es dann kurzfristig und womöglich unter Zeitdruck oder emotionalem Einfluß zu einer Neubewertung kommt, dann werden sie entsprechend schneller vernachlässigt. 1.2. Zum anderen sind die kurzfristigen Folgen, wenn die Handlungsausführung ansteht, unmittelbar präsent, die langfristigen hingegen nicht.

Ad 1.1: Daß die langfristigen Folgen kognitiv schwieriger zu berücksichtigen sind, beruht auf mehreren Gründen, die hier nur kurz erwähnt werden können. Zum einen sind langfristige Folgen in der Regel nicht unmittelbar sinnlich präsent, sondern abstrakt und komplex; gerade für Kinder und Jugendliche beruhen ihre langfristigen Folgenannahmen häufig nur auf Autoritätsargumenten, die sie von Erwachsenen übernommen haben und deren inhaltliche Korrektheit sie überhaupt nicht beurteilen können, schon weil sie die beteiligten Mechanismen – z.B. Selektionsmechanismen bei der schulischen, universitären und beruflichen Laufbahn, die langfristige Bedeutung bestimmter persönlicher Beziehungen – nicht durchschauen oder nicht einmal kennen. Selbst in Mischels (<1989> 1992) berühmten Experimenten ist ein Versprechen, bei kurzzeitigem Verzicht auf die sinnlich präsenten Marshmallows die doppelte Mengen von Marshmallows zu erhalten, ziemlich abstrakt: Die Kinder müssen sich auf die Institution des Versprechens und die Vertrauenswürdigkeit des Experimentators verlassen. Zum anderen sind die langfristigen Folgen darüber hinaus oft intellektuell im folgenden Sinne: Sie bleiben abstrakt; sie sind nur wahrscheinlich; man weiß nicht, welche Folgen oder ungeahnten Möglichkeiten sonst noch so eintreten werden, wodurch der ganze Kalkül hinfällig würde; es handelt sich nicht um *eine* konkrete Folge, sondern um riesige Bündel alternativer, aber ähnlicher Folgen mit jeweils minimalen Wahrscheinlichkeiten, die aber eben zu dann, weil nie im einzelnen vorgestellt, relativ abstrakten Bündeln zusammengefaßt werden; die Gefahr beispielsweise, die von schlecht gewarteten Bremsen eines Autos ausgeht, besteht in einer Unzahl von jeweils sehr unwahrscheinlichen einzelnen Unfallmöglichkeiten.

Zum anderen werden langfristige Folgen korrekterweise probabilistisch diskontiert. Wenn die intrinsisch relevanten Folgen darin bestehen, daß man etwas erlebt, also vor allem in angenehmen oder unangenehmen Gefühlen, sinkt schon ihre Erlebniswahrscheinlichkeit überproportional mit dem zeitlichen Abstand. Daneben gibt es noch andere Faktoren, wegen derer die Wahrscheinlichkeit bestimmter langfristiger Folgen sinkt: Das ganze (politische,

[7] Häufig sagt man dann, der Akratiker entschiede in solchen Fällen nur nach dem Lustprinzip. Tatsächlich besteht der Konflikt aber nicht zwischen hedonischen und nichthedonischen Interessen, denn auch bei den langfristigen Folgen geht es meist um hedonische Interessen: in den Beispielen etwa, sehr verkürzt: Sinnvolles mit der gewonnenen Zeit anzustellen, das einem nachher Befriedigung verschafft; Geld zu verdienen, mit dem man sich eine noch größere Lust verschaffen kann; den Rausch des Falls zu erleben, Stolz auf den Sprung zu empfinden, die Bewunderung der anderen zu genießen. Der Konflikt besteht also zwischen der Berücksichtigung kurz- und langfristiger Folgen.

juristische, physische, soziale) System, von dem die Folgen abhängen, wird zerstört; Personen, die als Garanten, Übermittler funktionieren, sind verschwunden, tot; man selbst oder die Gesellschaft findet später Mittel, bestimmte negative Folgen zu verhindern; usw. Rational erforderliche zeitabhängige probabilistische Diskontierungen sind bei wirklich langfristigen Entscheidungen (über mehrere Jahre hinweg) für ein normales Individuum überkomplex; man kann sie nicht wirklich berechnen. Die meisten Menschen nehmen deshalb – rational gesehen: als Ersatz – intuitiv eine diffuse, holistische zeitliche Diskontierung vor, mit der alle möglichen Absenkungen langfristiger Wahrscheinlichkeiten abgedeckt sein sollten. Die Diffusität und der Holismus öffnen kognitiven Verzerrungen Tür und Tor. Bei vielen Menschen ist die zeitliche Diskontierung zu groß, oftmals, vor allem bei jungen Menschen, viel zu groß.

Ad 1.2: Der andere Faktor, der zur zeitlichen Kurzsichtigkeit beiträgt, ist, wie gesagt, die Tatsache, daß die kurzfristigen Folgen, wenn die Handlungsausführung ansteht, unmittelbar präsent sind. Dies kann dann durch eine Art Rückkoppelung von den langfristigen Folgen ablenken: Das Schwelgen in den kurzfristig angenehmen Folgen ist selbst angenehm; dadurch wird dann die Betrachtung der langfristig unangenehmen Folgen abgewürgt. Ein Horror vor unangenehmen kurzfristigen Folgen kann sich zu einer Art panikartigen Obsession ausweiten, die ebenfalls die langfristigen Folgen aus dem Blick geraten läßt – man denke etwa an einen Studenten, der bei sich selbst Blut abnehmen muß (Mele 1987: 34 f.). Beides, das Schwelgen in angenehmen Gefühlen wie die Panik vor den unangenehmen negativen Folgen, kann zudem passende intrinsische Wünsche induzieren, den intrinsischen Wunsch nach Befriedigung bzw. nach Beseitigung der Gefahr, die also wieder die Handlung mit den kurzfristig besseren Folgen begünstigen. Nun könnte man ja auch in langfristigen angenehmen Folgen schwelgen oder obsessiv an die langfristig negativen Folgen denken und dadurch in Situationen drohender Willensschwäche genau in die rationaliter richtige Richtung gelenkt werden. Tatsächlich besteht hier aber schon aus den eben aufgelisteten kognitiven Gründen (1.1) keine Symmetrie. Die wirklich langfristigen Folgen sind zu vielfältig, zu viele und zu unwahrscheinlich, als daß sich hier einzelne von ihnen zum Ausmalen und emotionalen Erleben anbieten würden.

2. Der zweite Mechanismus, der uns in Versuchungen zur Willensschwäche führt, allerdings in Versuchungen ganz anderer Art, sind die gefühlsinduzierten Veränderungen intrinsischer Bewertungen. Oben (Abschnitt 2) hatte ich schon kurz den Mechanismus gefühlsinduzierter intrinsischer Wünsche erklärt. Dieser führt nun zum einen über den soeben angesprochenen Mechanismus des Schwelgens in angenehmen Folgen und der obsessiven Betrachtung der negativen Folgen zu einer übertriebenen Verstärkung der entsprechenden *hedonischen* Bewertung: Diese Bewertung ist extremer, als dies dem hedonischen Wert des Objekts entspricht. Dies ist u.a. bei typischem Suchtverhalten der Fall: Eß-, Trunk-, Sex-, Spiel-, Shopping-, Rauschgiftsucht. Zum anderen werden durch die gefühlsinduzierten Wünsche impulsive Unbeherrschtheiten ausgelöst, sozial am bedeutsamsten das auf Wut beruhende aggressive Verhalten: Die Wut induziert den intrinsischen Wunsch, das Objekt der Wut zu bestrafen, also einen Wunsch mit *nichthedonischem* Inhalt. Der irrationale gefühlsinduzierte Wunsch kann sich gegen die vorgängige rationale Absicht durchsetzen, weil er den schon in die damalige Absichtsbildung für *b* (und gegen *a*) eingeflossenen Wünschen einen neuen Wunsch hinzufügt, einen Wunsch nach *a*, was dann zur Umkehr der Präferenz führt. Die alte Absicht ist möglicherweise sogar noch bewußt, sie ist aber als solche entwertet, hat nur noch den Status einer Erinnerung an ein Werturteil. Um den kognitiven Widerspruch zu glätten, wird dieses Werturteil dann oft wenigstens für die aktuelle Situation kognitiv entwertet, also der Triumph der irrationalen Absicht rationalisiert: Der Vorsatz war nicht für derartig krasse Situationen wie die aktuelle gemacht, es liegt eine Ausnahmesituation vor usw.

## 5. Lösungsansätze zum Nichtdurchsetzungsproblem

Das Nichtdurchsetzungsproblem ist: Warum hat sich das besser begründete Urteil, daß *b* die bessere Alternative ist, nicht gegen die willensschwache Absicht durchgesetzt? Das Nichtdurchsetzungsproblem ist in der Tradition das zentrale Problem, zu dem alle Theorien einen Lösungsvorschlag machen. Entsprechend kann man diese Theorien danach einteilen, welche Antwort sie auf das Nichtdurchsetzungsproblem geben; allerdings lassen viele Autoren mehrere, sich nicht gegenseitig ausschließende Antworten auf das Nichtdurchsetzungsproblem zu. Diese Antworten können hier nicht alle besprochen werden. Tabelle 1 enthält lediglich eine Übersicht der Antworten. Die meisten von ihnen erklären nur Willensschwäche i.w.S., nicht aber Akrasie; sie nehmen z.B. an, daß das Optimalitätsurteil für *b* in letzter Sekunde doch noch revidiert wird (Fall 2.6 der Liste). Sodann gibt es eine Reihe von Erklärungen, nach denen zwar ein Optimalitätsurteil vorliegt, dieses aber gemäß der Optimalitätsurteilstheorie nicht die für eine Absicht erforderliche Form hat (Untergruppe 3 der Liste); solche Fälle sind mit der Optimalitätsurteilstheorie natürlich sehr leicht zu erklären. Ich möchte hier nur meine Erklärungen für die schwersten Formen von Akrasie darlegen, bei denen 1. der Antagonist ein Optimalitätsurteil mit (gemäß der Optimalitätsurteilstheorie) Absichtscharakter ist – dies sind die Fälle 4 bis 6 in der Liste – und 2. in denen der Akratiker mehr oder weniger sehenden Auges gegen das Optimalitätsurteil verstößt und absichtlich *a* tut; dies sind die Unterfälle h und i.

***Tabelle 1: Überblick über die Typen der Willensschwäche:***

*Gruppe I: Unmöglichkeit der willensstarken Handlung:*

1. Unmöglichkeit: *s* kann nicht *b* tun. (Bedingung A3 nicht erfüllt.)

*Gruppe II: Der Antagonist ist kein Optimalitätsurteil mit Absichtscharakter:*

2. Urteilslosigkeit: *s* hat gar kein besseres Urteil gefällt, das einer Absicht ähnelt, obwohl ein solches Urteil offensichtlich ist. (Bedingung A2 (Optimalitätsurteil) nicht erfüllt.)
   - 2.1. Gefühlsinduziertes Handeln gegen gute Argumente (motivierte Irrationalität), Schlußfolgerung nicht gezogen / auch Affekthandeln.
   - 2.2. Durch unbewußte Motive (mit-)verursachtes Handeln gegen bessere Gründe (motivierte Irrationalität).
   - 2.3. Bloß verbale Bekundungen des Optimalitätsurteils, aber kein Glaube.
   - 2.4. Optimalitätsurteil mit irrealer Bedingung.
   - 2.5. Apathie, Initiativlosigkeit: Gleichgültigkeit gegenüber langfristigen Folgen.
   - 2.6. (Kurzfristige) epistemisch irrationale Verwerfung des Optimalitätsurteils.
   - 2.7. Optimalitätsurteil mit Selbsttäuschung: *s* täuscht sich darüber (aus Interesse an einem besseren Selbstbild), daß er an das Optimalitätsurteil glaubt.

3. Absichtslosigkeit, nichtvoluntative Schwäche: *s* hat ein Optimalitätsurteil gefällt, aber nicht das für eine Absicht erforderliche Optimalitätsurteil. (Optimalitätsurteil vorhanden (manchmal Bedingung A2) erfüllt, aber das Urteil ist keine Absicht.)
   - 3.1. Moralische Schwäche: Das Optimalitätsurteil war nur ein moralisches. (A2 nicht erfüllt.)
   - 3.2. Epistemische Irrationalität: *s* hat nur ein bedingtes Optimalitätsurteil gefällt und hält die Bedingungen für erfüllt. (A2 nicht erfüllt.)
   - 3.3. Vage Vornahme: *s* hat eine „Vornahme“ ohne klare Ausführungsbedingungen gebildet – z.B.: ‘Es wäre optimal, demnächst / wenn ich mich stärker fühle, mit dem Rauchen aufzuhören.’ (Eventuell Akrasie.)
   - 3.4. Melioritätsurteil: *s* hält eine andere Handlung, *b*, für besser, ohne sie für optimal zu halten. (A2 nicht erfüllt.)
   - 3.5. Aspektbeschränktes Optimalitätsurteil: *s* hält *b* für in einer Hinsicht optimal, ohne *b* für insgesamt optimal zu halten. (A2 nicht erfüllt.)

3.6. Intellektualistisches Optimalitätsurteil / originäre Indolenz / Weichlichkeit (Handeln wider besseres Wissen): Das Optimalitätsurteil wird nicht zur Absicht, es erlangt keine motivierende Kraft; Gleichgültigkeit gegenüber langfristigen Folgen: 'Es wäre irgendwie besser, *b* zu tun'. – *s* glaubt nicht so recht an das Optimalitätsurteil. *s* hat gelernt, daß das Optimalitätsurteil richtig ist, es gibt Argumente dafür; *s* hat aber diffuse Einwände, die das Optimalitätsurteil nicht widerlegen, sondern es maximal in Zweifel ziehen können. (Akrasie.)

*Gruppe III: Der Antagonist ist ein Optimalitätsurteil mit Absichtscharakter, er wird aber übertrumpft:*

4. Das Optimalitätsurteil ist eine Grobvornahme. (Bedingung A2 zunächst erfüllt; und das Urteil ist eine Absicht, aber noch keine, die unmittelbar Handlungen auslösen kann.)

4.a-i.

5. Das Optimalitätsurteil ist eine Fein*vornahme*. (Bedingung A2 zunächst erfüllt; und das Urteil ist eine Absicht, Feinvornahme, aber die Absicht wird ausgestochen.)

5.b-i.

6. Das Optimalitätsurteil ist eine aktuelles (proximales) deiktisches Optimalitätsurteil. (Bedingung A2 zunächst erfüllt; und das Urteil ist eine Absicht, aber die Absicht wird ausgestochen.)

6.c-h.

6.HK Halb kognitivistische (nonkognitivistische) Absichtstheorien: Außer Optimalitätsurteilen können auch andere mentale Ereignisse Absichten sein. Eine solche Absicht, die kein Optimalitätsurteil ist, setzt sich gegen das Optimalitätsurteil durch. (Akrasie.)

6.NK Vollständig nonkognitivistische Absichtstheorien: Absichten sind keine Optimalitätsurteile. Eine solche Absicht setzt sich gegen das Optimalitätsurteil durch. (Akrasie.)

x.a-x.i Unterfälle zu 4-6.

a Projektemacherei, leerer Vorsatz: Die Grobvorname wird nicht zu einer Feinabsicht ausgearbeitet. (Akrasie.)

b Die Vornahme wird zum Handlungszeitpunkt nicht erinnert**,** oder *s* übersieht, daß die Bedingungen für die Vornahme erfüllt sind. (Bedingung A2 nicht erfüllt. Dies ist keine Willensschwäche, sondern Vergeßlichkeit, u.U. allerdings motivierte.)

c Willensschwund: Das Optimalitätsurteil wird (kurzfristig) bewußt revidiert. (A2 nicht erfüllt.)

d Willensschwund: Die Absicht wird in ein bloßes Optimalitätsurteil überführt, das keine Absicht mehr ist: Dadurch wird dieser Fall in Fall 3.6 überführt. (Akrasie.)

e Reflexbedingte schwache Willensschwäche: Ein unterdrückbarer Reflex setzt sich gegen eine Absicht durch. (Bedingung A1 nicht erfüllt.)

f Unbewußt verursachte schwache Willensschwäche: Eine unbewußte Absicht setzt sich gegen eine positive oder negative Grobvornahme durch. (A1 nicht erfüllt.)

g Gewohnheitsbedingte schwache Willensschwäche: Eine Habitualisierung setzt sich gegen eine positive oder negative Grobvornahme durch. (Akrasie.)

h Gefühlsbedingte schwache Willensschwäche: Eine gefühlsbedingte Absicht setzt sich gegen eine ihr widersprechende intellektuelle Vornahme durch. (Akrasie.)

i Schwache nachträgliche / sekundäre Indolenz: Willensschwäche durch Ignorieren: Die Grobvornahme wird verdrängt, ignoriert. (Akrasie oder Bedingung A2 nicht erfüllt.)

Der Unterfall *h* ist eben (Abschn. 4, Nr. 2) schon behandelt worden:

*h. Eine gefühlsbedingte Absicht setzt sich gegen eine ihr widersprechende intellektuellere Absicht durch:* Die neue Absicht kann sich durchsetzen, weil ihr einfach ein Wunsch mehr

zugrunde liegt, der für *a* spricht und der bei der Bildung der rationalen Absicht zu *b* nicht vorhanden war und wegen seiner Instabilität auch rationaliter nicht berücksichtigt werden sollte. Das alte Optimalitätsurteil ist nur noch als eine Erinnerung vorhanden, es wird nicht mehr richtig geglaubt – obwohl es besser begründet ist.

Fall *i* ist dann:
*i. Nachträgliche Indolenz: Akrasie durch Ignorieren: Eine Vornahme wird verdrängt, ignoriert:* Der Handelnde hat das Optimalitätsurteil für *b* früher gefällt, so daß dieses nur eine Vornahme darstellt und nicht aktuell deiktisch präsent ist; er erinnert sich dann an dieses Urteil, ist aber überwältigt von der Aussicht auf die kurzfristig bessere Alternative, wodurch ihm die alte Absicht plötzlich nichts mehr wert ist; er ignoriert sie, entwertet sie als Absicht, ohne aber das Urteil zu revidieren, gegen das er allerdings Einwände hat, die das Urteil jedoch nicht wirklich widerlegen. Ein Student hat sich beispielsweise vorgenommen, an diesem Abend eine der letzten Gelegenheiten wahrzunehmen, für eine wichtige Prüfung zu lernen; als die Arbeit schlecht vorangeht, entscheidet er sich spontan, ins Kino zu gehen (vgl. Wolf 1985: 25; ähnliches Beispiel: Baumeister et al. 1994: 20). Der Mechanismus ist (ähnlich wie im Fall 3.6): Der Handelnde glaubt nicht mehr richtig an das Optimalitätsurteil, er schiebt es weg, als hätte er es nie akzeptiert. Absichten funktionieren häufig so, daß sie im Moment der Handlung bewußt werden müssen. Wenn man die Absicht nicht bewußtwerden läßt oder sie an den Rand des Bewußtseins verdrängt und statt dessen eine andere Absicht fokussiert vor Augen hat, dann ist zu wenig Aufmerksamkeit für die weggedrängte Absicht da, um sich gegen die andere Absicht durchsetzen zu können.

## 6. Lösungsansätze zum Widerspruchsproblem

Das Widerspruchsproblem ist: Ist es möglich und, wenn ja, wie, daß Akratiker sich widersprechende Einstellungen haben? Da es plausibel ist anzunehmen, daß eine Art von kognitivem Widerspruch auch zwischen einem Optimalitätsurteil für *b* und der Absicht zu *a* sowie dem Tun von *a* besteht, haben auch Nonkognitivismen ein Widerspruchsproblem. Aber das Widerspruchsproblem ist für die Optimalitätsurteilstheorie größer, weil der kognitive Widerspruch nach dieser Konzeption direkter ist; es ist ein Widerspruch zwischen zwei Meinungen – nämlich, *a* sei optimal bzw. *b* sei optimal –, also zwischen Einstellungen gleichen Typs.

Meine erste Erklärung des Widerspruchsproblems ist trivial, nämlich daß die meisten willensschwachen Menschen, gerade weil sie um die fundamentalste Form der Rationalität, die doxastische Widerspruchsfreiheit, besorgt sind, wenn sie denn überhaupt ein Optimalitätsurteil für *b* gefällt haben, dieses kurzfristig revidieren und diese Revision eventuell rationalisieren. Der Willensschwache findet z.B. eine entschuldigende Ausnahme, wegen derer in dieser Situation – anders als bei ähnlichen Situationen – *b* doch nicht die beste Handlung ist. Nach den Bedingungen der Willensschwäche ist das Optimalitätsurteil ja besser begründet, die entschuldigende Ausnahme also falsch und eine Rationalisierung; die Revision ist epistemisch irrational. Nach der Tat mag der Willensschwache dies auch wieder klar sehen, deshalb sein Optimalitätsurteil restituieren und seine Handlung entsprechend bereuen. Wenn das Optimalitätsurteil kurzfristig revidiert wird, dann liegt zwar Willensschwäche vor, aber keine Akrasie. Der echte Akratiker stellt uns vor ein größeres Erklärungsproblem, das durch die zweite Erklärung gelöst werden soll.

Diese zweite Erklärung besagt, kurz gefaßt, der Akratiker hält den Widerspruch aus. Doxastische Zustände, unsere subjektiven Repräsentationen der Welt, sind nicht nur Speicher unmittelbarer Erlebnisse, sondern auch Interpretationen und theoretische Konstruktionen, mit denen wir eine Repräsentation weit über das unmittelbar Erlebte hinaus schaffen. Daß solche Konstruktionen, auf der Basis unterschiedlicher Indizien, zu

Widersprüchen führen, insbesondere auch zu ganzen Systemen oder Clustern oder Theorien aus konkurrierenden und untereinander inkonsistenten Meinungen, ist der Preis für diese Ausdehnung; und es ist keine Seltenheit. Schon für den Probabilismus ist es ja der Normalfall, daß wir inkonsistente Meinungen haben, die eben unterschiedliche Wahrscheinlichkeiten haben. Der Preis für die direkte Inkonsistenz ist, daß uns diese Systeme bei praktischen Entscheidungen dann keine Orientierung mehr liefern. Deshalb versuchen wir, Inkonsistenzen zu beseitigen, indem wir Begründungen überprüfen und daraufhin bestimmte Meinungen annullieren u.ä. Doch der intellektuelle Aufwand für solch eine Kohärentisierung ist hoch; schon wegen anderer Aufgaben müssen wir manche Inkonsistenzen erst einmal so stehen lassen und rationaliter unsere subjektiven Wahrscheinlichkeiten entsprechend reduzieren. Bei praktischen Entscheidungen müssen wir dann im Zweifelsfall, wenn der Widerspruch nicht aufgelöst werden konnte, einer der Meinungen oder einem System von Meinungen mehr oder weniger ad hoc den Vorzug geben.

Genau dies mag auch im Fall der Akrasie passieren. Der Akratiker war eh nie hundertprozentig von der Optimalität von *b* überzeugt; er entscheidet sich ad hoc für die Optimalität von *a* und läßt das Optimalitätsurteil für *b* links liegen. Diese Entscheidung mag insbesondere dadurch gebahnt sein, daß das Urteil zugunsten von *b* abstrakter ist, mit längeren Zeitperspektiven rechnet als das zugunsten von *a* (s.o., Abschnitt 4, Nr. 1). Wenn der Akratiker in dieser Situation gefragt werden würde, würde er dem Optimalitätsurteil für *b* wohl gar eine Wahrscheinlichkeit über 0,5 geben; aber er wird nicht gefragt und mag auch nicht daran denken.

## 7. Lösungsansätze zum Irrationalitätsproblem

Nach den Antworten auf das Nichtdurchsetzungs- und das Widerspruchsproblem kann das Irrationalitätsproblem wie folgt spezifiziert werden: Warum verwirft der Willensschwache kurzfristig das besser begründete Optimalitätsurteil, warum ignoriert der Akratiker es ad hoc jeweils zugunsten der schlechter begründeten Absicht, des schlechter begründeten Optimalitätsurteils? Daß diese Form von epistemischer Irrationalität möglich ist, ist relativ trivial und eine Alltagserfahrung: Menschen glauben Dinge ohne Begründung oder ohne gute Begründung; oder sie revidieren widerlegte Meinungen nicht; oder sie revidieren Meinungen, obwohl nichts oder nichts Wesentliches gegen die bisherige gute Begründung der These spricht usw.

Da dieser Schritt nicht epistemisch (kognitiv i.e.S.) begründet ist, dann ist er vermutlich motiviert oder beruht auf kognitiven Fehlern. An diesem Punkt liefern uns jedoch schon die Lösungsvorschläge zum Versuchungsproblem die Erklärung: *1. Zeitliche Kurzsichtigkeit* mit Schwelgen in den kurzfristigen Folgen bzw. Panik vor den kurzfristigen Folgen führt zu einer Emotionalisierung; in dieser emotional aufgeladenen Situation wird dann eine Neubewertung, Schnellüberprüfung der alten Urteile vorgenommen, bei der die emotional betonten Folgen übermäßig präsent sind und in Eile und mit extremer Risikoneigung die langfristigen Folgen ignoriert bis abgetan werden. *2. Gefühlsinduzierte intrinsische Wünsche* schaffen eine neue motivationale Lage mit zusätzlichen intrinsischen Wünschbarkeiten, die das Optimalitätsurteil für *b* sogar schlechter begründet erscheinen läßt; man muß dann schon sehr gezielt aus theoretischen Überlegungen heraus die frisch induzierten intrinsischen Bewertungen verwerfen, um das alte Optimalitätsurteil aufrechtzuerhalten.

## 8. Lösungsansätze zum Willensstärkeproblem

Nach den Definitionsbedingungen für ‘Akrasie’ konnte der Akratiker die besser begründete Handlung *b* ausführen, er konnte also die Akrasie überwinden. Was zunächst wegen der eben

(Abschn. 4-5; 7) gegebenen Erklärungen der Willensschwäche selbst wie ein kaum zu lösendes Problem aussieht, ist, da Willensschwäche durchaus auch ein praktisches psychologisches Problem ist, Thema psychologischer und psychotherapeutischer Forschung und Ratgeber, die erfolgreiche Anweisungen zur Überwindung der Willensschwäche liefern (z.B. Baumeister & Tierney 2011; Kuhl 1983: 251-301, 304-325; 1987: 286-288; 1996; s.a. Kennett 2001: 135-137). Ein Teil dieser Ratschläge betrifft asynchrone Maßnahmen der Selbststeuerung, die man also ergreift, bevor man in Versuchung gerät, z.B. Zigaretten oder Alkohol aus dem Haus verbannen. Im folgenden sollen nur synchrone Selbststeuerungsmechanismen gegen Akrasie vorgestellt werden, die Akrasie durch gefühlsinduzierte Wünsche oder Schwelgen in bzw. Panik vor den kurzfristigen Folgen der Alternativen verhindern mögen.

Eine wichtige Strategie gegen viele Formen der Akrasie durch gefühlsinduzierte Wünsche ist, einfach mit der Handlungsentscheidung zwischen *a* und *b* abzuwarten, also nicht im Affekt zu handeln, sondern zu warten, bis die Gefühle sich einigermaßen gelegt haben und somit auch der gefühlsinduzierte intrinsische Wunsch verschwunden oder zumindest deutlich abgeschwächt ist; anschließend kann dann wieder eine rationale Entscheidung gefällt werden. (Dies funktioniert natürlich nur bei Gefühlen, die von selbst abebben, wie etwa Wut, oder durch eigene Anstrengungen abgeschwächt werden können, nicht aber bei Gefühlen, die von innen oder außen laufend genährt oder gar verstärkt werden, wie z.B. Hunger, Durst bzw. Schmerzen durch Folter.) Zwei weitere Standardstrategien sind, sich auf die langfristigen Folgen der Handlungen zu konzentrieren bzw. die Begründung für die willensstarke Handlung *b* noch einmal durchzugehen. Damit wird dann der Tendenz zur zeitlichen Kurzsichtigkeit entgegengewirkt. Eine weniger verwendete und gelegentlich auch etwas risikoreiche Selbststeuerungsstrategie ist, dann, wenn man bemerkt, daß man aus einer emotionalen Versuchung heraus anfängt, eine Neubewertung vorzunehmen, diese Neubewertung abzubrechen und sich ihre Fortsetzung zu untersagen, weil sie unter dem emotionalen Einfluß nur schlechter begründet sein kann, als das gut begründete Optimalitätsurteil für *b*. Einige willensschwache Handlungen bestehen aus Exzessen, in die der Willensschwache nach und nach hineinschliddert: exzessiver Alkohol- oder Drogenkonsum, exzessive Einkäufe, Glücksspiele u.ä. Gegen diese Art von Willensschwäche ist es wichtig, den Grad des Konsums und die Annäherung an die kritische Grenze genau zu kontrollieren.

Dies sind nur einige Beispiele für alltägliche, einfache und erfolgreiche Selbststeuerungsstrategien gegen Willensschwäche. Es geht hier nicht um Vollständigkeit, sondern nur darum zu zeigen, daß der Willensschwache synchrone Mittel zur Willensstärke anwenden und damit willensstark handeln kann. Solche Strategien sind nicht utopisch, auch wenn sie einige Erfahrungen mit Willensschwäche und ein Lernen zum strategischen Umgang damit voraussetzen; ohne die strategischen Kenntnisse ist man aber den Versuchungen der Willensschwäche hilflos ausgeliefert. Wie Mischel zeigt, kennen aber schon die meisten Fünfjährigen Strategien zur Willensstärke und wenden sie auch an (Mischel <1989> 1992).

Das Paradoxon, daß die Überwindung der Willensschwäche anscheinend schon die genau nicht vorhandene Willensstärke voraussetzt, wird bei all diesen synchronen Strategien so aufgelöst: Zum einen ist der Einsatz der Strategie eine eigene, von *a* oder *b* unabhängige Handlung, die auch nicht per se schon die Wahl von *b* voraussetzt. Zum anderen enthalten die Erfahrung mit sowie das psychologische Hintergrundwissen zu Selbststeuerungsstrategien bei hinreichend informierten Personen nicht nur Kenntnisse solcher Handlungsmöglichkeiten, also synchroner Selbststeuerungsstrategien, sondern auch Wissen darüber, daß diese Strategien vorteilhaft sind und in welchen Situationen sie einzusetzen sind. In fortgeschrittenen Phasen der Selbststeuerung lösen kritische Situationen dann schon automatisch Gedanken aus, daß und welche geeignete Selbststeuerung jetzt angebracht wäre.

Nonkognitivistische Handlungstheorien berufen sich in ihren Erklärungen, wie Willensstärke möglich sei, auf die nämlichen Selbststeuerungsstrategien. Aber sie haben Schwierigkeiten, das Wirken dieser Selbststeuerungsstrategien zu erklären, genauso wie sie allgemein Schwierigkeiten haben, die Rationalität von Entscheidungen und Handlungen zu erklären. Oberflächlich können natürlich auch diese Theorien das sagen, was hier über die Selbststeuerungsstrategien ausgeführt wurde: Diese Strategien stehen einem selbstverständlich auch nach einer nonkognitivistischen Handlungstheorie offen, weil es ja eigene, von der Entscheidung zwischen *a* und *b* unabhängige, gute Handlungen sind. Aber tiefgründiger bleibt bei nonkognitivistischen Theorien das allgemeine Problem bestehen: Genau wie diese Theorien nicht erklären können, wie man dazu kommt, die als beste erkannte Handlung auszuführen, genausowenig können sie erklären, wie man dazu kommt, jetzt eine für gut erachtete Selbststeuerungsstrategie einzusetzen: Normalerweise führt das Optimalitätsurteil zur Absicht, bei Willensschwäche aber gerade nicht. Was kann man nun tun? Führt ein Optimalitätsurteil über den Einsatz von Selbststeuerungsmechanismen zu deren Ausführung? Zudem betreffen die aufgezeigten Selbststeuerungsmechanismen ja nicht ganz allgemein das Bilden einer Absicht, sondern das Bilden eines Optimalitätsurteils; diese Mechanismen ändern sehr präzise die Bedingungen für die Entscheidungsfindung, greifen in die Bildung des Optimalitätsurteils ein. Darüber sagen nonkognitivistische Modelle wie die Sui-generis-Theorie der Absicht ja gar nichts. Sie können nur aus der Literatur entnehmen, daß solche Mechanismen funktionieren; sie können ihre Funktionsweise aber nicht mit der eigenen Theorie erklären, weil diese Theorie eben nichts über den genauen Entscheidungsprozeß und insbesondere nichts über den Einfluß von Kognitionen auf die Entscheidung sagt. Nach der Strategie beispielsweise, den Einfluß der gefühlsinduzierten Wünsche auf die Entscheidung auszuschalten, muß man bei der Einleitung dieser Selbststeuerungsstrategie glauben, daß Bewertungen und Optimalitätsurteile, die nur auf stabilen intrinsischen Wünschen beruhen, besser sind, als solche, die auch auf kurzfristigen, gefühlsinduzierten intrinsischen Wünschen beruhen, und deswegen beschließen und die Absicht bilden, den Einfluß der gefühlsinduzierten Wünsche zurückzudrängen. Wie will eine nonkognitivistische Theorie diese Prozesse erklären? Selbst wenn sie sich auf eine unabhängige Theorie rationaler Bewertungen verläßt, kann sie qua Nonkognitivismus nicht erklären, wie deren Einsichten handlungswirksam werden können. Und Mele, der annimmt, daß Optimalitätsurteile per Voreinstellung („default“) automatisch zur Bildung der entsprechenden Sui-generis-Absichten führen (Mele 1992: 228 f., 230-232), könnte zwar die hier gelieferten Erklärungen der Bildung von Optimalitätsurteilen und der darin enthaltenen Möglichkeiten, der Willensschwäche zu entgehen, begrüßen; aber im Grunde würde seine Theorie dadurch kognitivistisch werden, und die Sui-generis-Absichten hätten nur noch eine überflüssige und obsolete Hilfsfunktion, z.B. das Entschiedene markant präsent zu halten.

## 9. Fazit

Ich bin am Ende meiner Darlegungen angelangt. Ich hoffe, damit allgemein eine Reihe wichtiger Erklärungen zum Phänomen der Akrasie geliefert zu haben und insbesondere gezeigt zu haben, daß die Optimalitätsurteilstheorie keine *besonderen* Schwierigkeiten hat, Akrasie zu erklären, während nonkognitivistische Theorien, wie die Sui-generis-Theorie der Absicht, grundsätzlich viel größere Erklärungsprobleme haben.

**Danksagung:** Ich danke den Hörern der Vortragsfassung dieses Artikels auf dem GAP-Kongreß 2012 für die fruchtbare Diskussion, insbesondere Rüdiger Bittner, Benjamin Kiesewetter, Anna Kusser und Walter Pfannkuche.

**Christoph Lumer**
Università di Siena
lumer@unisi.it

## Literatur

Audi, R. 1979: „Weakness of Will and Practical Judgement“, *Nous* 13, 173-196.

Baumeister, R. F., T. F. Heatherton und D. M. Tice 1994: *Losing control: How and why people fail at self-regulation*. San Diego, CA: Academic Press.

Baumeister, R. F. und J. Tierney 2011: *Willpower. Rediscovering the Greatest Human Strength*. London: Penguin. (Dt. Übers.: Die Macht der Disziplin: Wie wir unseren Willen trainieren können. Frankfurt (Main), New York: Campus 2012.)

Bratman, M. E. 1979: „Practical Reasoning and the Weakness of the Will“, *Nous* 13, 153-171.

— 1987: *Intention, Plans, and Practical Reason*. Cambridge (MA), London: Harvard University Press. (Neuausgabe: Cambridge: Cambridge U.P. 1999.)

Elster, J. 1999: *Strong Feelings. Emotion, Addiction and Human Behavior*. Cambridge (MA), London: MIT Press.

Holton, R. 1999: „Intention and Weakness of Will“, *The Journal of Philosophy* 96, 241-262.

— 2009: *Willing, Wanting, Waiting*. Oxford: Clarendon.

Kennett, J. 2001: *Agency and Responsibility. A Common-sense Moral Psychology*. Oxford [etc.]: Clarendon.

Kuhl, J. 1983: *Motivation, Konflikt und Handlungskontrolle*. Berlin [etc.]: Springer.

— 1987: „Action Control. The Maintenance of Motivational States“, in F. Halisch, J. Kuhl (Hrg.): *Motivation, Intention, and Volition*, Berlin [etc.]: Springer, 279-291.

— 1996: „Wille und Freiheitserleben. Formen der Selbststeuerung“, in J. Kuhl, H. Heckhausen (Hrg.): *Motivation, Volition und Handlung* (= *Enzyklopädie der Psychologie, Serie: Motivation und Emotion*, Bd. 4.), Göttingen, Bern, Toronto, Seattle: Hogrefe, 665-765.

Lumer, Ch. 1997: „The Content of Originally Intrinsic Desires and of Intrinsic Motivation“, *Acta analytica* 18, 107-121.

— 2005: „Intentions Are Optimality Beliefs - but Optimizing what?“, *Erkenntnis* 62, 235-262.

— 2007: „An Empirical Theory of Practical Reasons and its Use for Practical Philosophy“, in Ch. Lumer, S. Nannini (Hrg.): *Intentionality, Deliberation and Autonomy. The Action-Theoretic Basis of Practical Philosophy*, Aldershot: Ashgate, 157-186.

— 2009: *Rationaler Altruismus. Eine prudentielle Theorie der Rationalität und des Altruismus*. Zweite, durchgesehene und ergänzte Auflage. Paderborn: mentis.

— 2012: „Emotional Decisions. The Induction-of-Intrinsic-Desires Theory“, in A. Innocenti, A. Sirigu (Hrg.): *Neuroscience and the Economics of Decision Making*, Abingdon, New York: Routledge, 109-124.

McCann, H. J. 1998: „Practical Rationality and Weakness of Will“, in *The Works of Agency. On Human Action, Will, and Freedom*, Ithaca, London: Cornell University Press, 213-234.

Mele, A. R. 1987: *Irrationality. An Essay on Akrasia, Self-Deception, and Self-Control*. New York, Oxford: Oxford University Press.

— 1992: *Springs of Action. Understanding Intentional Behavior*. New York, Oxford: Oxford University Press.

Mischel, W., Y. Shoda und M. L. Rodriguez <1989> 1992: „Delay of Gratification in Children". In: G. Loewenstein, J. Elster (Hrg.): *Choice over time*, New York: Russell Sage Foundation 1992, 147-164.

Moore, M. S. <1993> 2010: *Act and Crime. The Philosophy of Action and its Implications for Criminal Law*. Oxford: Oxford University Press.

Platon: „Protagoras". In: Werke in acht Bänden. Griechisch und deutsch. Hrg. v. Gunther Eigler. (Griech. Text nach Edition Maurice Croiset. Dt. Übers. nach F. Schleiermacher.) Sonderausgabe. Darmstadt: Wissenschaftliche Buchgesellschaft 1977; 21990. Bd. 1, 83-217.

Smith, M. 2003: „Rational Capacities. Or: How to Distinguish Recklessness, Weakness, and Compulsion", in S. Stroud, Ch. Tappolet (Hrg.): *Weakness of Will and Practical Irrationality*, Oxford: Clarendon Press, 17-38. (Wiederabdruck in: Ders.: *Ethics and the Apriori. Selected Essays on Moral Psychology and Meta-Ethics*. Cambridge: Cambridge University Press 2004, 114-135.)

Watson, G. 1977: „Skepticism about Weakness of Will", *The Philosophical Review* 86, 316-339. (Wiederabdruck in G. Watson: *Agency and Answerability. Selected Essays*, Oxford: Clarendon 2004, 33-58. - Dt. Übers.: „Skepsis bezüglich Willensschwäche", in Th. Spitzley (Hrg.): *Willensschwäche*, Paderborn: Mentis 2005, 107-127.)

— 1999: „Disordered Appetites. Addiction, Compulsion, and Dependency", in J. Elster (Hrg.): *Addiction. Entries and Exits*, New York: Russell Sage Publictions, 3-28. (Wiederabdruck in Ders.: *Agency and Answerability. Selected Essays*, Oxford: Clarendon 2004, 59-87.)

Wolf, U. (1985): „Zum Problem der Willensschwäche" *Zeitschrift für Philosophische Forschung* 39, 21-33. (Wiederabdruck in: Th. Spitzley (Hrg.): *Willensschwäche*, Paderborn: Mentis 2005, 128-138.)

# The Case against Consequentialism: Methodological Issues

Nikil Mukerji

Over the years, consequentialism has been subjected to numerous serious objections. Its adherents, however, have been remarkably successful in fending them off. As I argue in this paper, the reason why the case against consequentialism has not been more successful lies, at least partly, in the methodological approach that critics have commonly used. Their arguments have usually proceeded in two steps. First, a definition of consequentialism is given. Then, objections are put forward based on that definition. This procedure runs into one of two problems. Substantive criticisms of consequentialism can only be formulated, if the posited definition is sufficiently concrete and narrow. In that case, however, consequentialists can defend themselves using a strategy that I call "interpretive divergence". They can simply point out that the critic's definition does not accord with their understanding of consequentialism to which criticisms do not apply. If, on the other hand, an all-encompassing definition is used, it is so abstract that it is doubtful whether any substantive criticisms can be formulated. To escape this dilemma, I sketch a methodological approach which drops the assumption that consequentialism should be defined. It assumes, rather, that the term "consequentialism" should be interpreted as a Wittgensteinian family resemblance term.

In recent decades the debate in normative ethics has in large part been a debate about the issue whether consequentialism is a tenable moral view. In this paper, I will not address this question. Rather, I will discuss how we should address it. In particular, I have three aims. I want to describe how the case against consequentialism has commonly been made. I want to explain why this procedure is problematic. And I want to sketch the rough outline of an alternative method that I take to be more fruitful. Throughout, I shall be interested solely in methodological issues. I shall remain agnostic, that is, about the substantive question whether consequentialism is, in fact, a tenable moral view.

The remainder falls into three sections. In the first section, I outline, discuss and reject the conventional approach which I call the Definitional Method (DM). It is based on the assumption that the idea of consequentialism should first be defined in terms of a necessary and sufficient feature shared by all consequentialist moral theories and then criticized. As I will show, this procedure runs into one of two problems. If a narrow definition is posited, substantive criticisms can be formulated; but consequentialists can defend themselves against these criticisms using a strategy that I call interpretive divergence. They can simply point out that the critic's definition does not accord with their understanding of consequentialism which, they can claim, is immune to these objections. If, on the other hand, the definition is all-encompassing, its abstractness seems to make it impossible to come up with any substantive criticisms. In the second section, I sketch out a new approach which seeks to address the dilemma of the DM. I call it the Family Resemblance Approach (FRA). It assumes that consequentialism should not be defined, but interpreted as a Wittgensteinian family resemblance term. In the third section, finally, I sum up and conclude.

## 1. The Definitional Method

Critics of consequentialism usually use the following method to argue against it.

**Definitional Method (DM)**

Step 1: Define Consequentialism in terms of a necessary and sufficient feature (or necessary and jointly sufficient set of features) that is shared by all consequentialist moral doctrines.

Step 2: Formulate a decisive objection against all doctrines which possess this definitional feature (or set of features).

In this section, I argue that the DM leads critics of consequentialism into a dilemma: If they choose a sufficiently general definition of consequentialism that envelops all forms of the doctrine, its abstractness makes it impossible to formulate any substantive criticisms. If, on the other hand, they start with an account of consequentialism which allows us to formulate substantive criticisms, then it does not capture all consequentialist moral theories and gives consequentialists the chance to defend themselves rather easily using a strategy that I call "interpretive divergence". To illustrate the problem at hand, I shall apply the method. To this end, let me introduce a definition of consequentialism, as it is proposed, roughly, by Hooker (2003).

**Narrow Definition of Consequentialism**

A moral theory is consequentialist if and only if it judges that an act is right if and only if it maximizes the impartial good.

A number of well-known criticisms apply to moral theories that come under this definition. Here are some examples. First, we may argue that consequentialist moral theories are overly demanding.[1] They judge that an act is wrong unless it maximizes the good of all weighted equally (whatever that good consists in). Hence, they condemn it, e.g., if you go to the cinema to see a movie, as there are obviously better ways for you to spend your money and time (Kagan 1998). You should rather, say, donate the money for the movie ticket to a charitable organisation. And you should preferably volunteer in a soup kitchen to help those in need, instead of spending your time in shallow amusement. Doing these things, it seems, would have better consequences, impartially considered. Now suppose you do both of these things. You work in the soup kitchen for two hours instead of seeing the movie and you donate the money that you would have spent on the ticket to a good cause. Are you now free to watch a movie? On consequentialism, as I have just defined it, it seems that the answer is no. Presumably, there is *always* something you could do which would do more impartial good than going to the cinema. So it seems that, on any consequentialist doctrine, you should *never* go to see that movie, because it is very unlikely that this is ever the best thing to do. But this seems absurd. Any doctrine which demands that moral agents constantly forgo the things that make their lives worth living (e.g. watching a movie every now and then) seems overly demanding. Consequentialist doctrines apparently are, then, overly demanding and should, therefore, be rejected.

Second, we may contend that consequentialist moral theories violate moral constraints, e.g. the constraint against killing an innocent person. To see this, consider Thomson's (1976) case "fat man". Imagine you are standing on a footbridge over a railway. You see a trolley approaching and can tell that it is out of control. There are five people on the tracks. The trolley will run over them and kill them unless it is stopped. You reason that the only way to stop it at this stage is to drop a heavy object in its path. There is a very fat man standing next to you on the footbridge. You could give him a shove. He would fall, land on the tracks and stop the trolley. This would, of course, kill the guy. But the five would go unharmed. Should you do it? On consequentialism, as I have defined it above, the answer is surely yes. Pushing the guy off the bridge results in less damage. One guy dies, instead of five. Hence, it obviously

[1] For a comprehensive discussion of the demandingness objection to consequentialism see, e.g., Hooker (2009) and Mulgan (2001).

maximizes the good (or minimizes the bad), impartially considered. But most of us believe that pushing the guy off the bridge would be immoral, as it would violate a moral constraint against killing an innocent person. Consequentialism seems to give the intuitively wrong answer in this case and should, hence, be rejected.

Thirdly, we may point out that consequentialist moral theories have no place for special obligations (Jeske 2008). This can be illustrated using the following case.[2] Imagine, e.g., that a friend of yours is in danger. She is in a building that has caught on fire and needs help to get out. But she is not the only person in there. Other people are also trapped in the building. And they need help as well. You are in a position to help any one of them. But you only have time to save one person before the building collapses. What should you do? According to consequentialism, as we have defined it above, you should attach the same weight to your friend's well-being as to everyone else's and do what will produce the most good, again impartially considered. If you happen to know, e.g., that there is an excellent surgeon amongst the people in the building, you should save that person instead of your friend who, we may assume, has a comparatively less important job. But this, it seems, would be immoral. In so acting you would fail to recognize the special moral obligation that you owe your friend. So consequentialism seems to lead you astray as to the real obligations that you have in this case. It should, therefore, be rejected.

These, I think, are valid concerns. Nevertheless, consequentialists can easily defend themselves against them using the strategy of "interpretive divergence".[3] All they need to do is to point out that the definition on which objections are premised does not accord with *their* definition of consequentialism.[4] Let us see how this would work in each of the three cases.

Consequentialists can rebut the demandingness objection, e.g., by pointing out that it crucially relies on the premise that all forms of consequentialism require the agent to do the *best* she can. Satisficing consequentialism, they can argue, does not require that (Slote 1984). On this doctrine, an act is right as long as its consequences are *good enough*. Satisficing consequentialism can, therefore, plausibly allow you to go to the cinema and watch a movie. Though doing that may not have the best consequences overall, it arguably has good enough consequences. E.g., you get to see a movie that you enjoy. The cinema turns a profit and can afford to employ people who need a job. They, in turn, can buy stuff from others and so on. That's not too bad, is it? But, then, it should be morally permissible. And this, in turn, means that the demandingness objection is not substantiated by the example.

As far as the second objection is concerned, consequentialists may argue that their theories can, in fact, incorporate moral constraints (Portmore 2005). Initially, this may sound strange. Moral constraints, one might say, apply to actions and not to their consequences. But consequentialists can point out that the line between the act and its consequences is a mere chimera and that the act itself should therefore routinely be included amongst its consequences.[5] This allows consequentialists, e.g., to consider the fact that the agent has to

---

[2] The description of the case is based on an example that I used in Mukerji (2013). It is fashioned after an infamous illustration by Godwin (1793) that has come to be known as the "famous fire cause" Barry (1995: 222).

[3] McNaughton & Rawling (1991) and Ridge (2005) call it the "the consequentialist vacuum cleaner".

[4] There are various examples of consequentialists applying this strategy. John Broome, e.g., applies it when he concedes that "[m]any serious doubts have been raised about consequentialism" before hastening to add that "they are not about consequentialism as *I* defined it." (Broome 2004: 42; emphasis added) And Walter Sinnott-Armstrong does so too when he says: "Even if other philosophers mean something else by 'consequentialism', I will be satisfied if my argument supports the view that *I* labelled 'consequentialism'." (Sinnott-Armstrong 2001: 345; emphasis added)

[5] What I do "may be described variously as making marks on a piece of paper, signing a cheque, paying a bribe, or ensuring the survival of my business." (Sumner 1987: 166) In other words, the boundary between the act and its consequences can be pushed back and forth, depending on the chosen

*kill* one person in order to save five lives as part of her act's "comprehensive outcome" (Sen 2009). In the "fat man" case they can, hence, maintain that the fact that you have to *kill* the fat guy in order to save the five on the tracks is accessible to consequentialist moral evaluation. Moreover, they can take an agent-relative stance on the evaluation of this consequence. Consequentialists can argue, not only that *killings* are worse than mere deaths, but also that killings *done by the agent* are worse, from the moral perspective of the agent, than killings done by other persons. This, in turn, makes it possible for them to say that you may be forbidden, on certain versions of consequentialism, to push the fat man off the bridge. And this is precisely what common sense suggests.

In regards to the third objection, consequentialists can claim that critics falsely assume consequentialist theories to be strictly impartial. As David Brink explains, this need not be so. In fact, consequentialists can adopt a position like C. D. Broad's (1971) "self-referential altruism". It requires a universal concern for all, but allows different weights to be attached to the welfare of different individuals according to "the nature of the relationship in which the agent stands to potential beneficiaries" (Brink 2006: 382). In the case of the burning building you may, then, attach greater weight to the welfare of your friend and save her. Consequentialists can, hence, make room for special obligations.

As it turns out, then, it is easy for consequentialists to dodge objections if they are founded on the narrow definition of consequentialism that I have just used. Of course, this, by itself, does not show that there is no definition which can avoid this problem. Perhaps the problem that I have illustrated is a specific complication of the particular definition that I have chosen. But this seems not to be the case. If there was, in fact, a definitional feature of consequentialist doctrines, we should be able to discover it by examining the characteristics of a paradigmatic consequentialist doctrine, such as classic utilitarianism.

**Classic Utilitarianism (CU)**

An act is right if and only if it maximizes the sum total of sensory happiness of all sentient creatures.

This doctrine possesses a number of characteristics. We can examine them one by one and ask, in each case, whether it may serve as a defining feature of consequentialism. The first property of CU that leaps to the eye is its maximizing nature. It requires that the agent do the *best* she can. As I have already explained, however, this cannot be regarded as a defining characteristic of consequentialism, because there are also satisficing doctrines.[6]

A further feature of CU is the hedonistic idea that the only intrinsic good is sensory happiness. This, too, cannot be a defining feature of consequentialist doctrines, because there are also non-hedonistic ideas about the ultimate good. Some theorists, e.g., hold the view that the only intrinsic good is desire-satisfaction (e.g. Hare 1981 and Singer 1979/1993). But, perhaps, all consequentialist theories are united by a more general idea about goodness. Both hedonism and desire-satisfactionism are special variants of the notion that the only intrinsic good is individual welfare, a view that goes by the name "welfarism" (Sen 1979). But welfarism cannot be seen as a defining characteristic of consequentialism either, since there are non-welfarist forms of consequentialism which recognize goods other than individual welfare, e.g., knowledge, accomplishments and so on.

A further aspect of CU is that it takes the overall good to be the *sum* of individual parts. This, however, cannot be seen as a defining characteristic of consequentialism either, as G. E.

---

description of the events. For this reason consequentialists can simply claim that the act itself should routinely be included amongst its consequences (Scheffler 1982/1994).

[6] Many theorists, however, take the maximization principle to be an essential feature of all consequentialist doctrines and define consequentialism in reference to it. See, e.g., Arneson (2004), Nida-Rümelin (1993), Scheffler (1982/1994), Williams (1973).

Moore's (1903/1959) consequentialist view makes clear. Moore famously opposed the view that the value of a whole is not identical to the sum of the values of its parts. But again, we may suspect that there is a more general idea behind the principle of summation that might be suitable to define consequentialism. Perhaps all consequentialist doctrines share an *aggregative* conception of the good, meaning, roughly, that they allow losses to one individual to be compensated by benefits to another. This idea will not do either, because consequentialism is not tied to aggregation (Hirose 2004). One can be a consequentialist while rejecting aggregation.[7]

Yet another feature of CU that is often emphasized (e.g. by Williams 1981) is its impartiality. This notion can be factorized into two separate views. It involves, firstly, the idea of universalism, *viz*. that the well-being of all sentient beings is to be taken into consideration and, secondly, the idea of equal treatment, *viz*. that everybody's well-being is to be weighted equally (Mukerji 2013). Neither of these ideas, however, can be seen as a defining characteristic of consequentialism generally, as consequentialist moral theories can depart from both of these views. Self-referentially altruistic versions of consequentialism depart from the idea of equal treatment. Egoistic forms of consequentialism deny the principle of universalism.

A further noteworthy property of CU is that the rightness of an act is judged solely on the basis of its objective outcome, *viz*. the happiness that is actually produced. Perhaps this idea may serve as a definitional characteristic. But we should not get our hopes up, because there are, of course, versions of consequentialism which evaluate acts based on their *expected* consequences.[8]

As Sinnott-Armstrong (2011) points out, the only feature that CU seems to share with all other consequentialist theories is the very basic and abstract idea that the only thing which determines whether an act is right or wrong is the (expected or actual) amount of goodness that it brings about. We can, of course, define consequentialism in terms of this idea.

**Broad Definition of Consequentialism**

A moral theory is consequentialist if and only if it judges whether an act is right only based on the extent to which it *promotes goodness*.

This definition does seem to include all conceivable consequentialist theories.[9] But it gives rise to another problem. It appears to be close to vacuous. It is questionable, therefore, whether one could, in fact, formulate any meaningful criticisms on its basis. I, for one, am at a loss when it comes to thinking of any. This may initially seem implausible. But remember that any criticism formulated on the basis of such a broad definition has to be independent of all ideas that I have ruled out as defining features of consequentialism. We cannot assume, e.g., that consequentialist doctrines require that the agent bring about the *best* consequences. We cannot assume that only individual welfare matters. We cannot assume that the good is aggregative and so on. I do not see how we could conceivably make a case against consequentialism that is independent of all these substantive ideas.[10]

---

[7] Such a position is held, e.g., by Mendola (2006).

[8] Feldman (2006) calls these forms of consequentialism "expected utility consequentialism".

[9] The only notable exception I know of is Portmore's (2011) moral theory. It does not evaluate actions based on a single measure of goodness. Nevertheless, Portmore claims that it is a form of consequentialism.

[10] It is, of course, possible to object to consequentialism on the ground that it requires – very generally – that moral agents consider only consequences and because it is, hence, incompatible with our conviction that "that there are certain things forbidden whatever consequences threaten." (Anscombe 1958: 10) This criticism relies only on the basic idea behind consequentialism that is captured in the broad definition. The objection is question-begging, though. It merely states that consequentialism is

To sum up, then, it seems that the DM runs into one of two problems. The definition that is used is either too narrow or too wide. If we posit a narrow definition, we can formulate substantive criticisms. But, as I have shown, consequentialists can dodge them using the strategy of interpretive divergence. If, on the other hand, we choose a broad definition that encompasses all forms of consequentialism, it is so abstract that no substantive criticisms apply. Given this dilemma, I would like to suggest a new and more promising alternative method for criticizing consequentialism. Unlike the DM, it does not require us to give a definition of consequentialism.

## 2. The Family Resemblance Approach

It may be hard to understand how a philosophical investigation can get off the ground, if the object to be investigated is not defined.[11] But this resistance seems to be rooted in a warped view of language. It is assumed that there are only two kinds of general terms, *viz*. "basic terms" and "composite terms" (Sluga 2006). The former are used to pick out observable characteristics, while the latter refer to more complex objects and are defined in terms of the former. On this view, it seems to be inexplicable how the term "consequentialism" can be made sense of, if it does not fall into either category. But we do not need to buy into this ontology of language. As Ludwig Wittgenstein has famously suggested, many general terms may fall into a third category. They may be "family resemblance terms". Recently, some philosophers have suggested that "consequentialism" should be interpreted as such a family resemblance term (Portmore 2007; Sinnott-Armstrong 2011).[12] In what follows, I shall explore this idea and examine how we can use it to devise a methodical approach for criticizing consequentialism.

To start, it is important to get clear on the idea of family resemblance. As Wittgenstein explains, family resemblance obtains between the objects of a given class, if they form "a complicated network of similarities overlapping and criss-crossing" (Wittgenstein 1953/1986: §66), while there is *no single feature* they all share in common which could serve as the basis of a definition.[13]

To be sure, overlapping and criss-crossing are distinct ideas.[14] Consider three objects *a*, *b* and *c*. Each of them possesses three out of six components *A*, *B*, *C*, *D*, *E* and *F*, as the following table shows.

1. Table: An Illustration of Family Resemblance: Criss-Crossing

| *a* | *b* | *c* |
|---|---|---|
| *ABC* | *ADE* | *BDF* |

incompatible with an ethic that forbids certain acts categorically. In and of itself, this is not a reason that consequentialists need to accept as speaking against their doctrine.

[11] This view goes back at least to Plato. See, e.g., the dialogue *Euthyphro*. Here, Socrates insists that his interlocutor produce a definition of piety. See also Woodruff (2010), esp. his remarks on Socratic definition and the priority of definition (sec. 3-4).

[12] The notion of family resemblance is usually seen as originating in Wittgenstein's *Blue Book* (1933-4/1960). An oft-cited discussion can be found in the *Philosophical Investigations* (1953/1986: §§66,67).

[13] Wittgenstein includes further characteristics. He suggests, e.g., that family resemblance terms are also vague, i.e. their extensions are indeterminate. As Michael Forster argues, however, this is a mistake. He says that "it would in principle be quite consistent with Wittgenstein's core model of family resemblance concepts (...) that it leaves the extension of such a concept perfectly determinate" (Forster 2010: 67). In what follows, I follow Forster's view.

[14] Both of the following examples are adapted versions of examples used by Forster (2010).

The similarities that are characteristic of the family *abc* criss-cross in this case. That is, the components that the objects share are different throughout pairs. *a* and *b* share component *A*. *b* and *c* share *D*. And *a* and *c* share component *B*. Overlapping, on the other hand, is illustrated by the following example.

2. Table: An Illustration of Family Resemblance: Overlapping

| *a* | *b* | *c* | *d* |
|---|---|---|---|
| *ABC* | *BCD* | *CDE* | *DEF* |

Here, the similarities between objects extend not only throughout pairs. They overlap – at least in the case of components *C* and *D*. These run through *a*, *b*, *c* and *b*, *c*, *d*, respectively.

Having characterized the idea of family resemblance, we should ask, then, whether it may be adequate to interpret the class of consequentialist doctrines as a family. There may, of course, be doubts. In the previous section, I said that there is, in fact, a basic idea that lies behind *all* variants of consequentialism, *viz*. that the moral status of an act depends only on the extent to which it promotes goodness. This seems to make it problematic to interpret "consequentialism" as a family resemblance term. For if there is one basic idea that lies behind *all* consequentialist moral doctrines, these doctrines apparently *do* share one characteristic in common. This, in turn, seems to rule out that we can think of them as being related *via* family resemblance which would imply that there is no such feature.

At this point, it seems to be instructive to introduce the distinction between the substantive *content* of a moral theory and its *structure*, as it is drawn, e.g., by (Hurka 1992). To this end, let us go back to the first example of family resemblance which illustrates criss-crossing. We can interpret *a*, *b*, and *c* as moral theories and *A*, *B*, *C*, *D*, *E* and *F* as their logical components. Consider the statement:

(1) _ contains components *A*, *B* and *C*.

The placeholder _ stands for a moral doctrine. If we plug in *a* for _ in (1), we get a true statement about the *content* of *a*, *viz*.:

(2) *a* contains components *A*, *B* and *C*.

If we plug in *b* for _ in (1), however, we get a false statement, *viz*.

(3) *b* contains components *A*, *B* and *C*.

In contrast, consider

(4) _ contains three out of six components *A*, *B*, *C*, *D*, *E* and *F*.

If we substitute *a* for _, we get a true statement about *a* again, *viz*. a true statement about the structure of *a*.

(5) *a* contains three out of six components *A*, *B*, *C*, *D*, *E* and *F*.

But we can also plug *b* in for _ and get a true statement.

(6) *b* contains three out of six components *A*, *B*, *C*, *D*, *E* and *F*.

The fact that (2) and (3) are true and false, respectively, though (5) and (6) are both true, suggests that the members of a family of moral doctrines *a*, *b*, *c*, ... may share a given *structural* feature, even though there is no *substantive* feature that all of them share. If the basic idea behind consequentialism relates, then, to the structure of consequentialist doctrines rather than to their content, it is not ruled out that consequentialism can be construed as a family of moral doctrines. This condition seems, indeed, to be fulfilled. The basic idea behind consequentialism that is captured in the broad definition that I have

introduced towards to the end of the previous section says merely that the rightness of an act depends only on the extent to which that act promotes goodness. This, in effect, says only that every consequentialist doctrine has to possess certain *kinds* of components – *viz*. a theory that explains how goodness is measured and a theory that explains precisely how the goodness of an act relates to its moral status.[15] The definition is, hence, analogous to proposition (4). It only concerns the structure of consequentialist doctrines which is compatible with the idea that consequentialist moral theories form a family.

It seems, therefore, that consequentialism can be characterized as a family of moral doctrines. Now how can we use this result to devise a methodological approach for the case against consequentialism? We need, it seems, a clearer idea as to how the consequentialist family can be delimited. As a first step, it should hence be useful to conduct an inquiry into the logical structure of consequentialist doctrines. To this end, we should look towards a doctrine that is undoubtedly a version of consequentialism, *viz*. CU. The first step is, then, to

(i) Factorize classic utilitarianism into logically independent components, $C_{11}$, …, $C_{n1}$.

This first step will reveal a number of claims, $C_{11}$, …, $C_{n1,}$ that are *typically* involved in a consequentialist doctrine. This will *ipso facto* reveal the logical structure that is common to all consequentialist doctrines. That is, it will tell us the precise *number*, *n*, of components that is involved in a consequentialist doctrine. And it will tell us which *kinds* of components are contained in a consequentialist theory, *viz*. components that are of the same kind as $C_{11}$, …, $C_{n1}$, respectively.

In a next step, we can make use of a logical consequence of our assumption that consequentialism is a *family*. It implies that there have to be alternatives to every paradigmatic component. That is, for each component of CU, $C_{i1}$, there has to be at least one alternative component, $C_{i2}$. In order to understand the range of possibilities that consequentialism allows we should, therefore, investigate which non-standard alternatives there are to each of the paradigmatic components $C_{11}$, …, $C_{n1}$. In the second step, we should

(ii) Take stock of all alternatives, $C_{i2}$ ,…, $C_{im}$, to each of the paradigmatic components $C_{i1}$, $i=1, \ldots, n$.

Upon completing steps (i) and (ii), we end up, as it were, with a construction kit for consequentialist doctrines, as it is shown in the table below. It registers all components that consequentialists can embrace. The first column shows all the paradigmatic components, i.e. those of CU. Each row shows one paradigmatic component and all its alternatives. To construct a (any) consequentialist doctrine from the kit we simply choose one component from each row. If the rows are, in fact, logically independent, then it should be possible for consequentialists to combine every two components, $C_{ij}$ and $C_{kl}$ from two different rows *i* and *k* in a consequentialist moral theory.[16]

[15] This distinction has famously been emphasized by Rawls (1971/1999).

[16] Note that logical independence between rows should be distinguished from logical independence between components. Logical independence obtains between two components, $C_{ij}$ and $C_{kl}$, if endorsing (or rejecting) $C_{ij}$ does not necessitate endorsing (or rejecting) $C_{kl}$ and *vice versa*. Logical independence obtains between two rows, *i* and *k*, if and only if there is no component in row *i* that commits one to a component (or range of components) in row *k*. Note that the latter implies the former, but not *vice versa*. That is, the fact that two rows, *i* and *k*, are logically independent implies that any two components $C_{ij}$ and $C_{kl}$ in these rows are logically independent of one another. But the fact that two components $C_{ij}$ and $C_{kl}$ from rows *i* and *k* are logically independent does not imply that the rows themselves are logically independent.

3. Table: Construction Kit for Consequentialist Doctrines

| | | | | |
|---|---|---|---|---|
| $C_{11}$ | $C_{12}$ | ... | ... | $C_{1a}$ |
| $C_{21}$ | $C_{22}$ | ... | ... | $C_{2b}$ |
| ... | ... | ... | ... | ... |
| $C_{i1}$ | ... | $C_{ij}$ | ... | ... |
| ... | ... | ... | ... | ... |
| $C_{n1}$ | $C_{n2}$ | ... | ... | |

At this point, then, I can explain what a comprehensive case against consequentialism requires according to the *Family Resemblance Approach* (FRA). In order to show that consequentialism is an untenable moral view we need to show that there is a convincing and decisive objection to each possible combination of components that makes up a consequentialist moral theory. The question is how this can be done methodically. The family of consequentialist doctrines is very large.[17] So, obviously, going through all theories one by one is not an option. Here is a suggestion, however. It starts from the observation that every consequentialist doctrine, *C*, has to endorse exactly one component, $C_{ij}$, from each row. In order to show that all consequentialist theories are untenable we merely need to show, it seems, that all components in a given row are objectionable. This can be accomplished by climbing onto the shoulders of those who have already put forward convincing objections to (particular versions of) consequentialism. What we need to work out is which components these objections target. In a third step, we should, therefore,

(iii) Survey objections $O_1$, $O_2$, ...$O_o$ to consequentialism and correlate them with determinate components, $C_{ij}$.

After that, we can piece together a comprehensive case against consequentialism by combining objections that target components in a given row. So in the fourth and final step, we

(iv) Put together a set of objections $O=(O_1, ..., O_m)$ such that there is at least one objection, $O_l$, for all components, $C_{i1}, ..., C_{im}$, of a given row, $i$, $l=1, ..., m$.

It is not guaranteed, of course, that this four-step-procedure of the FRA will be successful. That will depend on whether or not there are, in fact, enough substantive arguments against consequentialism. But it is clear, at least, that by proceeding along the lines of this method we will not encounter the problems that are associated with the DM. As I explained above, the DM either omits versions of consequentialism, if the definition is narrow. Or it becomes impossible to formulate substantive criticisms, if the definition is broad. If applied properly, the FRA avoids both these problems. It avoids the first, because it takes into account all versions of consequentialism. Consequentialists are, hence, not able to use their strategy of interpretive divergence because, ideally, no version of consequentialism has been left out. It avoids the second problem too, because arguments do not have to rely on a vague and abstract idea that lies behind consequentialist doctrines. They can, rather, directly target concrete components of consequentialist doctrines.

[17] The following reasoning can get us a rough estimate of how large the consequentialist family is. CU, let us assume, is the combination of eight logically independent claims. (Sinnott-Armstrong (2011) offers a characterization of CU that contains even more independent claims, *viz*. eleven.) If each of the rows contains merely two components (which is certainly an underestimate), we would end up with $2^8$=256 individual moral theories.

## 3. Conclusion

In this paper, I have analysed the way in which the case against consequentialism is commonly made. As I explained in the first section, the conventional approach involves two steps. First, consequentialism is defined in terms of necessary and sufficient features. Then, criticisms are formulated on the basis of that definition. These criticisms are intended to show that all moral theories which possess the defining features of consequentialism are objectionable and should be rejected. This two-step method, I argued, runs into one of two difficulties. Substantive criticisms can be formulated, if a narrow definition is posited. But such a definition gives consequentialists the opportunity to defend themselves against criticisms using the strategy of interpretive divergence. They can object that their critic's definition is based on an impoverished understanding of consequentialism and that the objections proposed do not apply to a more appropriate interpretation of the creed. If, on the other hand, the definition is all-encompassing, its abstractness seems to make it impossible to come up with any substantive criticisms at all. I illustrated the problem by way of an example. I posited a common definition of consequentialism, formulated three well-known criticisms and showed how consequentialists can dismantle these criticisms using the strategy of interpretive divergence. As I argued further, the problem illustrated by these examples does not seem to be a specific phenomenon of the particular definition that I used in my example. If there was a defining feature of consequentialism that all consequentialist moral theories must embrace, we should detect it by looking towards the properties of a paradigmatic consequentialist theory, *viz*. CU. CU can be characterized as a maximizing, welfarist, hedonistic, aggregative, impartial doctrine. None of these qualities, however, seem to be defining features of consequentialism in general. The only aspect of CU that, indeed, seems to be shared by all other consequentialist moral theories is the very abstract idea that the moral status of an act depends only on the extent to which it promotes the good. But this opaque idea seems impossible to criticize. So the DM does, in fact, seem to run into either one of the two problems I have described.

In the second section, I argued that consequentialism should be regarded as a family of moral theories which are united – not by a single defining feature – but by a common structure. I explored the methodological consequences of this idea and suggested that a case against consequentialism should proceed in the following way. First, critics should determine the structure of a typical consequentialist doctrine, e.g. CU, by factorizing it into logically distinct moral claims. Then, they should consider which alternative components consequentialists may embrace in place of any of the paradigmatic components. Upon completing these two steps, critics possess a construction kit for consequentialist doctrines and can view the case against consequentialism in a new light. What critics are required to do is to show that all doctrines that can be constructed from the kit are subject to a decisive objection. This, I suggested, may be done by surveying objections to consequentialism, correlating them with individual components and by showing that all components of a given row are subject to a decisive objection. If successful, this would refute consequentialism as a whole, because all consequentialist doctrines must endorse at least one component from each row. This FRA is, as I have argued, superior to the conventional procedure of the DM, because it avoids both of the two problems that the latter unavoidably runs into. It makes it possible to address *all* consequentialist doctrines and, hence, does not give consequentialists the chance to defend themselves by diverging to a version of consequentialism to which criticisms do not apply. And it gives critics the chance to criticize consequentialist doctrines in terms of substantive content rather than bare logical structure.

In saying that the FRA solves these problems I do not mean to claim, of course, that it is itself free from problems. In this short piece, I have merely tried to explain why we need a new method for the critical study of consequentialism. And I have laid out the bare bones of what

is, I hope, a promising approach. My suggestion certainly raises new issues that need addressing. And there are surely many gaps in the reasoning that need to be filled. I have left aside, e.g., general moral-epistemological issues, such as the question what constitutes an objection to a moral doctrine anyway. I have not examined how it can be established that a given objection relates to a given component. Furthermore, there are certainly additional problems that still lie in the dark. The best way to deal with them is, I think, to apply the general approach that I have laid out and to see whether it gets us anywhere. I surely hope it does. But here, as anywhere in philosophy, the proof of the pudding is in the eating![18]

**Nikil Mukerji**

Technische Universität München
School of Education
Peter-Löscher-Lehrstuhl für Wirtschaftsethik
Marsstraße 20-22, 80335 Münche bzw.
Arcisstr. 21, 80333 München (postal address)
nikil.mukerji@tum.de

## References

Anscombe, G. E. M. 1958: 'Modern Moral Philosophy', *Philosophy* 33, 1–19.

Arneson, R. 2004: 'Moral Limits on the Demands of Beneficence?' in Chatterjee, D.K. (ed.) *The Ethics of Assistance. Morality and the Distant Needy*. Cambridge: Cambridge University Press, 33–58.

Barry, B. 1995: *Justice as Impartiality,* Oxford: Clarendon Press.

Brink, D. O. 2006: 'Some Forms and Limits of Consequentialism' in Copp, D. (ed.) 2006: *The Oxford Handbook of Ethical Theory*. Oxford: Oxford University Press, 380–423.

Broad, C. D. 1971: 'Self and Others' in Broad, C. D.; and D. R. Cheney (eds.): *Broad's Critical Essays in Moral Philosophy*. London: Allen and Unwin, 262–282.

Broome, J. 2004: *Weighing lives.* Oxford: Oxford University Press.

Feldman, F. 2006: 'Actual Utility, The Objection from Impracticality, and the Move to Expected Utility', *Philosophical Studies* 129, 49–79.

Forster, M. 2010: 'Wittgenstein on Family Resemblance Concepts' in Ahmed, A. (ed.) 2010: *Wittgenstein's Philosophical Investigations. A Critical Guide*. Cambridge: Cambridge University Press, 66–87.

Godwin, W. 1793: *An Enquiry Concerning Political Justice and Its Influence on General Virtue and Happiness*. London: G. G. J. and J. Robinson.

Hare, R. M. 1981: *Moral Thinking. Its Levels, Method and Point*. Oxford: Clarendon Press.

Hirose, I. 2004: 'Aggregation and Numbers', *Utilitas* 16, 62–79.

Hooker, B. 2003: *Ideal Code, Real World: A Rule-Consequentialist Theory of Morality*, Oxford: Oxford University Press.

— 2009: 'The Demandingness Objection' in Chappell, T. D. J. (ed.): *The Problem of Moral Demandingness. New Philosophical Essays.* Hampshire: Palgrave Macmillan, 148–162.

Hurka, T. 1992: 'Consequentialism and Content', *American Philosophical Quarterly* 29, 71-78.

Jeske, D. 2008. 'Special Obligations' in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Fall 2008 Edition).* http://plato.stanford.edu/entries/special-obligations.

[18] I would like to thank Matthew Braham, Thomas Kaczmarek, Julian Müller, Julian Nida-Rümelin, Martin Rechenauer and Geoffrey Sayre-McCord for their generous comments.

Kagan, S. 1998: Normative Ethics. Boulder: Westview Press.

McNaughton, D.; and P. Rawling. 1991: 'Agent-Relativity and the Doing-Happening Distinction', *Philosophical Studies* 63, 167-185.

Mendola, J. 2006: *Goodness and Justice. A Consequentialist Moral Theory*. Cambridge: Cambridge University Press.

Moore, G. E. 1903/1959. *Principia Ethica*. Cambridge: Cambridge University Press.

Mukerji, N. 2013: 'Utilitarianism' in Lütge, C. (ed.) 2013. *Handbook of the Philosophical Foundations of Business Ethics, Vol. 1*, Dordrecht: Springer, 297–312.

Mulgan, T. 2001: *The Demands of Consequentialism*. Oxford: Clarendon Press.

Nida-Rümelin, J. (1993): *Kritik des Konsequentialismus*. München: Oldenburg Verlag.

Portmore, D. W. 2005: 'Combining Teleological Ethics with Evaluator Relativism. A Promising Result', *Pacific Philosophical Quarterly* 86, 95–113.

— 2007: 'Consequentializing Moral Theories', *Pacific Philosophical Quarterly* 88, 39–73.

— 2011: *Commonsense Consequentialism. Wherein Morality Meets Rationality*, New York: Oxford University Press.

Rawls, J. 1971/1999: *A Theory of Justice (Revised Edition)*, Cambridge, MA: Harvard University Press.

Ridge, M. 2005: 'Review of Pleasure and the Good Life by Fred Feldman', *Mind* 114, 414–417.

Scheffler, S. 1982/1994: *The Rejection of Consequentialism. Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*, Oxford: Clarendon Press.

Sen, A. K. 1979: 'Utilitarianism and Welfarism', *The Journal of Philosophy* 76, 463–489.

— 2009: *The Idea of Justice*. Cambridge, MA: Harvard University Press.

Singer, P. 1979/1993, *Practical Ethics (Second Edition)*. Cambridge: Cambridge University Press.

Sinnott-Armstrong, W 2001, 'What is Consequentialism? A Reply to Howard-Snyder', *Utilitas* 13, 342–349.

— 2011: 'Consequentialism' in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition)*. http://plato.stanford.edu/entries/consequentialism.

Slote, M. A. 1984: 'Satisficing Consequentialism', *Proceedings of the Aristotelian Society* 58, 139–163.

Sluga, H. 2006: 'Family Resemblance', *Grazer Philosophische Studien* 71, 1–21.

Sumner, L. W. 1987: *The Moral Foundation of Rights*. Oxford: Clarendon Press.

Thomson, J. J. 1976: 'Killing, Letting Die, and the Trolley Problem', *The Monist* 59, 204–217.

Williams, B. A. O. 1973, 'A Critique of Utilitarianism' in Smart, J. J. C.; B. A. O. Williams (ed.) 1973: *Utilitarianism – For and Against*. Cambridge: Cambridge University Press, 75–155.

— 1981: 'Persons, Character and Morality' in *Moral Luck: Philosophical Papers 1973-1980*. Cambridge: Cambridge University Press, 1–19.

Wittgenstein, L. 1953/1986: *Philosophical Investigations. Translated by G.E.M. Anscombe*, Oxford: Basil Blackwell.

Wittgenstein, L. 1960: *Preliminary Studies for the "Philosophical Investigations". Generally Known as the Blue and Brown Books*. New York: Harper.

Woodruff, P. 2010: 'Plato's Shorter Ethical Works' in in E. N. Zalta (ed.) *The Stanford Encyclopedia of Philosophy*, (Summer 2010 Edition), URL = http://plato.stanford.edu/entries/plato-ethics-shorter.

# Moralischer Zufall und Kontrolle

Julius Schälike

Der Aufsatz befasst sich mit den unterschiedlichen Formen des moralischen Zufalls. Ich vertrete die These, dass lediglich konstitutiver Zufall möglich ist, ergebnis- und situationsbezogener hingegen nicht. Moralischer Zufall liegt vor, wenn Faktoren, die sich der Kontrolle des Akteurs entziehen, seine moralische Qualität beeinflussen. Die Möglichkeit moralischen Zufalls ist unvereinbar dem Kontrollprinzip, dem zufolge lediglich Faktoren, die ein Subjekt kontrolliert, auf seine moralische Qualität „abfärben". Ich versuche zu zeigen, dass das Kontrollprinzip aufgegeben werden muss. Das Fehlen von Kontrolle in Bezug auf Resultate führt dazu, dass die Resultate nicht aussagekräftig sind für die moralische Qualität des Akteurs. Das Fehlen von Kontrolle unterminiert jedoch nicht immer eine aussagekräftige Relation. Mentale Zustände wie Entscheidungen und auch Handlungs*dispositionen sind* Ausdruck der moralischen Identität eines Subjekts und sind somit auch aussagekräftig – ganz gleich, ob das Subjekt ihre Genese kontrolliert. Die Frage ist lediglich, ob Entscheidungen eine zusätzliche, unabhängige Rolle spielen, oder ob sich ihre Relevanz auf das Epistemische beschränkt. Im Anschluss an Hume argumentiere ich, dass flüchtige Ereignisse wie Entscheidungen nur dann moralisch auf den Akteur abfärben, wenn ihnen etwas Persistierendes – eine Disposition – korrespondiert. Daraus aber folgt: es gibt allein *konstitutiven* Zufall.

## 1. Einleitung

Zwei LKW-Fahrer fahren fahrlässig etwas zu schnell; der eine hat jedoch das Pech, dass ihm ein Kind vor das Fahrzeug läuft, das an den Unfallfolgen stirbt, während der zweite Fahrer sein Ziel unfallfrei erreicht.[1] Beide werden nicht nur strafrechtlich ganz unterschiedlich behandelt. Wir beurteilen sie auch moralisch nicht gleich. Und auch sie selbst tun dies nicht. Während der eine ruhigen Gewissens bleibt, wird der andere von Schuldgefühlen gepeinigt. Und dies erscheint uns auch vollkommen angemessen. Würde der Unfallfahrer lediglich mit dem Bedauern eines Außenstehenden reagieren, ergänzt durch den Hauch eines schlechten Gewissens, das sich angesichts der leichten Geschwindigkeitsübertretung einstellen mag, so wären wir empört. Mit dem zweiten Fahrer, der Glück hat, verfahren wir viel großzügiger, wir erwarten von ihm keine heftigen Schuldgefühle, denn wir halten seine Schuld für gering.

Vielleicht ist es angemessen, die beiden Fahrer unterschiedlich zu behandeln. Dies gilt etwa, wenn man an Schmerzensgeld denkt. Da der zweite Fahrer kein Leid verursacht hat, besteht kein Anlass, ihm Schmerzensgeld abzuverlangen. Wie aber verhält es sich mit der moralischen Schuld? Verdient der glückliche Fahrer ein milderes moralisches Urteil als der unglückliche? Hat er weniger Anlass für Gewissensbisse?

Scheinbar nein: Beide Akteure, so nehmen wir an, haben exakt die gleichen intrinsischen Eigenschaften, haben also den gleichen Charakter, sie vollziehen die gleichen Körperbewegungen, ihre mentalen Zustände sind gleich. Dass es im einen Szenario zum Unfall kommt, hat der Fahrer nicht vorausgesehen. Es stand nicht unter seiner Kontrolle. Wie sollte es dann einen Einfluss auf seine moralische Qualität haben? Sind beide Fahrer nicht gleich schlecht? Verdienen sie nicht das gleiche Maß an Kritik? Wenn man dem unglücklichen Fahrer das Auftauchen des Kindes auf der Strasse anlastete, also ein Ereignis, das mit ihm im Grunde nichts zu tun hat – warum sollte man ihn dann nicht auch für beliebige andere

[1] Williams 1976.

negative Ereignisse tadeln, etwa für den Tsunami vor Fukushima? Steht er nicht zu diesem Ereignis in einem in relevanter Hinsicht ganz ähnlichen Verhältnis wie zu dem Erscheinen des Kindes? Analysiert man diese Beispiele, so scheint sich ein Prinzip herauszukristallisieren, das *Kontrollprinzip*:

**Kontrollprinzip**

Nur Faktoren, die jemand kontrolliert, beeinflussen seine moralische Qualität, liefern Gründe für Lob und Tadel.

Das Kontrollprinzip ist jedoch nicht so unproblematisch, wie es zunächst den Anschein hat. De facto genügen unsere moralischen Urteile diesem Prinzip nämlich, wie wir gesehen haben, keinesfalls. Das Problem, das sich hier abzeichnet, kennt man als das Problem des moralischen Zufalls. Was versteht man unter moralischem Zufall? Und warum stellt er ein philosophisches Problem dar?

Moralischer Zufall ist die Übersetzung von moral luck. Leider ist moral luck ein irreführender Begriff. Es geht nicht um luck im Sinne von Glück oder Pech oder auch – allgemeiner gesprochen – von Zufall, sondern es geht um Kontrolle. Thomas Nagel definiert moralischen Zufall wie folgt:

**Moralischer Zufall**

Faktoren, die sich der Kontrolle des Akteurs entziehen, beeinflussen seine moralische Qualität.[2]

Nicht immer, wenn Zufall vorliegt, fehlt Kontrolle,[3] und nicht immer, wenn kein Zufall vorliegt, besteht Kontrolle.[4] Moralischer Zufall ist somit ein terminus technicus: es geht der Sache nach nicht um Zufall, sondern um Kontrolle.

Das philosophische Problem des moralischen Zufalls besteht nun darin, dass eine sehr grundlegende Intuition in Spannung zu einem höchst plausiblen Prinzip zu stehen scheint. Zum einen ist es angesichts von Beispielen wie dem der LKW-Fahrer überaus einleuchtend, dass moralischer Zufall möglich ist. Zum anderen lässt sich nicht leicht eine Begründung hierfür finden. Denn das Kontrollprinzip ist äußerst plausibel: Nichts, was sich unserer Kontrolle entzieht, kann unsere moralische Qualität prägen. Erst Kontrolle stiftet ja eine Verbindung zwischen uns und etwas anderem, durch die dieses andere aussagekräftig für unsere moralische Qualität sein kann, und dies schließt moralischen Zufall aus. Die Frage ist also: Lässt sich die Intuition, dass moralischer Zufall möglich ist, in einer Weise rechtfertigen, die im Einklang mit plausiblen Prinzipien bezüglich der moralischen Qualität steht? Muss das Kontrollprinzip womöglich ersetzt werden? Die Tragweite des Problems ist nicht zu unterschätzen. Sollte sich das Kontrollprinzip als unaufgebbar, moralischer Zufall als inexistent erweisen, wäre ein großer Bereich der Moral obsolet. Jedes Urteil über die moralische Qualität von Personen müsste als gegenstandslos angesehen werden. Dies liegt daran, dass die Suche nach Faktoren, die tatsächlich unter der Kontrolle eines Akteurs stehen und damit gemäß dem Kontrollprinzip als Grundlage für die Beurteilung seiner moralischen Qualität dienen könnten, notwendig erfolglos verlaufen würde. Denn betrachten wir erneut die beiden Fahrer. Sie haben zwar nicht unter Kontrolle, ob ein Kind vor ihr Fahrzeug läuft,

---

[2] Nagel 1976.

[3] Tendenziell bedroht der Zufall zwar Kontrolle. Robert Kane hat jedoch zu zeigen versucht, dass ein Ereignis zufällig (indeterminiert) eintreten kann, ohne dass der Akteur, der es verursacht, die Kontrolle verliert (Kane 1998, Kap. 8).

[4] So kann man in Bezug auf die eigene Identität nicht von Zufall sprechen (Rescher 1993). Es ist kein Zufall, dass ich die Eigenschaften habe, die mich ausmachen, dies ist vielmehr notwendig; hätte ich andere Eigenschaften, wäre ich ja ein anderer. Dies impliziert jedoch nicht, dass ich *kontrollieren* kann, welche Eigenschaften mich konstituieren. Charaktereigenschaften sind typischerweise willentlich kaum veränderbar.

wohl aber – so scheint es –, ob sie absichtlich zu schnell fahren. Stellen wir uns jedoch einen dritten LKW-Fahrer vor, der in charakterlicher Hinsicht den anderen beiden gleicht. Auch er würde fahrlässig zu schnell fahren, wenn sich ihm die Gelegenheit böte. Aber sie bietet sich ihm nicht, da sein Motor nicht anspringt, mit der Folge, dass er nicht zu schnell fährt. Dass die beiden anderen die Absicht bildeten, zu schnell zu fahren, stand dann jedoch auch nicht unter ihrer Kontrolle, denn auch bei ihnen hätte etwas dazwischen kommen können, es war einfach Glück bzw. Pech, dass es nicht geschah. Und dies gilt für alle Faktoren, die man als Basis für das moralische Urteil heranziehen könnte. Es gilt etwa für Charakterdispositionen. Manchmal mögen wir die Absicht bilden, unseren Charakter zu verändern, aber auch bei der Bildung dieser Absicht hätte etwas dazwischen kommen können, sodass man uns diese Absicht – und damit auch den durch sie erzeugten Charakterzug – mangels Kontrolle nicht zugute halten kann. Hier zeigt sich Folgendes: Nimmt man das Kontrollprinzip ernst, so entgleitet einem der Bezugspunkt der moralischen Bewertung einer Person. Urteile über die moralische Qualität wären unmöglich und damit ein zentraler Aspekt der Moral gegenstandslos.

## 2. Drei Formen moralischen Zufalls

Auf das Problem des moralischen Zufalls sind unterschiedliche Reaktionen möglich. Man könnte das Kontrollprinzip akzeptieren und eine massive Revision unseres Selbstbildes und unserer moralischen Praxis in Kauf nehmen: Personen ließen sich moralisch nicht beurteilen.[5] Will man diese Revision vermeiden, so muss man das Kontrollprinzip ablehnen und den moralischen Zufall akzeptieren. Wie wir gesehen haben, schließen sich das Kontrollprinzip und der moralische Zufall aus: Die These, moralischer Zufall sei möglich, ist genau das, was das Kontrollprinzip bestreitet. Aber welche Formen des moralischen Zufalls sind möglich?

Die Liste der Kandidaten enthält drei Einträge: 1. Bei resultatebezogenem Zufall (resultant luck) geht es um die unkontrollierten Handlungsfolgen. Fließen sie ins moralische Urteil ein, so lassen sich die ersten beiden Fahrer ungleich zu beurteilen, da ihr Handeln unterschiedliche Folgen hat. 2. Situationsbezogener Zufall (circumstantial luck) betrifft die Auswirkungen der unkontrollierten Handlungssituationen. Erinnern wir uns an den dritten LKW-Fahrer, dessen Motor nicht anspringt. Seine Handlungssituation ist anders, und das hat zur Folge dass er nicht zu schnell fährt. Hält man ihm dies zu Gute, so könnte er, obgleich sein Charakter den gleichen Fehler hat, als besser gelten als die beiden anderen. 3. Konstitutiver Zufall (constitutive luck) fokussiert auf unkontrollierte Prozesse, die zur Bildung der Eigenschaften beitragen, die uns als moralische Subjekte ausmachen, insbesondere unsere Charakterdispositionen. Konstitutiver moralischer Zufall liegt vor, wenn diese Dispositionen unsere moralische Qualität auch dann beeinflussen, wenn wir sie nicht kontrolliert erzeugt haben. Alle drei LKW-Fahrer müssten, da sie charakterlich gleich sind, gleich beurteilt werden.[6]

Ich möchte im Folgenden prüfen, welche Gründe für und gegen die einzelnen Formen sprechen. Was sie voneinander unterscheidet, lässt sich anhand der Rolle darlegen, die sie den Handlungen beimessen. Für resultatebezogenen Zufall sind auch unabsichtliche Handlungsfolgen für die moralische Qualität eines Akteurs relevant. Für situationsbezogenen Zufall zählen nur absichtliche Handlungen, genauer gesagt: Handlungsentscheidungen.

[5] Zimmerman (2002) akzeptiert das Kontrollprinzip und plädiert für weitgehende Revisionen, übersieht jedoch m.E., wie weit diese gehen müssten.

[6] Manchmal wird kausaler Zufall (*causal luck*) als vierte Variante genannt (Nagel 1976, 35): Zufall bezüglich der Determination durch antezedente Umstände. Diese Zufallsform stellt jedoch lediglich eine Komposition von situativem und konstitutivem Zufall dar.

Angewendet auf das Beispiel der LKW-Fahrer bedeutet dies, dass die beiden Fahrer, die sich dazu entschieden haben, zu schnell zu fahren, moralisch gleich zu beurteilen sind. Der Umstand, dass der eine Fahrer in einen Unfall verwickelt wird, spielt keine Rolle. Der dritte Fahrer, dessen Motor nicht anspringt, profitiert moralisch somit von seinem Glück, denn da er faktisch die Entscheidung, zu schnell zu fahren, zwar getroffen hätte, aber nicht getroffen hat, verdient er ein milderes Urteil. Für konstitutiven Zufall sind Handlungen irrelevant. Charaktereigenschaften, die uns durch Glück oder Pech zugefallen sind, beeinflussen unsere moralische Qualität.

### *2.1 Situationsbezogener und ergebnisbezogener Zufall*

Ist es plausibel, den situationsbezogenen Zufall zu akzeptieren, den ergebnisbezogenen jedoch zu leugnen? Prima facie spricht einiges dafür: Die Resultate des Handelns sind, sofern sie nicht vorhergesehen werden, nur sehr schwach mit dem Akteur verbunden. Sicher, er hat sie verursacht, aber da die Kontrolle fehlte, haben sie doch moralisch nichts mit ihm zu tun. Wenn die Mafia ohne mein Wissen die Bremsen meines Autos zerstört hat, mit der Folge, dass ich einen Unfall verursache, kann man mir keinen Vorwurf machen.[7]

Anders verhält es sich bezüglich der Frage, ob mentale Einstellungen wie Entscheidungen und Absichten dem Subjekt zugerechnet werden können. Auch dann, wenn sie insofern nicht unter der Kontrolle des Subjekts stehen, als ihre Existenz abhängt von unkontrollierbaren Faktoren wie dem Anspringen eines LKW-Motors, sind sie doch Ausdruck des Subjekts, seiner praktischen Identität.[8]

Hier zeichnet sich ab, dass das Kontrollprinzip eine irrige Verallgemeinerung einer Einsicht darstellt, die sich am LKW-Fahrer-Beispiel zeigte: Der unglückliche Fahrer verdient nicht deshalb das gleiche Urteil wie der glückliche, weil der Unfall nicht unter seiner Kontrolle stand, sondern weil das Fehlen von Kontrolle in Bezug auf Resultate dazu führt, dass die Resultate nicht aussagekräftig sind für die moralische Qualität des Akteurs. Das Fehlen von Kontrolle unterminiert jedoch nicht immer eine aussagekräftige Relation. Mentale Zustände wie Entscheidungen sind Ausdruck der moralischen Identität eines Subjekts und sind somit auch aussagekräftig – ganz gleich, ob das Subjekt sie kontrolliert. Es ist somit nicht ad hoc, sondern gut begründet, ergebnisbezogenen moralischen Zufall zu leugnen, situationsbezogenen Zufall jedoch zu akzeptieren. Was tatsächlich relevant ist, ist nicht Kontrolle, sondern eine moralisch aussagekräftige Relation zwischen moralischer Qualität und bestimmten Faktoren wie Handlungsresultaten oder mentalen Zuständen. Äußere Ereignisse stehen nur unter der Bedingung von Kontrolle in dieser Relation, wobei Kontrolle hier in einer ganz schwachen, unproblematischen Weise zu verstehen ist: Steuerung der Ereignisse durch den Willen, wie sie bei absichtlichem Handeln gewährleistet ist. Innere Ereignisse und Zustände wie Entscheidungen und Absichten bedürfen der Kontrollbedingung nicht, um ein moralisches Licht auf das Subjekt zu werfen.

---

[7] So auch Adam Smith: „Die Folgen [...] sind sogar – wenn das möglich ist – noch belangloser für Lob und Tadel als die äußere Körperbewegung. Da sie nicht vom Handelnden abhängen, sondern vom Zufall, können sie nicht die angemessene Grundlage für eine Empfindung abgeben, deren eigentliche Gegenstände Charakter und Verhalten des Handelnden sind“ (1790, 148).

[8] Cf. Kant: „Der gute Wille ist nicht durch das, was er bewirkt, oder ausrichtet, [...] sondern allein durch das Wollen, d. i. an sich, gut [...]. Wenngleich durch eine besondere Ungunst des Schicksals, oder durch eine kärgliche Ausstattung einer stiefmütterlichen Natur, es diesem Willen gänzlich an Vermögen fehlete, seine Absicht durchzusetzen; wenn bei seiner größten Bestrebung dennoch nichts von ihm ausgerichtet würde, und nur der gute Wille (freilich nicht etwa ein bloßer Wunsch, sondern als die Aufbietung aller Mittel, soweit sie in unserer Gewalt sind) übrig bliebe: so würde er wie ein Juwel doch für sich selbst glänzen, als etwas, das seinen vollen Wert in sich selbst hat.“ Cf. auch Smith 1790, 148: „Auf die Absicht oder auf die Gesinnung [...] muss sich in erster Linie alles Lob und Tadel [...] richten“; Thomson 1993, 199: „Whatever we do, our doing of it is no more to our discredit than are those purely mental acts by which we do it“; ähnlich Glover 1970, 64–66.

Wir können das Kontrollprinzip somit durch das Expressionsprinzip ersetzen:

> **Expressionsprinzip**
>
> Nur Faktoren, die Ausdruck der moralischen Identität eines Subjekts sind, beeinflussen seine moralische Qualität, liefern Gründe für Lob und Tadel. Subjektexterne Ereignisse sind Ausdruck der moralischen Identität eines Subjekts, wenn das Subjekt sie durch ihren Willen kontrolliert.

### *2.2 Konstitutiver Zufall*

Was spricht für die Möglichkeit von konstitutivem Zufall? Konstitutiver Zufall liegt vor, wenn Handlungsdispositionen, etwa Tugenden und Laster, die moralische Qualität von Personen beeinflussen, und zwar unabhängig von ihrer Aktualisierung und obwohl sie nicht unter der Kontrolle des Subjekts stehen. Wenn konstitutiver Zufall möglich ist, dann gilt: Auch wenn jemand faktisch gar nichts Schlechtes tut, ist er moralisch genau so kritikwürdig wie jemand, der dies tut, wenn es wahr ist, dass er es täte, wenn er könnte. Für die moralische Qualität ausschlaggebend wären Handlungs- bzw. Entscheidungsdispositionen, nicht Handlungen oder Entscheidungen; letztere wären lediglich als Symptome dieser Dispositionen bedeutsam. Diese Position wird von Aristoteles akzeptiert:

> Die Taten sind Zeichen einer inneren Haltung, denn wir loben wohl auch einen, der nichts vollbracht hat, wenn wir annehmen können, er habe die Disposition dazu.[9]

Konstitutiver Zufall ist offensichtlich aus den gleichen Gründen möglich wie situativer Zufall. Die Dispositionen eines Subjekts stehen in einer ähnlich engen Relation zu seiner moralischen Qualität wie seine Entscheidungen. Wenn die Dispositionen zur Konstitution der Identität des Subjekts zumindest beitragen, so auch zu seiner moralischen Identität und somit auch zu seiner moralischen Qualität. Eine engere Relation als die zwischen Dispositionen und moralischer Qualität, so scheint es, ist gar nicht möglich. Wie Judith Jarvis Thomson bemerkt, wäre es unsinnig zu sagen, jemand habe einen miserablen Charakter, sei aber ein durch und durch guter Mensch.[10] Auch in seinen Dispositionen drückt sich die moralische Qualität eines Subjekts aus, sodass sie unter das Expressionsprinzip fallen. Die Frage scheint nur noch zu sein, ob Handlungen bzw. Entscheidungen eine zusätzliche, unabhängige Rolle spielen, oder ob sich ihre Relevanz auf das Epistemische beschränkt. Sind faktische Entscheidungen lediglich Indizien für Entscheidungsdispositionen? Dann gäbe es, genauer betrachtet, gar keinen situativen, sondern allein konstitutiven Zufall. Alle drei LKW-Fahrer wären moralisch gleich zu bewerten, da sie alle die Disposition haben, zu schnell zu fahren, auch wenn nur zwei von ihnen tatsächlich zu schnell fahren und nur einer einen Unfall verursacht.

## 3. Tadel, Handlungen und Charakter

David Hume hat sich im Treatise mit dem Problem auseinandergesetzt, wie es möglich ist, jemanden für ein so flüchtiges Phänomen wie eine Handlung zu tadeln. Dies ist, so Hume, dann angemessen, wenn der Faktor, der die Handlung verursacht hat, in der Person persistiert.

> Actions are by their very nature temporary and perishing; and where they proceed not from some cause in the characters and disposition of the person, who perform'd them, they infix not themselves upon him, and can neither redound to his honour, if good, nor infamy, if evil. The action itself may be blameable; it may be contrary to all the

[9] Aristoteles, *Rhet*. I.9, 1367b33f.; ähnlich Hume, *Treatise* 3.2.1, 477f.; 3.3.1., 575.
[10] Thomson 1993, 208.

> rules of morality and religion: But the person is not responsible for it; and as it proceeded from nothing in him, that is durable or constant, and leaves nothing of that nature behind it, 'tis impossible he can, upon its account, become the object of punishment or vengeance.[11]

Für diese Analyse spricht, dass sie es verständlich macht, warum jemandem Handlungen, die er unter Hypnose vollzogen hat, nicht zugerechnet werden können. Solche Handlungen entspringen nicht aus Dispositionen, die die Hypnose überdauern, und gehören deshalb nicht in einer Weise zum Subjekt, die ein evaluatives Licht auf es wirft, nachdem die Hypnose geendet hat. Verständlich wird auch, warum wir manchmal bereit sind, jemandem seine Taten nach einiger Zeit zu verzeihen. Dies erscheint gerechtfertigt, wenn die Person sich so verändert hat, dass die Dispositionen, in denen die Taten gründen, verschwunden sind.

Humes Insistieren darauf, dass man, wenn man einen Akteur moralisch beurteilt, auf ein persistierendes Merkmal Bezug nehmen muss, scheint mir unbestreitbar. Aber könnte diesem Erfordernis nicht auch dadurch Genüge getan werden, dass man auf die Kontinuität der personalen Identität hinweist? Diese kann ja auch dann gewährleistet sein, wenn sich einige Aspekte des Subjekts verändert haben – etwa seine okkurenten mentalen Zustände, seine Absichten, sein Wille. Dann könnte man einräumen, dass die mentalen Zustände, die die Handlung ausgelöst haben, zwar nicht mehr bestehen, und womöglich auch die zugrunde liegenden Dispositionen nicht; aber der Akteur ist noch dieselbe Person wie die, die die Handlung vollzogen hat. An diese persistierende personale Identität könnte man die moralische Zurechnung knüpfen. Die Zuschreibung hätte dann eine viel breitere Basis als im humeschen Modell. Zwar ist die personale Identität von den Dispositionen abhängig; wenn sich der Charakter vollständig änderte, wäre die Person nicht mehr die selbe, und die Zurechnung vergangener Handlungen wäre nicht mehr möglich. Aber wenn sich nur einige Dispositionen änderten, darunter die, aus denen die Handlung H hervorging, bliebe die personale Identität möglicherweise erhalten; H könnte dann zugerechnet werden.

Diese Argumentation halte ich nicht für überzeugend. Zwar trifft es zu, dass personale Identität auch über die Änderung von Eigenschaften hinaus fortbestehen kann, doch kommt es in manchen Kontexten nicht darauf an, ob sich die personale Identität geändert hat, sondern darauf, ob die relevanten Eigenschaften persistieren. Angenommen jemand ist kahlköpfig, hatte vor 20 Jahren jedoch volles Haar. Wenn es darum geht, zu seiner heutigen Frisur Stellung zu nehmen, ist es klarerweise völlig irrelevant, wie er vor 20 Jahren aussah. Der Umstand, dass es sich noch um dieselbe Person handelt, ändert nichts daran, dass es unangemessen wäre, zu sagen, er hätte heute volles Haar. Analoges gilt für das Handeln und das Entscheiden: Auch wenn man sagen könnte, dass eine gestrige Entscheidung gestern zum Subjekt gehörte, so „schmückt" diese Entscheidung das Subjekt doch heute so wenig, wie den Kahlen heute sein Haar schmückt, das er vor 20 Jahren besaß.[12] Freilich kann man auch fragen, wie die moralische Qualität des gesamten Lebens eines Subjekts zu beurteilen ist. Hier fallen auch vergangene Episoden ins Gewicht, und zwar auch dann, wenn die relevanten Dispositionen verschwunden sind. Aber solche Dispositionen können kein Grund sein, gut oder schlecht über die heutige moralische Qualität des Subjekts zu denken. Schlagend deutlich wird dies, wenn man sich vorstellt, eine Entscheidung sei durch eine Hypnose evoziert worden, die nicht so weitreichend war, dass sie die Kontinuität der personalen Identität unterbrochen hat. Nachdem die Hypnose endet, wird die Zuschreibung unangemessen, und dies lässt sich nicht unter Hinweis auf die personale Identität erklären, sondern allein unter Hinweis auf die Dispositionen.

---

[11] *Treatise* 2.3.2, 411. Cf. auch *Treatise* 3.3.1, 575.

[12] Cf. auch Lohmar 2005, Kap. VII; Sher 2006, Fn. 2, S. 35f.

Somit gilt: Entscheidungen, Handlungen und ihre Folgen sind lediglich Indizien für beurteilungsrelevante Faktoren, sie haben keine eigenständige Bedeutung für die moralische Qualität von Personen.

## 4. Kontrolle und Dispositionen höherer Ordnung

Resultate- und situationsbezogenen moralischen Zufall gibt es mithin nicht, konstitutionsbezogenen schon. Diese Behauptung ist nicht willkürlich oder inkohärent (nach dem Motto: „Wenn der Zufall moralisch irrelevant ist, dann muss er es überall sein"), da es ein gesichertes Kriterium gibt, wie der Einfluss des Zufalls einzuschätzen ist: Unterbricht er die vom Expressionsprinzip benannte Verbindung, die zwischen einem Ereignis oder einem Sachverhalt einerseits und einem Subjekt andererseits besteht und die es ermöglicht, dass das Ereignis oder der Sachverhalt moralisch auf das Subjekt „abfärbt"? Handlungen und ihre Konsequenzen sind nur dann auf diese Weise mit dem Akteur verbunden, wenn er sie kontrolliert, also der Zufall ausgeschaltet ist, wobei diese Kontrolle bei absichtlichem Handeln gewährleistet ist. Die eigenen Handlungsdispositionen hingegen erweisen sich zwar als kaum kontrollierbar, doch sie stehen in einer unmittelbaren Beziehung zum Subjekt, da sie es moralisch konstituieren. Hier bedarf es keiner Kontrolle, um eine moralisch aussagekräftige Beziehung herzustellen, da diese Beziehung immer schon und unaufhebbar besteht.

Nun ist es sicher so, dass jemand, der gegenwärtig nicht (leicht) dazu zu bringen ist, grausam zu sein, diese Disposition oftmals erwerben kann. Als Beispiel kann der Fall von Robert Harris dienen, den Gary Watson schildert: Harris war als Kind sensibel und sanftmütig, aber durch die Umstände wurde er zu einem grausamen Mörder. Eine Frage, die sich bei der Betrachtung dieses Falles stellt, ist: Wäre aus mir unter diesen Bedingungen nicht auch so jemand geworden?[13] Wenn die Antwort „ja" lautete, so hätte auch ich die Disposition, grausam zu sein. Wenn nicht der okkurente Wille, sondern die volitiven Dispositionen das Kriterium für moralische Urteile über Personen darstellen, warum sollte man mich dann nicht für einen grausamen Menschen halten?

Hierzu ist zu sagen, dass es einen Unterschied macht, ob man hier und jetzt bereit ist, unter bestimmten Bedingungen grausam zu handeln, oder ob erst umfangreiche, zeitlich ausgedehnte kausale Einflüsse erforderlich sind, um diese Bereitschaft zu erzeugen. Wer hier und jetzt nicht grausam wäre, auch wenn man ihm etwa eine hohe Belohnung in Aussicht stellte, der hat gegenwärtig nicht die Disposition der Grausamkeit. Er könnte jedoch die Disposition haben, diese Disposition zu erwerben; dann hätte er eine Disposition zweiter Ordnung. Jemand hat die Disposition zweiter Ordnung, x zu tun, wenn er, wenn bestimmte Bedingungen erfüllt wären, zu einem späteren Zeitpunkt die Disposition ausbilden würde, x zu tun.[14]

Dispositionen zweiter Ordnung sind nicht direkt aussagekräftig für die moralische Qualität; von ihnen hängt vielmehr ab, wie sich die moralische Qualität ändern kann. So kann man sagen, dass Robert Harris schon in jungen Jahren die Anlage hatte, grausam zu sein, doch damit ist nicht gesagt, dass er in seiner Kindheit schon so schlecht war, wie als Erwachsener. Entsprechend kann man von manchem Nichtschwimmer sagen, er habe die Disposition, gut

[13] „[T]he thought that if *I* had been subjected to such circumstances, I might well have become as vile [...] induces [...] a sense of equality with the other: I too am a potential sinner" (Watson 1987, 245).

[14] Hiermit modifiziere ich meine Position in Schälike 2010, 271; dort habe ich Dispositionen höherer Ordnung („Meta-Dispositionen") nicht temporal definiert, sondern sie dadurch von Dispositionen erster Ordnung unterschieden, dass bei diesen nur ein Schritt zur Aktualisierung erforderlich sei, bei jenen mindestens zwei. Ich danke Gilbert Scharifi für diesbezügliche Kritik.

zu schwimmen, da er fähig ist, gut schwimmen zu lernen; doch erst nachdem er es gelernt hat, wird man ihn als guten Schwimmer loben können.

Dispositionen zweiter Ordnung sind somit nicht direkt für die moralische Qualität relevant; indirekt geben sie hingegen durchaus oftmals Aufschluss über die moralische Qualität. Dispositionen zweiter Stufe bezüglich Charakterdispositionen sind nämlich typischerweise Funktionen der Charakterdispositionen: Je ausgeprägter etwa die Disposition des Mitleids ist, desto schwerer wird sie sich zerstören und durch eine Disposition zu Grausamkeit ersetzen lassen.[15] Wer somit die Disposition zweiter Stufe hat, sehr leicht die Disposition zu Grausamkeit zu erwerben, ist in der Regel weniger mitleidig als jemand, bei dem dies schwerer fällt, und darum kann er als schlechter gelten.

Folgt man meiner Argumentation, akzeptiert den konstitutionsbezogenen Zufall und bestreitet die anderen beiden Formen moralischen Zufalls, so hat dies Konsequenzen, die unplausibel anmuten, wie sich etwa an folgendem Beispiel Thomas Nagels zeigt. Angenommen jemand, der Offizier in einem Konzentrationslager war, hätte ein ruhiges und harmloses Leben in Deutschland geführt, wenn die Nazis nie an die Macht gekommen wären. Angenommen außerdem ein anderer, der ein ruhiges und harmloses Leben in Argentinien führte, wäre Offizier in einem Konzentrationslager geworden, wenn er Deutschland nicht 1930 aus geschäftlichen Gründen verlassen hätte.[16] Konsequenterweise muss man auch hier sagen, dass der faktische und der kontrafaktische Verbrecher gleich tadelnswert sind. Dies wirft, wie George Sher bemerkt, die Frage auf, ob das Bild, das wir uns über unsere eigene moralische Qualität und die der meisten anderen Menschen machen, nicht völlig inadäquat ist. Wer weiß schon, wie er sich verhalten hätte, hätte er unter der Nazidiktatur leben müssen? Hat die Position, für die ich argumentiert habe, die Konsequenz, dass wir alle „moralische Monster“ sind?[17] Käme diese Implikation einer *reductio ad absurdum* meiner Thesen gleich?

In der Tat wäre es sicherlich eine gravierende Korrektur unseres Bildes von uns selbst und fast allen anderen, wenn wir uns als moralische Monster sehen müssten. Ob und in welchem Grade diese Korrektur erforderlich ist, ist jedoch schwer zu sagen. Es ist ja gar nicht ausgemacht, dass tatsächlich alle zu allem fähig sind. Auch unter den Bedingungen einer Diktatur versagen längst nicht alle, und nicht alle, die versagen, versagen im gleichen Maße.

Auch wenn die charakterbezogene Sichtweise somit nicht in dem Maße revisionär ist, wie Sher argwöhnt, erweist sich unser Selbstbild doch in einem nicht unerheblichen Maße tatsächlich als illusionär. Diese Implikation aber betrachte ich nicht als eine *reductio* meiner Argumentation, sondern als eine unbequeme Wahrheit über uns.

---

[15] Dies ist allerdings im Fall von Dr. Jekyll und Mr. Hyde anders. Der kultivierte, tugendhafte Dr. Jekyll hat eine Substanz zu sich genommen, die dazu führt, dass er sich unabsichtlich immer wieder in den lasterhaften Mr. Hyde verwandelt. Diese Verwandlung resultiert nicht aus einer moralischen Defizienz der Dispositionen Dr. Jekylls – nehmen wir an, diese Dispositionen seien perfekt. Hier spricht Jekylls Disposition zweiter Stufe, zu Hyde zu werden, nicht gegen seine moralische Qualität, solange er Jekyll ist. Und auch wenn man annähme, die Einnahme der Substanz hätte die Disposition erzeugt, bei einem Fingerschnippen *jederzeit sofort* zu Hyde zu mutieren, würde dies Jekylls moralische Qualität nicht beeinträchtigen, jedoch aus dem besonderen Grund, dass Jekyll und Hyde zwei unterschiedliche Personen sind. Wenn das Fingerschnippen Jekyll nicht in eine andere Person verwandelte, sondern ihn nur grausam machte, so hätten wir es nicht mit einer Disposition zweiter, sondern erster Stufe zu tun, die sich negativ auf die moralische Qualität Jekylls auswirkte.

[16] Nagel 1976, 26.

[17] Sher 2006, 26.

## 5. Psychologische Nachbemerkung

Ergebnisbezogener und situativer moralischer Zufall existieren nicht, konstitutiver schon. Ist damit das Problem des moralischen Zufalls vollständig gelöst? Noch nicht ganz. Die Intuitionen wehren sich heftig gegen die Leugnung der beiden genannten Zufallsformen, wie sich am Beispiel der LKW-Fahrer deutlich zeigt. Wir fänden es unangemessen, wenn der Unfallfahrer eine ähnliche Haltung zu dem Unfall einnähme wie der glückliche Fahrer, der von dem Unfall in der Zeitung liest. Wir erwarten von ihm stärkere Schuldgefühle, als sie angesichts bloßer Geschwindigkeitsüberschreitung angemessen wären. Wie lässt sich das erklären?

Meine Vermutung ist folgende: Wir deuten die Ausbildung von Schuldgefühlen als Indikator für bestimmte moralische Einstellungen. Wer bei der Verletzung von moralischen Normen keinerlei Schuldgefühle verspürt, der hat diese Normen nicht internalisiert und ist deshalb ein schlechter Mensch. Allerdings ist der Mensch kein vollständig rationales Wesen. Manchmal reagiert er auf bestimmte Anlässe zu stark. Dies geschieht jedoch nicht vollkommen zufällig. Bestimmte Erlebnisse stoßen uns auf unsere Fehler mit einer emotionalen Wucht, die diese Fehler wie unter einer Lupe vergrößert erscheinen lassen. Dieser Vergrößerungseffekt muss bei dem Schluss von Symptomen auf Dispositionen berücksichtigt werden. Wenn der Unfallfahrer nicht mit viel stärkerer Bestürzung und größeren Selbstvorwürfen auf den Tod des Kindes reagierte als der glückliche Fahrer, dann läge das vermutlich daran, dass er die relevanten Normen nicht hinreichend internalisiert hat. Tatsächlich der Schuld angemessen sind die größeren Schuldgefühle zwar nicht, aber der Mensch ist teilweise eben auf systematische Weise irrational, und wir wissen das. Jemand, der die richtigen moralischen Dispositionen hat, wird mit übertriebenem Schuldgefühl reagieren, wenn er kausal in ein Unglück involviert ist. Wir schließen daraus, dass jemand angemessen reagiert, dass seine moralischen Dispositionen zu schwach ausgebildet sind, da wir annehmen, dass er ebenso irrational ist wie die meisten Menschen, sodass er, wenn er die richtigen Dispositionen hätte, überreagieren würde.

**Julius Schälike**

Fachbereich Philosophie
Universität Konstanz
julius.schaelike@gmail.com

## Literatur

Aristoteles (Rhet.): *Rhetorik*.

Glover, J. 1970: *Responsibility*. London: Routledge&Kegan Paul.

Hume, D. (*Treatise*): *A Treatise of Human Nature*, hg. von P. H. Nidditch. Oxford: Oxford University Press ²1978.

Kane, R. 1998: *The Significance of Free Will*. Oxford: Oxford University Press.

Kant, I. (GMS): *Grundlegung zur Metaphysik der Sitten*. In: *Kants Werke*. Akademie-Textausgabe. Berlin: de Gruyter, Bd. 4.

Lohmar, A. 2005: *Moralische Verantwortlichkeit ohne Willensfreiheit*. Frankfurt/M.: Klostermann.

Nagel, T. 1976: „Moral Luck“, in: ders., *Mortal Questions*. Cambridge: Cambridge University Press 1979, 24–38.

Rescher, N. 1993: „Moral Luck“, in: Statman 1993, 141–166.

Schälike, J. 2010: Spielräume und Spuren des Willens. Eine Theorie der Freiheit und der moralischen Verantwortung. Paderborn: Mentis.

Sher, G. 2006: *In Praise of Blame*. Oxford: Oxford University Press.

Smith, A. 1790: *Theorie der ethischen Gefühle* (übers. v. W. Eckstein). Hamburg: Meiner 2010.

Statman, D. (Hg.) 1993: *Moral Luck*, Albany: State University of New York Press.

Thomson, J. J. 1993: „Morality and Bad Luck", in: Statman 1993, 195–215.

Watson, G. 1987: „Responsibility and the Limits of Evil", in: ders., *Agency and Answerability*. *Selected Essays*. Oxford: Oxford University Press, 219–259.

Williams, B. 1976: „Moral Luck", in: Statman 1993, 33–55.

Zimmerman, M. J. 2002: „Taking Luck Seriously", in: *The Journal of Philosophy* 99, 553–576.

# What Makes Moral Values Queer?

Julius Schönherr

John Mackie's argument from moral queerness has traditionally been taken to raise doubts about the existence of moral properties, facts and values. This is based on Mackie's claim that, if moral properties existed, they would have to instantiate some property Q which would have to be "utterly different from anything else in the universe" (Mackie 1977: 38). Traditionally, Q has been taken to be the intrinsic reason-givingness of moral properties. In this paper, I will consider the possibility that the problematic feature is the unexplainable supervenience of moral properties on non-moral properties. Concerning this thesis, I will do two things. First, I will give an argument for why it is advantageous to focus the metaphysical argument for queerness on supervenience and not on moral properties' reason-givingness. Second, I will consider whether there is a compelling account of explanation that renders the queerness charge from supervenience true.

## 1. Introduction

John Mackie's argument from moral queerness has been traditionally taken to raise doubts about the existence of moral properties, facts and values. This argument is based on his claim that, if moral properties existed, they would have to be non-natural entities, "different from anything else in the universe" (Mackie 1977: 38). This standard take on the matter can be expressed by the following thesis:

**Queer properties**

Moral properties would have to be non-natural entities. Therefore, they are metaphysically queer.

In opposition to this traditional picture, I will, in this paper, provide an argument for why it is advantageous to focus the charge of moral queerness on the supervenience of moral properties on non-moral properties rather than on moral properties themselves. To do this, I will examine a version of the following alternative thesis from Mackie:

**Queer supervenience**

Non-natural moral properties would have to supervene on non-moral properties. Therefore, they are metaphysically queer.

But Queer supervenience alone, without further specification(s), can hardly count as an accurate expression of moral queerness. Baldness supervenes on the number of hairs, density supervenes on mass and volume, and, ideally, grades might even supervene on performance. Clearly, there is nothing queer about these supervenient properties and many philosophers have indeed employed supervenience as a distinct tool to render entities not queer. Therefore, I will turn to a refined version of Queer supervenience: Queer unexplainable supervenience, which is more promising version of the argument from moral queerness against the non-natural moral realist.

**Queer unexplainable supervenience**

Non-natural moral properties would have to unexplainably supervene on non-moral properties. Therefore, they are metaphysically queer.

However, despite its initial appeal, Queer unexplainable supervenience is likely to fail because no account of 'explanation' that would make this thesis true is available. Therefore, non-natural realism can, until further notice, escape the charge.

First, I will, in an exegetical effort, locate Queer unexplainable supervenience in the wider array of queerness charges as Mackie conceives of them. Next, I will give the argument for why the charge Queer supervenience is superior to Queer properties. Third, I will explain the notion of moral supervenience in more detail in order to, fourth, address two difficulties that arise when asking for an explanation of supervenience: On the one hand, choosing a demanding account of 'explanation' might leave property types other than moral properties unexplained; on the other hand, employing a lax account of 'explanation' might enable non-natural moral realists to buy into it. Either way, the moral realist would be able to escape the charge.

## 2. Locating Queer Supervenience: The Geography of the Argument from Moral Queerness

In his book *Ethics, Inventing Right and Wrong* Mackie advocates the so called "error theory". This theory holds that while moral judgment is truth apt, purporting to represent the world, it in fact never succeeds in doing so. Consequently, all moral judgments are systematically false. In judging the world to be a certain way morally, we are constantly victim to an error. This is because "there are no objective values" (Mackie 1977: 15).

One of Mackie's influential arguments in favor of that thesis is the so called 'argument from moral queerness', according to one metaphysical interpretation of which moral realism would be committed to entities "utterly different from anything else in the universe" – so different in fact that a commitment to these entities would render moral realism conclusively implausible.[1] This metaphysical argument is two-pronged. With regard to the first prong, Mackie raises worries with regard to moral values *themselves*:

> If there were objective values, then they would be entities or qualities or relations of a very strange sort, utterly different from anything else in the universe. (Ibid: 38)

Famously, according to Mackie, what makes these moral values so "utterly different from anything else in the universe" is, on the one hand, the fact that they would provide anyone who knows about them with an "overriding motive" for action and, on the other hand, the fact that these objective moral values would have to have "to-be-pursuedness somehow built into them" (Ibid: 40). This latter feature introduces the idea that moral values would have to be intrinsically normative which, in the very least, means that values provide agents with at least *some*[2] *good*[3] *reason* for action. The intrinsic reason-givingness of objective

[1] Mackie takes the force of the argument from moral queerness to be rather strong. In his eyes, the argument renders the existence of moral properties *impossible*. However, a more moderate solution is plausible. While an entity's being utterly different from anything else makes a commitment to such entities metaphysically pricy, the existence of these entities is thereby not rendered impossible. If this were the case, then nobody could ever rationally believe in the existence of an ontologically extravagant entity; which seems false.

[2] The argument does not require the strong thesis that moral reasons always have to be *overriding* reasons. While overriding reasons outweigh all other reasons, moral obligations, on a weak reading, might only *contribute* to an overall judgment. A contributing reason can, however, be overpowered by other reasons, leading to an action that in the end defies the contributing reason's suggestion - in this sense, it has been overridden.

[3] Good reasons are normative reasons that speak in favor of something. They stand opposed to explanatory reasons, which are reasons that explain people's behaviors. Explanatory reasons answer the questions: why did she do X? Such reasons are usually pairs of desires and means-end beliefs. On the other hand, good reasons answer the question: why ought she to do that? (See. Lenman 2009)

moral properties has, in recent literature, become the standard interpretation of the metaphysical interpretation of the argument from moral queerness.[4]

Regarding the second prong, Mackie raises worries with regard to how these putative moral values *supervene* on natural[5] features:

> Another way to bring out this queerness is to ask, about anything that it is supposed to have some objective moral quality, how this is linked with its natural features. What is the connection between the natural fact that an action is a piece of deliberate cruelty—say, causing pain just for fun—and the moral fact that it is wrong? It cannot be an entailment, a logical or semantic necessity. Yet it is not merely that the two features occur together. The wrongness must somehow be "consequential" or "supervenient"; it is wrong because it is a piece of deliberate cruelty. But just what in the world is signified by this "because"? (Ibid: 41)

I think that there are at least three interesting points to be gleaned from this (and the preceding) passage(s). First, according to Mackie, it is not *merely* the fact that moral properties supervene which makes them queer but, rather, the fact that this supervenience cannot be explained by, say, "entailment, a logical or semantic necessity". Cases of supervenience which turn out to be semantic phenomena do not strike us as in any way as extravagant, nor do they seem unique. It is, for instance, not odd that a bachelor cannot cease to be an unmarried man without also ceasing to be a bachelor, because being unmarried is semantically entailed by the description "bachelor". On the contrary, the supervenience of moral properties on natural properties, widely held to be a true consequence of Moore's open question argument, cannot be explained in semantical terms (see Moore 1912: 58–61).

Second, Mackie's qualms about supervenience seem to depend on his doubts about moral properties themselves.[6] Arguably, it is precisely *because* moral properties are intrinsically reason-giving that their supervenience cannot be explained. If these values weren't so queer in the first place, we might, after all, have no trouble explaining why they supervene. Reductive forms of moral realism do not obviously encounter this problem. Naturalist reductivists for instance claim that moral properties (goodness, say) are *identical* to their descriptive[7] supervenience bases (happiness, say). And since everything depends on itself, the dependence of moral properties on their descriptive basis would be no mystery. The target of the argument from queerness is not any old form of moral realism. Rather, the argument targets non-reductive types of normative realism (also known as non-natural moral realism), because these forms of realism hold that *normative* properties supervene on *non-*

---

[4] See e.g. Michael Smith, *Beyond the Error Theory* p. 2; Bart Streumer 2011, 3.

[5] In philosophy the base properties relevant to moral supervenience have been characterized as descriptive properties (Jackson), natural properties (Sturgeon), non-moral properties (Railton). In his 2007 paper *Anti-Reductionism and Supervenience* Michael Ridge provides a compelling analysis of these differences and he points to problems with all of them. Ridge argues that reductive naturalists such as Jackson can't draw the contrast between moral and non-moral because they believe that they are the identical. Choosing *the natural* as the supervenience base ties in neatly with the aim to reconcile moral supervenience with a naturalist worldview (see. Horgan 1993).However, conceptually there is no reason that the moral way things are should not depend on non-natural things such as gods commandments. In this sense the natural would be too narrow of a base. However, keeping these problems in mind, I will, to make for an easier read, in this paper refer to the moral supervenience base as "non-moral".

[6] This is just to say that the supervenience worry is not independent from one's account of moral properties. It not to say that *one must have* a certain account of moral properties in order to get a supervenience problem. It is, for instance, silent about whether moral expressivists – philosophers who don't believe that there are moral properties at all – can explain supervenience.

[7] See footnote 6 for why the reductive naturalist specify the supervenience base as descriptive and not as natural.

*normative* properties. While normative properties tell us what we have reason to do, non-normative properties do not. As such, moral properties and non-moral properties really are, as Michael Ridge notes, "distinct existences" (Ridge 2008). Therefore, it might prima facie seem puzzling and in need of explanation that there should be necessary connections (as supervenience tells us) among them.

A third point to learn from Mackie is that in rendering unexplained supervenience queer, one adduces a *reason* for why it is implausible; it is not the lack of explanation per se that is implausible but, rather, that this lack of explanation is queer, i.e. it is unprecedented.

Mackie notes that realists[8] only have a problem explaining supervenience if they are committed to moral properties being *intrinsically normative*; that is, whenever the realist has a problem with queer moral properties, she might have a corresponding problem with queer moral supervenience. Though they are two sides of the same coin, metaphysical arguments from moral queerness have traditionally focused on moral properties themselves and not on supervenience[9]. In the next chapter, I will give an argument for why it is advantageous to formulate queerness based on supervenience and not, as it is traditionally conceived, based on moral properties.

## 3. Motivating Queer Supervenience: A Methodological Argument in Favour of the Focus on Moral Supervenience

In this section I will present a methodological argument for why it is advantageous to base the argument form moral queerness on supervenience of moral properties, rather than on moral properties themselves.

*The* important aspect of queerness is that the object under scrutiny is "utterly different from anything else in the universe". I want to argue for a methodological constraint on this difference: being different ought not to be trivially true. Here is an example of something's being trivially different from anything else: gold is different from anything else in the universe, because it is the only metal with the atomic number of 79. The reason why this is trivial is that, upon knowing what kind of thing gold is, we already know that it is the only metal with that atomic number. Being a metal with the atomic number 79 just *is* being gold[10]. This constraint can be formulated as follows:

(i) X's being utterly different from anything else ought not to be formulated such that the difference is trivial.

As I have said, gold's having the atomic number 79 makes it trivially different from anything else; this is because it is *sufficient* to make it gold. This sufficiency condition for being trivially different from anything else can be stated as follows:

(i)C X will be trivially different from everything else if the feature Y that accounts for the difference is a sufficient X-maker.

For moral properties that would mean:

---

[8] Expressivists don't believe that there are moral properties at all. Foremostly realists have noted that expressivists equally have a problem explaining supervenience. I don't wish to take a stance on this issue here.

[9] See for instance Michael Smith, Beyond the Error Theory p. 2; Bart Streumer 2011, 3.

[10] The formulation "what kind of thing X is" cannot be replaced by "understanding the term gold". Before it was discovered that gold has a certain atomic structure, many did understand the word "gold" properly. But, it is also true that they did not know what kind of thing it is.

(i)P Moral properties will be trivially different from anything else if the feature that accounts for the difference, the intrinsic reason-givingness, is sufficient to make something a moral property.

For moral supervenience it would mean:

(i)S Moral properties will be trivially different from anything else if the feature that accounts for the difference, the supervenience of the moral on the non-moral, is sufficient to make something a moral property.

Here is the explanation of why (i) is reasonable: If some feature (Y) is a sufficient X-maker, it will be simply inconceivable – ruled out to begin with – that something has Y while still not being X. One could take anything one likes (*anything*!) and adduce a sufficient different-making feature Y. In the case of gold, having the atomic number 79 is such a sufficient condition. Here is another example:

(A) The existence of volcanoes is queer because they are the only openings, or ruptures, in a planet's surface or crust, that allow hot magma, ash and gases to escape from below the planet's surface.

Obviously (A) is false. Volcanoes are not queer; they are a normal geological structure. But it is true that they are the only openings in the planet's surface that allow hot magma, ash and gases to escape from below the surface. This is because being an opening in a planet's surface that allows magma, ash and gases to escape just *is* being a volcano. It is clear that via this method, we could render queer anything we want. That, however, is not what the argument from queerness should be about. Here is an example of something that, intuitively, really is queer:

(A)* The existence of witches is queer because they are the only women who can inflict harm via magical powers and fly on brooms.

In all probability, (A)* is true. Witches really are queer. But, like the volcano example, it is also true that every woman that can inflict harm via magic powers, by definition, just *is* a witch. Both phenomena, witches and volcanoes, will, trivially, come out different from anything else. But only witches are queer. Therefore, it turns out that the weakness of trivial difference is not that it could not possibly track queer features, but rather that expressing the fact that something is different from everything else in the style of (i)C *does not help us identify these features*. Being different from anything does not get the argument going anymore.

Let's apply this to the moral case. It turns out that the formulation of queerness for moral properties, formulation (i)P, might suffer this methodological weakness. Is being intrinsically reason-giving sufficient for being a moral property? The case is at least debatable. Think of any, however absurd, belief about intrinsic reason for action, say, the intrinsic reason to open as many doors as possible. Is this a *false moral view*, or is it *not a moral view at all*? I find this hard to decide. If this case, as well as all cases of intrinsic-reasons for action, is a moral view and turns out to be sufficient to make something a moral property, then (i)P would suffer the methodological weakness sketched out above. The possibility that anything else other than moral properties could have intrinsic reason-giving properties would be trivially ruled out.

The supervenience approach is superior in this regard. We would by no means say that the fact that a property strongly supervenes on another property is sufficient to make it a moral property. Supervenience is a formal relation among properties that is not restricted to any domain whatsoever. Therefore, we actually could (with some optimism) wander around the universe – to use Mackie's metaphor – searching for other properties that unexplainably supervene on lower level properties; but, so goes Mackie's diagnosis, we wouldn't find any.

However, the fact that we would not find any would not be ruled out by definition; it could, conceivably, occur anywhere where there is causation. But it just doesn't.

## 4. A More Precise Account of Moral Supervenience

Moral properties supervene on descriptive properties.[11] On the most basic level, this means that "things cannot differ with respect to some moral characteristic unless there is some natural property with respect to which they differ" (Kim 1984: 153 - 176).[12] One famous distinction Kim makes is between weak and strong supervenience. Unlike weak supervenience, strong supervenience makes the dependence of the higher level properties on the lower level properties *necessary*, whereas weak supervenience does not. The strong version, as will shortly become clear, is the type of supervenience relevant for *moral* properties. Therefore, I will restrict talk about supervenience in this paper to strong supervenience, which consists in the following claim:

> A strongly supervenes on B just in case, *necessarily*, for each x and each property F in A, if x has F, then there is a property G in B such that x has G, and *necessarily* if any y has G, it has F. (My italics) (see Ibid: 165)

This can be expressed in this formula of modal predicate logic:

**SS**

$\Box\forall x\forall F in A[Fx\rightarrow\exists G in B(Gx\wedge\Box\forall y(Gy\rightarrow Fy))]$

Kim defines supervenience not as a 'characteristic' (as Hare did), but rather as a *relation* among families of properties.[13] The relevant families for the moral case are the family of moral properties (A) and the family of non-moral properties (B). The supervenience relation among property families (A and B) is defined in terms of the dependence of individual properties (individual Fs for A-properties and individual Gs for B-properties).

Furthermore, Kim, in his definition, talks about the dependence of properties instantiated in objects (x). The variable "x" can be assigned any bearer of A- and B-properties. One such bearer might, for instance, be a particular action. If x has a moral property, say, that of being right, then it has a particular non-moral property and, necessarily, any action that has the same non-moral property will also be right. The range of x, however, does not need to be restricted to actions; it could be any non-moral state that fixes a moral property (such as attitudes, states and actions embedded in meticulously described circumstances).

The formulation "if any y has G, it has F" in Kim's definition states that each non-moral base property (G) determines a supervenient property (F). So if, for instance, a certain

---

[11] Not all philosophers construe supervenience as a relation among *properties*. Non-cognitivists, for instance, don't believe that there are moral properties at all. Non-cognitivistsargue that supervenience is a requirement of moral discourse (Timmons, Horgan 1992, 231). Therefore, for these philosophers, there are no objective moral properties that can supervene on anything at all. In this paper, I will restrict myself to the property talk of supervenience.

[12] Famously, R. M. Hare introduced the notion of supervenience to metaethical philosophy:"First, let us take that characteristic of "good" which has been called its supervenience. Suppose that we say "St. Francis was a good man." It is logically impossible to say this and to maintain at the same time that there might have been another man placed exactly in the same circumstances as St. Francis, and who behaved in exactly the same way, but who differed from St. Francis in this respect only, that he was not a good man." (Hare 1952, 145)

[13] There might not be a clear difference between all of these terms, but there are some uncontroversial examples of moral terms and descriptive terms. The term "loyalty" might be blurry in that it is largely descriptive but adds a moral connotation. But other terms such as "right", "wrong", "good, and "bad" are clearly moral terms, referring to moral properties. They supervene on descriptive properties such as being the intentional infliction of pain or being a lie. (see Jackson 1998)

action has the moral quality of being wrong based on the fact that it has a certain non-moral property (such as being an intentional killing in circumstance C), then each action in the exact the same circumstances determines the moral quality of being wrong.

Kim's formulation of strong supervenience includes two necessity operators. The modal force of these operators is, in principle, debatable. Nick Zangwill tailors a convincing interpretation of these operators when applied to the moral case. He interprets the first operator as *conceptual* necessity and the second one as *metaphysical* necessity (see Zangwill 1995: 374). The amended formula for moral supervenience should therefore be:

**SS***

$\Box_c\forall x\forall F\text{in}F[Fx\rightarrow\exists G\text{in}G(Gx\wedge\Box_M\forall y(Gy\rightarrow Fy))]$

Where "$\Box_C$" stands for "it is conceptually necessary that" and "$\Box_M$" stands for 'it is metaphysically necessary that'.

Why does the second operator, according to Zangwill, stand for metaphysical necessity? Let me explain: Some philosophers think that favoring our friends and family is morally permissible. They might also think that favoring our fellow countrymen is not morally permissible. Others might have thought this to be the other way round, opting for country first and friendship second. In this case, we would judge there to be substantial normative disagreement *within morality* between the two parties. Only one party can be right. And at least one of them is wrong. But being wrong, in this case, does not mean that one has committed any conceptual mistake, i.e. it does not mean that one does not know what morality is about. (see Ibid: 280ff.) Nevertheless, we would say that if one party were in fact right and, say, the friendship-first position were correct, then this would be a necessary truth: It could not possibly be otherwise,[14] even though it is conceivable that it could have been otherwise. And those who thought it permissible to favor countrymen over friends *did* conceive of that option. Zangwill interprets the second necessity operator as indicating metaphysical necessity because it is *possible to conceive* of alternative ways the world might have been morally; but despite this, the world could *not possibly be* different in moral respects than it actually is.[15]

Why does the first necessity operator, according to Zangwill, stand for conceptual necessity? It is an interesting observation that, among metaethicists, the conceptual necessity of supervenience is broadly accepted: Independently of which things are in fact right, we know solely by understanding the concept of morality that moral truths are necessarily true. Upon understanding the concept of morality, we have a "conceptual grasp that there are some metaphysical necessities in the offing – although we may not know which" (Zangwill 1995: 374). Imagine Pete, who knows that Mary lied when she told her husband that she had worked all day. Imagine that he does not quite know whether it was right or wrong of her to do that. This by itself would not be a conceptual mistake. Now, imagine that Pete thinks that even if lying, in this case, is the right thing to do, *it* could just as well have been wrong (without a change in the circumstances). He does not know that in morality, truths are necessary. Such a view presumes that moral value is not conceptually bound by non-moral properties. Intuitively, most philosophers would assume that Pete did not understand what

[14] Consequentialists might argue that, the question of whether the friendship first or the country first position is right depends on the context, which determines which of both tends to maximize the good. Therefore, they might argue that it is not true that each of these positions is either necessarily true or necessarily false. But even for consequentialists, the necessity holds if one embeds the principles in contexts. The consequentialist would not argue that, given a stable context, either position could be necessarily true or necessarily false.

[15] Zangwill remarks: "Which metaphysical necessities obtain is, as it were, conceptually contingent. But even though it is not built into our concepts that everything which is an instance of causing-pain-for-fun is evil, it may be necessary all the same." (Zangwill 1995: 374)

morality is about. Morality loses its point if it is not fixed by actions, states of affairs, or attitudes. If morality did not require moral properties to be fixed by non-moral properties, then it would allow for moral properties to be free floating, detached from the non-moral world. Furthermore, it would not be conceptually guaranteed that morality gives as reliable a verdict about what we ought to do. Therefore, Jackson calls this part of supervenience "the most salient and least controversial part of folk moral theory" (Jackson 1998: 118) and Michael Smith writes, "Everyone agrees that the moral features of things supervene on their natural features [...]. For recognition of the way in which the moral supervenes on the natural is a constraint on the proper use of moral concepts" (Smith 1994: 21).

Why are the two necessity operators critical in the context of this paper? The *metaphysical necessity* claim purports to describe part of the behavior of moral properties, i.e. how they relate to descriptive properties. It is this part of moral supervenience that stands in need of explanation. As I showed in the first paragraph, Mackie argues that if these necessities remain unexplained, they would be queer. Second, *conceptual necessity* is about what sort of relation morality incontrovertibly *requires*. Mackie's error theory, for example, embraces the conceptual constraint arguing that it is an integral part of morality for moral properties to be supervenient (he argues that moral value "would have to be" supervenient). But, he denies that unexplained metaphysical necessities exist, as the existence of such a relation would be queer and implausible. To sum up, unexplained metaphysical necessity is the target of the queerness argument and the conceptual necessity secures its importance.

To remind ourselves, from the formula for strong moral supervenience the embedded metaphysical necessities – $\Box_{\mathbf{M}}\forall y(Gy \rightarrow Fy)$ – stand in need for explanation. Suppose there is something (say, X) that explains these necessities. Then, an explanation for those necessities could be expressed as follows:

**EXP**

X explains that it is metaphysically necessary that everything that is G is also F.

But EXP could mean either of two things. In one sense, X might explain why being G necessitates the instantiation of a particular property F and not that of another property such as E (where F and E are not identical). In another sense, X might explain why something's being G necessitates anything at all and not just nothing. In this latter case, X would explain why a certain relation obtains, namely that of necessitation.

## 5. Explaining Supervenience: Two Accounts

Any metaethical account is conceptually committed to moral supervenience and normative realists in particular to supervenience being a relation between distinct types of properties. However, the charge goes that if these necessities remain unexplained, they are queer, e.g. they are utterly different from anything else in the universe and a commitment to these entities is therefore implausible.

Turning this uneasiness into a serious queerness argument against normative realism requires finding a plausible account of 'explanation' that explains all remaining cases of supervenience but which leaves moral supervenience in particular as an unexplained queer residue. The rest of the paper will be concerned with pointing to two difficulties in meeting this challenge: First, if one were to choose a demanding account of explanation, one might manage to leave the moral case unexplained. But such an account might also leave other types of facts or properties such as mental properties unexplained. Hence, if we do not want to give up our commitment to these other properties, moral supervenience could, though unexplained, escape queerness. On the other hand, if one chooses an account of 'explanation'

that is too lax, one might end up with something that the moral realist can readily accommodate. In this case, again, supervenience would escape queerness.

### *5.1 Conceptual Entailment – a Demanding Account of Explanation*

One thought, most closely associated with Simon Blackburn, is that the explanatory failure of moral supervenience consists in the fact that the metaphysical necessities implied by supervenience are not accompanied by corresponding epistemic necessities; that is, even though the instantiation of some non-moral properties necessitate the instantiation of some moral properties, there remains an epistemic "lack of entailment" from the former to the latter. Put in yet another way, no knowledge of non-moral facts automatically endows rational agents with knowledge of the moral facts. Blackburn describes the lack of entailment as follows:

> There is no moral proposition whose truth is entailed by any proposition ascribing naturalistic properties to its subject. (Blackburn 1971: 118)

And he points out that this stands in conflict with the metaphysical necessities implied by moral supervenience:

> If A has some naturalistic properties, and is also good, but its goodness is a distinct further fact not following from its naturalistic features, and if B has those features as well, then it follows that B also is good. And this is a puzzle for the realist, because there is no reason at all, on his theory, why this should follow. (Ibid: 119)

The reason that Blackburn implies is lacking is epistemic entailment. If it is possible to conceive of a single non-moral state N with either some moral value M or, alternatively, with M*, we do not have an explanation for why it is exactly one of these that is, in fact, necessitated by N. And if it is possible to conceive of N without any further moral property at all, we do not have an explanation for why N necessitates any moral. Earlier, when talking about supervenience, I have shown that there are moral necessities that are not also conceptual necessities. Therefore, I take Blackburn's claim to be on target. In the guise of the formulation EXP from above, we can render the failure to explain moral supervenience as follows:

> **EXP***
>
> In the moral case, conceptual necessities do not explain that it is metaphysically necessary that everything that is G is also F.

But is this a problem unique to morality? In the philosophy of mind there has been a famous corresponding debate about whether the instantiation of physical properties such as brain states necessitate the instantiation of mental properties despite a lack of epistemic entailment. This debate has been famously revolving around the so-called zombie thought experiment.

A zombie is a physical duplicate of a normal person, which lacks (at least some) phenomenal experiences but which is physically and functionally identical to this person (see Chalmers 1996: 84). For instance, when I sniff a juicy lemon there is a particular experience that I undergo; but, when my zombie-twin sniffs that lemon he does not experience any smell. However, if I said that this lemon smells fresh, my zombie twin would surely say the same thing and it would react just as that normal person did. The only difference would be that the zombie would not have experienced any smell. Some philosophers have argued that these zombies are epistemically possible without being metaphysically possible, i.e. they say that we

an coherently imagine a detailed world in which such zombies exist, but that, nevertheless, they are not possible.[16]

If these philosophers were right, then moral facts and mental states would seem very much on par with regard to the 'lack of entailment' and it would seem difficult to render moral supervenience queer for this 'lack of entailment'.

In contemporary philosophy of mind, it is a moot point as to whether zombies are conceivable and possible, conceivable or possible, or neither. And as long as the debate isn't decided, the normative realist can, it seems, adopt a wait-and-see attitude about the issue. Shafer-Landau takes this cue and defends his version of non-reductive moral realism exactly in this way:

> The problem, then, should be that competent speakers of a language can conceive of a world in which the base properties that actually underlie particular moral ones fail to do so. But there is no mystery here, since people can conceive of many things that are not metaphysically possible. If certain base properties metaphysically necessitate the presence of specified moral properties, then the conceptual possibility that they fail to do so reveals only a limitation on our appreciation of the relevant metaphysical relations. There is no deep explanatory puzzle resisting resolution here. (Shafer-Landau 2004: 86)

On the one hand, I think that Shafer Landau is right: as long as it is a moot issue as to whether "people can conceive of many things that are not metaphysically possible" this 'lack of entailment' argument seems somehow ill-suited to bringing out the queerness of moral supervenience. On the other hand, Shafer-Landau's assessment of the situation is almost too optimistic. It is far from clear whether people can conceive "many things" that aren't metaphysically possible. After all, there seems to be something about the subjective experience of mental states that makes them *especially* problematic. Whether the analogy between the moral and the mental can withstand critical examination would require a thorough analysis of the zombie thought experiment debate. This is outside the scope of this paper.

### *5.2 Explanation in Terms of the Base – A Minimal Requirement on Explanation*

In *The Impossibility of Superdupervenience*, Michael Lynch and Joshua Glasgow put forward an interesting argument defending the idea that supervenience among properties can never be explained in a satisfying way. They ask us to imagine a supervenience relation, such as the supervenience of moral properties/facts (call them B-facts) on non-moral properties/facts (call them A-facts), and then they ask us to imagine that this supervenience can be explained by some further fact; call this explainer fact an 'S-fact'. Now they point to the following dilemma:

> Either (i) these S-facts themselves supervene on the A-facts or (ii) they do not. We shall argue that the nonreductive materialist cannot claim that S-facts supervene on the A-facts on pain of a regress. This leaves (ii). If S-facts don't supervene on A-facts then either (iia) the S-facts are members of the set of A-facts; or (iib) the S-facts are sui generis. We shall claim that neither (iia) nor (iib) is a plausible option for the nonreductive materialist. If so, then the S-facts cannot be explained in a materialistically respectable way. Superdupervenience is impossible. (Lynch and Glasgow 2002: 208f)

Before discussing this passage, I would like to point out that (ii) and (iia) do not go together well because, if S-facts were a member of A facts, then they would be trivially supervenient on

[16] See Kirk 1974 & Chalmers 1996, 93–171.

them. Hence, if (iia) were true, (ii) would be false. However, this glitch will not destroy their argument; it simply requires refiling (iia) under (i).

Here is how I understand their argument: Regarding option (i), consider the fact '*B supervenes on A*' and suppose now that S explains this fact, thereby creating the new fact "*S explains that B*' *supervenes on A*'. Call this latter fact the S-fact and suppose that the S-fact is itself supervenient on (and non-identical with) A, thereby producing the fact "*S explains that B supervenes on A*' *supervenes on A*'. If we are entitled to ask what explained the first supervenience fact, we should now be equally entitled to ask what explains this new supervenience fact. If we invoke a new supervenient explainer S* we could, in turn, ask what explains it. Therefore, it seems that as long as explainers themselves supervene, there will always remain an unexplained supervenience fact.

Now, consider the case (iib). If the explaining S-facts are neither identical to some base facts nor supervenient on them, they can only be sui generis. Here is an example: We take it that moral facts (and properties) supervene on non-moral ones. Religious people, for instance, might say that it is God who makes it the case that B-facts supervene on A-facts. God, however, is neither part of the moral supervenience base nor is he supervenient on it. But invoking non-natural, non-supervenient entities such as Him to explain supervenience makes things even queerer than they were in the first place. After all, supervenience theses were supposed to provide a recipe to tie higher level entities to lower level entities, thereby making them more palatable. If this attempt to provide a tie between higher and lower level entities invites, at the same time, a further untied entity, the original intent of posing supervenience theses has clearly failed.

The only solution left in the Lynch/Glasgow taxonomy is (iia). They also dismiss this option for the reason that "S-facts could be a type of A-facts for any value of "A"; for, how could the facts which connect the B-facts to the A-facts be A-facts themselves? Intuitively, the metaphysical S-facts must be connecting facts, and therefore ontologically distinct from the facts they are connecting" (Ibid: 210). To me, the idea that base facts connect themselves to higher-level facts does seem puzzling, though not necessarily wrong. Though it would be nice to have some more elaboration on this thought, Lynch and Glasgow do not provide further arguments for why base facts cannot also be such connecting facts. In the end, their reasoning is that neither of the options for explaining supervenience is satisfactory; therefore they conclude:

> **NO EXP**
>
> Nothing explains that it is metaphysically necessary that everything that is G is also F.

Dissatisfied with not explaining the necessitation relation involved in supervenience, Karen Bennett has recently argued that this relation – which she calls 'grounding relation' – must be explainable in terms of the lower level entities (e. g. facts and properties) alone[17].

> Grounding, I claim, is not (only) internal, but superinternal. A superinternal relation is one such that the intrinsic nature of only one of the relata — or, better, one side of the relation — guarantees not only that the relation holds, but also that the other relatum(a) exists and has the intrinsic nature it does. [...] Everything is settled by the base, by the first relatum(a).(Bennett 2011: 35f)

As mentioned earlier, explaining necessitation requires two things: First, explaining why the relation obtains at all, and second, explaining why *exactly these* properties are necessitated.

[17] Louis DeRosset in *Grounding Explanations* has argued for a tight connection between grounding and the explanation of necessitation relations. He states: "All proponents of grounding agree that grounding relates facts, and that the facts that ground a fact are the facts that explain it." (DeRosset 2012, 5) In this paper, I'll take the connection between grounding and explanation for granted.

According to the quoted passage, it is the base entities themselves that provide answers to both questions. As a result, we may formulate Bennett's conclusion as follows:

**BASE EXP**

*Being G* explains that it is metaphysically necessary that everything that is G is also F.

Her reasoning is this: Take an arbitrary grounding/necessitation fact, say the fact '*A grounds B*', i.e. '*B exists in virtue of A*', i.e. '*A necessitates B*'. Bennett argues that if there is nothing in virtue of which this fact is true (in her terms: if there is nothing that grounds this grounding fact) then it is fundamental, i.e. it is not true in virtue of anything. However, this is implausible because, if the fact that '*B exists in virtue of A*' is fundamental, then so are the objects that are involved in that fact.[18] Consider the fact, '*Berlin exists in virtue of an ensemble of buildings and people*'. If this grounding fact were fundamental, so would be the entities that compose that fact; hence, it would follow that Berlin is a fundamental object, which is an absurd conclusion. Therefore, (at least some) necessitation facts cannot remain unexplained. An explanation for 'A grounds B' invokes a new entity, e.g. X, thereby generating a new fact '*X grounds* '*A grounds B*''.[19] Since we have already decided that these grounding facts are not fundamental, this new grounding fact would, on its part, needs something (say, Y) in virtue of which it is true. This generates a regress. If these explainers (X, Y, ...) are *new* entities at each level of explanation, the regress is vicious, leading to an infinite series of novel entities. In order to avoid this conclusion, Bennett assumes that the invoked explainers are not new entities at each level. Rather, they are the base entities themselves. Hence, the fact that '*B exists in virtue of A*' would be true in virtue of A alone and the fact that ''*B exists in virtue of A*' *exists in virtue of A*' would also be true in virtue of A alone. In terms of the Lynch/Glasgow framework, we might say that Bennett accepts (iia) – she believes that the explainers are part of the base themselves.

The dialectical situation at this point is this: It might be the case that Lynch and Glasgow are right, in which case supervenience could never be explained in a satisfying way. In this case, the queer supervenience charge against the moral realist would be spurious. But it might also be that Bennett is right, and explanations of supervenience have to take the form of BASE EXP. In this case we could ask: Could the non-reductive normative realist embrace BASE EXP or would she have to reject it? I can think of no argument that would successfully deny her embracing it. I certainly *want* to say that, as a property dualist, she should not be allowed to claim that the non-moral based facts explain the nature of the moral facts, and that the necessitation relation holds. However, it is question-begging to employ her dualism to support the idea that she cannot explain necessitation when, originally, the lack of explanation was supposed to undermine her dualism. We cannot forge an argument against dualism and, at the same time, support it with the assumption that dualism is false; this begs the question.

## 6. Conclusion

Normative realists hold that moral properties (or facts) and non-moral properties (or facts) are ontologically distinct. Furthermore, they hold that moral properties strongly supervene on non-moral properties, which entails the view that the instantiation of some non-moral properties necessitates the instantiation of moral properties. This view stokes uneasiness: if non-moral properties and moral properties really are distinct, then the moral realist cannot

[18] She supports this thought with a principle put forward by Ted Sider according to which "the fundamental truths involve only fundamental notions" (Sider 2011, 126ff).

[19] Alternatively, one could formulate 'B exists in virtue of A' and ''B exists in virtue of A' exists in virtue of X'.

tell us *why* the one necessitates the other, i.e. she cannot explain that the one supervenes on the other.

However intuitive, the attempt to express this uneasiness by reformulating it in terms John Mackie's argument from moral queerness proves to be problematic. There does not seem to be a straightforward account of 'explanation' that normative properties fail to satisfy, and that all other strongly supervenient properties do satisfy. On the one hand, employing conceptual entailment as a form explanation runs the risk of also rendering mental properties queer, and on the other hand, the theory of grounding which requires supervenience to be explained in terms of the base properties alone can readily be embraced by the normative realist.

Despite these difficulties, the normative realist's defense is built on somewhat shaky ground. First of all, ultimately, rescuing moral supervenience by referring to mental supervenience makes the plausibility of normative realism hostage to developments in a totally unrelated branch of philosophy. After all, philosophers might come to the conclusion that zombies are either not conceivable and not possible, or both conceivable and possible. In both cases, the normative realist would lose her companion. This is all the more plausible because supervenience is not a conceptual requirement on mental properties, therefore leaving more conceptual space for supervenience to fail in the mental case.

Concerning the second way to explain supervenience, metaphysical grounding has only very recently been a focus of metaphysicians. And while most philosophers agree that grounding facts have to be grounded in base-fact alone, they have not yet tackled the question whether grounding can serve as a touchstone to separate those instances of metaphysical necessitation that do hold from those that do not hold; an answer to this question will likely lead to a reevaluation of whether moral supervenience is metaphysically queer.

But, for now, the normative realist has escaped the charge of queer moral supervenience.

**Julius Schönherr**

Humboldt-Universität zu Berlin
Schoenherrjulius@gmail.com

## Bibliography

Bennett, K. 2011: 'By Our Bootstraps'. *Philosophical Perspectives* 25 (1), 27-41.

Blackburn, S. 1971: 'Moral Realism'. In J. Casey (eds.): Morality and Moral Reasoning.

Chalmers, D. 1996. 'The Conscious Mind: In Search of a Fundamental Theory', Oxford: Oxford University Press.

deRosset, L. forthcoming: 'Grounding Explanations'. *Philosophers' Imprint*.

Horgan, T. & Timmons, M. 1992: 'Troubles on Moral Twin Earth: Moral Queerness Revived', Synthese 92 (2), 221 – 260.

Kim, J. 1984: 'Concepts of Supervenience', Philosophy and Phenomenological Research 45, 153-76.

Kirk, R. 1974: 'Zombies vs Materialists', *Proceedings of the Aristotelian Society* 48, 135-52.

Levine, J. 1993: 'On Leaving Out What It's Like'. In: Martin Davies & Glyn W. Humphreys (eds.), Consciousness: Psychological and Philosophical Essays. Blackwell.

Lynch M. P. & Glasgow J. 2003: 'The Impossibility of Superdupervenience', Philosophical Studies 113 (3), 201-221.

Mackie, J. L. 1977: Ethics: Inventing Right and Wrong, Penguin.

Moore, G. E. 1903: Principia Ethica, Dover Publications.

Ridge, M 2007: 'Anti-Reductionism and Supervenience'. Journal of Moral Philosophy 4 (3), 330-348.

Sider, T. 2011: Writing the Book of the World. Oxford University Press.

Shafer-Landau, R. 2003: Moral Realism: A Defense. Oxford: Oxford University Press.

Zangwill, N. 2002: 'Moral Supervenience'. In: J. Kim (ed.)" Supervenience, International Research Library of Philosophy, Ashford: Dartmouth Publishing Company.

# Konsequentialistische Theorien und der Besondere-Pflichten-Einwand

Marcel Warmt

Ein häufiger Einwand gegen konsequentialistische Theorien ist der Besondere-Pflichten-Einwand. Dieser besagt, dass konsequentialistische Theorien unplausibel (kontraintuitiv) sind, weil sie keine besonderen Pflichten generieren können, obwohl besondere Pflichten bzw. die daraus resultierenden Handlungen moralisch wünschenswert sind. Es ist das Ziel dieses Artikels, zu zeigen, dass auch Konsequentialisten eine Form von besonderen Pflichten – quasi-besondere Pflichten – anerkennen können. Diese Pflichten sind stark genug, um den Besondere-Pflichten-Einwand zu entkräften. Die Widerlegung des Besondere-Pflichten-Einwands wird im Wesentlichen in der Rekonstruktion einer konsequentialistischen Zwei-Ebenen-Theorie à la R. M. Hare und einer sich daraus ergebenden Handlungscharakterisierung bestehen.

## 1. Einleitung

Ein häufiger Einwand gegen konsequentialistische Theorien ist der Besondere-Pflichten-Einwand. Dieser besagt, dass konsequentialistische Theorien unplausibel (kontraintuitiv) sind, weil sie keine besonderen Pflichten generieren können, obwohl besondere Pflichten bzw. die daraus resultierenden Handlungen moralisch wünschenswert sind. Unter einer besonderen Pflicht verstehe ich, im Gegensatz zu einer allgemeinen Pflicht, diejenige Pflicht, die man gegenüber Personen auf Grund einer besonderen Beziehung oder vorausgehenden Handlung hat. Unter einer konsequentialistischen Theorie verstehe ich eine Theorie, die die moralische Richtigkeit und Falschheit von Handlungen ausschließlich aufgrund der (wahrscheinlichen) Handlungsfolgen beurteilt.[1]

Mit ihrem Aufsatz *Relatives and Relativism* haben Diane Jeske und Richard Fumerton (1997) den Besondere-Pflichten-Einwand neu formuliert. Das Ziel der beiden ist es, zu zeigen, dass auch Konsequentialisten anerkennen müssen, dass es Situationen gibt, in denen es *erlaubt* ist, zu Gunsten seiner Vertrauten zu handeln, obwohl eine andere Handlung das allgemeine Wohlergehen mehr fördern würde (vgl. Jeske/Fumerton 1997: 154). Hierauf aufbauend folgern sie, dass Konsequentialisten entweder den Konsequentialismus zu Gunsten einer deontologischen Ethik fallen lassen müssen oder aber ein radikal relativistisches Konzept von Werten zu vertreten haben, will heißen, dass die richtige Handlung für den Akteur diejenige ist, die das maximiert, was für den Akteur einen Wert hat (vgl. Jeske/Fumerton 1997: 145).

Um ihre These zu untermauern, arbeiten sich Jeske und Fumerton an folgendem Szenario ab:

> Suppose, for example, you took your child canoeing. After taking the wrong fork in the river, your canoe overturns in the rapids. As it turns out, another canoe with two children has been caught in the same rapids and has suffered the same fate. You judge (correctly) that you can either save your child or save the two strangers but you cannot do both. (The two other children are relatively close to you but you will be unable to

[1] Der Einfachheit halber werde ich in diesem Artikel ein utilitaristisches Aggregationsprinzip zu Grunde legen und im Folgenden „Wohlergehen“ als eine mögliche Konkretisierung für ein intrinsisch wertvolles Gut verwenden.

save your child who has drifted further away if you first save those other children.) What should you do? (Jeske/Fumerton 1997: 146)

Das Hauptziel dieser Arbeit besteht in der Zurückweisung des Besondere-Pflichten-Einwands. Im Gegensatz zu der Schlussfolgerung von Jeske und Fumerton werde ich dafür argumentieren, dass auch ein Konsequentialist – der weder ein Konzept radikal relativer Werte vertritt, noch den Konsequentialismus zu Gunsten einer deontologischen Ethik fallen lässt – widerspruchsfrei anerkennen kann, dass es in einigen Situation moralisch erlaubt ist, das eigene Kind zu retten, selbst wenn eine Alternativhandlung das Aggregationsergebnis verbessert hätte. Um dieses Ziel zu erreichen, werde ich im dritten bis fünften Abschnitt die folgenden drei Hauptthesen verteidigen:

(1) In einer moralisch perfekten Welt generieren konsequentialistische Theorien eine Gruppe von Pflichten – die quasi-besonderen Pflichten –, die ihrem Inhalt nach mit einigen besonderen Pflichten identisch sind.

(2) Der Übergang von der moralisch perfekten Welt zur realen Welt macht es notwendig, dass konsequentialistische (Ein-Ebenen-)Theorien zu Zwei-Ebenen-Theorien weiterentwickelt werden.

(3) Konsequentialistische Zwei-Ebenen-Theorien generieren auch in der realen Welt einige quasi-besondere Pflichten.

Im Anschluss an die dritte These werde ich im sechsten Teil den Besondere-Pflichten-Einwand zurückweisen und vier Einwände gegen meine Schlussfolgerung diskutieren.

Selbst wenn Kritiker diese drei Thesen und die darauf basierende Schlussfolgerung akzeptieren, könnten sie noch immer versuchen, den Besondere-Pflichten-Einwand in einer abgeschwächten Form aufrechtzuerhalten. In diesem Fall würden sie dafür argumentieren, dass die diskutierten quasi-besonderen Pflichten zu eng gefasst sind und es darüber hinaus weitere wünschenswerte besondere Pflichten gibt, die innerhalb des dargestellten Konsequentialismus nicht generiert werden können. Das zweite Ziel dieser Arbeit besteht daher darin, auch diese abgeschwächte Form des Besondere-Pflichten-Einwands zu entkräften. Hierzu werde ich im siebten Abschnitt eine vierte Hauptthese diskutieren:

(4) Die generierten quasi-besonderen Pflichten – sowohl der moralisch perfekten Welt als auch der realen Welt – decken den Umfang an besonderen Pflichten ab, die innerhalb jeder moralischen Theorie, die das Prinzip der Unparteilichkeit[2] anerkennt, widerspruchsfrei gefordert werden können.

Den Abschluss dieser Arbeit bildet eine Schlussbetrachtung, in der ich die erarbeiteten Ergebnisse in Bezug zu einer möglichen Erwiderung von Jeske und Fumerton setze. Doch zunächst werde ich im zweiten Teil eine Besonderheit der besonderen Pflichten thematisieren und der Frage nachgehen, welche Kritik demzufolge tatsächlich mit dem Besondere-Pflichten-Einwand am Konsequentialismus geübt wird.

## 2. Die Struktur des Besondere-Pflichten-Einwands

Der Besondere-Pflichten-Einwand bringt eine seltsame Eigentümlichkeit mit sich. Für gewöhnlich empfinden Menschen moralische Pflichten – mit Ausnahme derjenigen Pflichten, die für ein friedliches Zusammenleben notwendig sind – als etwas, dass eher beschwerlich und lästig als wünschenswert ist. Es mag zwar sein, dass jemand im Nachhinein froh und

[2] Ohne im Folgenden näher darauf einzugehen, setze ich voraus, dass das Prinzip der Unparteilichkeit zu jeder plausiblen moralischen Theorie dazugehört und dass eine gerechtfertigte Parteilichkeit sich nur aus dem Prinzip der Unparteilichkeit ableiten lässt (vgl. Singer 2004: 14).

stolz darüber ist, dass er seine moralische Pflicht erfüllt hat, aber niemand würde ernsthaft behaupten, dass er heute seinen Vergnügungen nachgeht, *indem* er seine moralischen Pflichten erfüllt.

Hinzu kommt, dass der Konsequentialismus nicht in sich widersprüchlich wäre, wenn er keine besonderen Pflichten generieren könnte. Konsequentialistische Theorien treten für gewöhnlich mit einem sehr starken Unparteilichkeitsanspruch auf, so lautet seit Bentham ein häufig vorgebrachtes Diktum: „Jeder zählt für einen, keiner für mehr als einen" (zit. nach Mill 2006: 185). Die Anerkennung besonderer Pflichten erscheint demnach als eine halbherzige Ad-Hoc-Antwort, die gegen eine Grundidee des Konsequentialismus verstößt. Aber worauf zielt der Besondere-Pflichten-Einwand dann ab?

Letztlich kann er nur zeigen, dass der Konsequentialist zu bestimmten Handlungen moralisch verpflichtet ist bzw. ihm bestimmte Handlungen moralisch verboten sind und dass diese Handlungen nicht mit unseren gewöhnlichen Intuitionen vereinbar sind. Die meisten Menschen haben das klare Gefühl, dass es falsch ist, zu Gunsten eines Fremden zu handeln, wenn man stattdessen zu Gunsten des eigenen Kindes handeln kann, nur weil man damit das gesamte Wohlergehen (etwas) mehr fördert. Dass es moralisch verboten sein soll, zu Gunsten der Unsrigen zu handeln, ist für viele Menschen ein überaus quälender Gedanke. Eine Moral die fordert, einem geliebten Menschen nicht zu helfen, obwohl man es könnte, erscheint vielen Menschen inakzeptabel. Demnach wird mit dem Besondere-Pflichten-Einwand nicht versucht, die logische Falschheit des Konsequentialismus aufzuzeigen, sondern es wird appelliert, auf unser tief verwurzeltes Gefühl, dass wir für die Unsrigen sorgen müssen, zu hören. Dabei kommt es nicht auf die Art[3] der Pflicht an, die es dem moralischen Akteur erlaubt – oder besser: ihn verpflichtet – für die Seinigen zu sorgen, sondern auf ihren Inhalt[4]. Würde eine moralische Theorie den Inhalt besonderer Pflichten generieren, auch ohne dabei auf die spezifische Art der besonderen Pflichten zurückzugreifen, wäre der Besondere-Pflichten-Einwand hinfällig.

## 3. Zur ersten These

Meine erste Hauptthese lautet, dass konsequentialistische Theorien in einer moralisch perfekten Welt eine Gruppe von Pflichten generieren – die quasi-besonderen Pflichten –, die ihrem Inhalt nach mit einigen besonderen Pflichten identisch sind.

Zunächst ist zu klären, was im Folgenden unter einer moralisch perfekten Welt zu verstehen ist. Unter einer moralisch perfekten Welt verstehe ich eine Welt wie die Unsrige, mit dem Unterschied, dass alle Menschen durchgehend dazu fähig sind, die moralisch richtige Handlung zu erkennen, und außerdem durchgehend eine derart starke Motivation zum moralischen Handeln haben, dass diese alle anderen Handlungsmotivationen übertrumpft.[5]

In einer nach konsequentialistischen Gesichtspunkten moralisch perfekten Welt würde das Wohlergehen aller Menschen unparteiisch gefördert werden. Um beispielsweise das Wohlergehen von Kindern zu gewährleisten, müssen bestimmte Personen die Aufgabe übernehmen, für die Erfüllung des Wohlergehens der Kinder Sorge zu tragen. Da sich Eltern im Familienverband effektiv um die Bedürfnisse ihrer Kinder kümmern können und dies von den Eltern zumeist auch als ein wichtiger Teil eines erfüllenden Lebens wahrgenommen wird,

[3] Hiermit meine ich den Unterschied zwischen besonderen und allgemeinen Pflichten, so wie ich ihn in der Einleitung definiert habe.

[4] Hierunter zähle ich zum Beispiel, sich vorrangig um die eigenen Kinder kümmern zu müssen.

[5] Geht man von dem Gedanken aus, dass sich die reale Welt in eine moralisch perfekte Welt wandelt, dann muss als ein weiteres Kriterium die „stabile Integration" hinzugedacht werden. Hierunter verstehe ich, dass beispielsweise das Problem der (absoluten) Armut, soweit es sich innerhalb einer moralisch perfekten Welt lösen lässt, bereits gelöst ist.

würde die Sorgepflicht bei Kindern, deren Eltern noch leben und in der Lage sind, sie zu versorgen, auch in einer moralisch perfekten Welt durch die leiblichen Eltern übernommen werden. In einer derartigen Welt wäre das Ideal der Unparteilichkeit verwirklicht, aber bestimmte Menschen hätten dennoch die Pflicht, für das Wohlergehen derjenigen Kinder Sorge zu tragen, für die sie die Verantwortung übernommen haben (vgl. z.B. Koller 1998: 452-457; Rachels 2008: 266ff.). Innerhalb einer konsequentialistischen Theorie sind diese Pflichten ihrer Art nach zwar allgemeine Pflichten (vgl. Jeske 2008: 220), aber ihrem Inhalt nach entsprechen sie besonderen Pflichten, weshalb ich sie quasi-besondere Pflichten nenne.[6]

Es ist wichtig zu sehen, dass Kinder in der moralisch perfekten Welt diese Aufmerksamkeit nicht deshalb erfahren, weil sie zur Gruppe der „Kinder" gehören, so wie andere Menschen zu anderen Gruppen – Eltern, Frauen, Alte usw. – gehören. Sondern weil sie, jedes einzelne von ihnen, in einem besonderen Maße hilfsbedürftig sind. Sie können ihre elementarsten Grundbedürfnisse – Nahrung, Obdach, medizinische Versorgung usw. – nicht alleine befriedigen. Um das Wohlergehen insgesamt zu maximieren, muss die nötige Zuwendung aber noch über die Versorgung mit materiellen Gütern hinausgehen. Insbesondere für eine gesunde psychische Entwicklung ist soziale und emotionale Nähe von großer Bedeutung.[7] Soziale und emotionale Nähe kann dezentral, also beispielsweise von Eltern, wesentlich besser geleistet werden als in zentralen Großeinrichtungen. Hinzu kommt, dass Eltern, also diejenigen Menschen, die sich dafür entschieden haben, ein Kind zu bekommen, für gewöhnlich ein Interesse haben, sich um genau dieses Kind zu kümmern und im besonderen Maße darüber glücklich sind, zu sehen, wie es ihrem Kind gut geht.

## 4. Zur zweiten These

Die zweite Hauptthese besagt, dass der Übergang von der moralisch perfekten Welt zur realen Welt es notwendig macht, dass konsequentialistische (Ein-Ebenen-)Theorien zu Zwei-Ebenen-Theorien weiterentwickelt werden. Die Begründung dieser These basiert im Wesentlichen auf Hares Zwei-Ebenen-Theorie, die er in seinem Buch *Moralisches Denken* (1992) ausgeführt hat[8]. Kurzum: Der Mensch hat zahlreiche Schwächen, durch die ihm der Eintritt ins Paradies der moralisch perfekten Welt versperrt ist. Zu den zentralen Schwächen gehört, dass der Mensch nicht immer in der Lage ist, die moralisch beste Handlung auszuführen – oder in Hares Terminologie: Menschen können über weite Strecken nicht perfekt kritisch denken. Sie erfassen nicht immer alle relevanten Merkmale einer Situation oder gewichten die erfassten Merkmale falsch. Infolgedessen kommt es selbst bei Menschen mit einer Motivation zum moralischen Handeln zu moralisch ineffizienten Handlungen (vgl. Hare 1992: 91-94). Verstärkend kommt hinzu, dass selbst das Erkennen der moralisch besten Handlung nicht zwangsläufig zu einem Motivationspotenzial führt, das die Ausführung der moralisch besten Handlung notwendig nach sich zieht. Die Motivation zum moralischen Handeln kann durch andere Interessen überlagert werden.

Um den moralischen Output zu verbessern, ist es daher notwendig, dass das kritische Denken durch ein intuitives Denken unterstützt wird. Die grundlegende Aufgabe des intuitiven

---

[6] Bis hierhin wurde noch nichts darüber ausgesagt, wie in einer moralisch perfekten Welt ein Konsequentialist in der Kanusituation zu handeln hat (siehe hierzu Anmerkung 12). Es wurde lediglich gezeigt, dass innerhalb einer moralisch perfekten Welt *einige* quasi-besondere Pflichten gerechtfertigt werden können.

[7] Hier ist an psychologische Phänomene wie Hospitalismus zu denken (z.B. Spitz 1974).

[8] Hare steht mit der Ausarbeitung seiner Zwei-Ebenen-Theorie in einer langen utilitaristischen Tradition (vgl. Fehige/Frank 2010). Für eine generelle Kritik an der Zwei-Ebenen-Theorie von Hare siehe exemplarisch Scanlon (1990). Für eine Kritik am Anspruch und der Leistungsfähigkeit der kritischen Ebene siehe beispielsweise Nida-Rümelin (1995) und für die intuitive Ebene Birnbacher (1995).

Denkens ist es, moralische Handlungen in Situationen anzuleiten, in denen wir den dargestellten menschlichen Schwächen zu unterliegen drohen. Um diese Aufgabe zu erfüllen, müssen bei der moralischen Erziehung Prinzipien und Pflichten bzw. Intuitionen und Handlungsdispositionen verankert werden, welche das Finden der moralischen Handlung erleichtert, die zu einer Annäherung an das optimale Aggregationsergebnis führt (vgl. Hare 1992: 94). Dabei ist zu beachten, dass das Verankern derartiger Intuitionen und Handlungsdispositionen bereits im Kindesalter und nicht erst im Erwachsenenalter beginnt. Da im Kindesalter aber noch nicht abzusehen ist, welche Fähigkeit zum kritischen Denken im Erwachsenenalter bei den jeweiligen Individuen besteht, können die verankerten Prinzipien nicht auf das tatsächliche Vermögen abgestimmt werden. Daher müssen sich diese an einer durchschnittlichen Fähigkeit orientieren, um eine Annäherung an das optimale Aggregationsergebnis zu gewährleisten.

## 5. Zur dritten These

In Bezug auf den Besondere-Pflichten-Einwand ist die dritte Hauptthese von besonderer Bedeutung, sie lautet: Konsequentialistische Zwei-Ebenen-Theorien generieren auch in der realen Welt einige quasi-besondere Pflichten.

Die entscheidende Frage ist: Welche Prinzipien und Pflichten sind auf der intuitiven Ebene zu verankern bzw. in Bezug auf den Besondere-Pflichten-Einwand: Welche *quasi-besonderen* Pflichten sind zu verankern? Hare hat dazu selbst ein eindringliches Beispiel gegeben:

> Hätten Mütter einen Hang dazu, sich um alle Kinder dieser Welt gleich viel zu kümmern, so ist es unwahrscheinlich, daß Kinder auch nur so gut behandelt würden wie derzeit. Die Verantwortung wäre bis zur Nichtexistenz verdünnt. (Hare 1992: 199f.)

Dementsprechend ist die quasi-besondere Pflicht, dass sich Mütter vorrangig um ihre Kinder kümmern müssen, auf der intuitiven Ebene zu verankern.[9]

Auch wenn ich es in dieser Arbeit nicht umfassend ausführen kann, ist die Annahme, dass es auf der Basis besonderer Beziehungen weitere quasi-besondere Pflichten gibt, nicht unberechtigt. Man denke diesbezüglich nur an das durch stabile Freundschaften gesteigerte Wohlergehen. Stabile Freundschaften sind aber nur möglich, wenn man Freunden zumindest eine gewisse Priorität zukommen lässt.

Des Weiteren gehe ich davon aus, dass innerhalb einer konsequentialistischen Zwei-Ebenen-Theorie auch quasi-besondere Pflichten auf der Basis vorausgehender Handlungen gerechtfertigt werden können. Es ist ganz richtig, wie Nida-Rümelin in seiner Kritik an Hare ausführt, dass das Versprechen für den (gewöhnlichen) Menschen eine fundamentale Regel der Handlungskoordination ist und dass daher anzunehmen ist,

> daß das in einer Gesellschaft insgesamt verwirklichte Maß der Präferenzerfüllung abnimmt, wenn sich die Institution des Versprechens nicht aufrechterhalten ließe. (Nida-Rümelin 1995: 46)

Es ist nicht ganz klar, ob das Versprechen in einer moralisch perfekten Welt von gleicher Bedeutung wie in unserer realen Welt ist, aber gerade in unserer Welt, in der die menschlichen Schwächen einen festen Platz haben, wird eine fest verankerte Institution des

[9] Zwei Punkte seien angemerkt: Erstens, ich gehe davon aus, dass diese Pflicht nicht nur für Mütter sondern auch für Väter gültig ist. Zweitens, es ist wichtig zu sehen, dass die quasi-besondere Pflicht, sich vorrangig um die eigenen Kinder zu kümmern, noch spezifiziert werden muss. Dass eine absolute Priorisierung nicht zu einer Annäherung an das optimale Aggregationsergebnis führt, sieht man beispielsweise an Fällen von Nepotismus. Zur tiefergehenden Begründung der genannten quasi-besonderen Pflicht siehe ergänzend die Diskussion im Rahmen des dritten Abschnitts.

Versprechens das Wohlergehen stark fördern. Daher ist die Pflicht, seine Versprechen zu halten, ebenso auf der intuitiven Ebene zu verankern.

## 6. Zurückweisung des Besondere-Pflichten-Einwands

Nach der Verteidigung der ersten drei Thesen kann der Besondere-Pflichten-Einwand diskutiert werden. Hierzu werde ich zunächst das Kanubeispiel mit weiteren Situationsmerkmalen anreichern, dann eine weitere These formulieren und vier Einwände gegen diese These vorbringen. Anhand der Diskussion der Einwände werde ich Argumente für meine neue These nachreichen.

Zunächst sei daran erinnert, dass auf der intuitiven Ebene eine Form der quasi-besonderen Pflicht, sich vorrangig um die eigenen Kinder zu kümmern, verankert wurde. Um den Kritikern möglichst entgegen zu kommen, sollen weitere Bedingungen erfüllt sein: Erstens, kritisches Denken kommt zu dem Ergebnis, dass es moralisch das Richtige ist, die beiden anderen Kinder zu retten. Zweitens, die erwachsene Person ist sich sowohl dem Ergebnis des kritischen Denkens bewusst als auch der Tatsache, dass sie zum gegenwärtigen Zeitpunkt in der Lage ist, perfekt kritisch zu denken.[10]

Wie ist nun die Rettung des eigenen bzw. die Rettung der beiden fremden Kinder innerhalb der vorgestellten Zwei-Ebenen-Theorie zu bewerten? Meine These lautet: Sowohl die Rettung des eigenen Kindes als auch die Rettung der beiden fremden Kinder ist moralisch *erlaubt*.

### *6.1 Erster Einwand*

Der erste Einwand könnte wie folgt lauten: Aus konsequentialistischer Perspektive muss es moralisch verboten sein, das eigene Kind in der beschriebenen Kanusituation zu retten: Der Konsequentialismus ist eine Theorie, die die moralische Richtigkeit und Falschheit von Handlungen ausschließlich aufgrund der (wahrscheinlichen) Handlungsfolgen beurteilt. Das eigene Kind zu retten, obwohl man erkannt hat, dass diese Handlung (wahrscheinlich) nicht das Wohlergehen maximiert, widerspricht dem Aggregationsprinzip. Demnach kann es nicht moralisch erlaubt sein, das eigene Kind zu retten.

Dieser Einwand ist jedoch nicht zutreffend. Um dies zu verstehen, ist es wichtig, sich noch einmal eine der bereits angesprochenen menschlichen Schwächen in Erinnerung zu rufen: Das Erkennen der moralisch besten Handlung schließt nicht zwangsläufig die Motivation ein, diese Handlung auch auszuführen. Um derartige Schwächen möglichst gut aufzufangen, wurde die intuitive Ebene eingeführt, die die Handlungsmotivation in eine Richtung lenken soll, mit der eine Annäherung an das optimale Aggregationsergebnis erreicht wird. Die verankerten Pflichten, Intuitionen und Dispositionen der intuitiven Ebene können aber nur auf Situationen vorbereiten, die einigermaßen häufig vorkommen, also eine gewisse Regelmäßigkeit an den Tag legen. Es wird daher Situationen geben, in denen die verankerten Pflichten eine Handlung fordern, die in diesen ganz speziellen Situationen zu einem nicht-optimalen Aggregationsergebnis führen. Wenn aber

(1) das Erkennen der moralisch besten Handlung nicht zwangsläufig zu einem Motivationspotenzial führt, das die Ausführung der Handlung notwendig nach sich zieht

(2) und die Intuitionen und Handlungsdispositionen der intuitiven Ebene eine starke Quelle der Handlungsmotivationen sind

[10] Es ist allerdings fraglich, ob diese Bedingung in der realen Welt bei einem vergleichbaren Szenario überhaupt erfüllbar ist.

(3) und gemäß dem kritischen Denken die Intuition und Handlungsdisposition „Eltern sollen sich vorrangig um ihre Kinder kümmern“ auf der intuitiven Ebene verankert wurde,

(4) dann muss eine Handlung, die diesem Prinzip folgt, auch dann erlaubt sein, wenn kritisches Denken in der konkreten Situation eine andere Handlung als moralisch beste Handlung identifiziert. Denn die Rettung des eigenen Kindes ist letztlich auf die größere Motivation für diese Handlung zurückzuführen. Diese Motivation ist aber selbst durch (ein vorausgehendes) kritisches Denken gerechtfertigt.[11]

Meines Erachtens kommt die Schwierigkeit im Kanubeispiel daher, dass die Situationsbeschreibung nicht angemessen sein kann. Es stellt sich die folgende Frage: Wie ist es zu verstehen, dass das kritische Denken in der Kanusituation zum Ergebnis kommt, dass es moralisch das Richtige ist, die beiden fremden Kinder zu retten, obwohl kritisches Denken gleichzeitig die Intuition, dass sich Eltern vorrangig um ihre Kinder kümmern sollen, rechtfertigt? Hier bieten sich eine schwache und eine starke Interpretationen für das Ergebnis des kritischen Denkens an:

Die starke Interpretation: Unter Berücksichtigung aller relevanten Fakten ist es moralisch das Richtige, die beiden fremden Kinder zu retten.

Die schwache Interpretation: Unter idealisierten Voraussetzungen ist es moralisch das Richtige, die beiden fremden Kinder zu retten.

Meine nächste These ist nun, dass, bezogen auf die Kanusituation, das kritische Denken nur im Sinne der schwachen Interpretation zu verstehen ist. Die Berücksichtigung aller relevanten Fakten schließt nämlich mit ein, dass auch die menschliche Natur mit den menschlichen Schwächen berücksichtigt wird. Welche Handlung aber unter Berücksichtigung der menschlichen Natur mit den menschlichen Schwächen die Richtige ist, ist bereits mit der quasi-besonderen Pflicht auf der intuitiven Ebene geklärt. Mit der Verankerung der entsprechenden Intuition und Handlungsdisposition wurde bereits in Kauf genommen, dass es *einige* Situationen gibt, in denen das allgemeine Wohlergehen durch die Ausführung dieser Pflicht nicht direkt maximiert wird, sondern nur über die langfristigen Folgen. Die Kanusituation ist nichts weiter als eine Beschreibung einer derartigen Situation. Unter der Voraussetzung, dass selbst in Anbetracht der Kanusituation kritisches Denken zu dem Ergebnis kommt, dass die quasi-besondere Pflicht – Eltern sollen sich vorrangig um ihre Kinder kümmern – weiterhin auf der intuitiven Ebene verankert sein soll, kann kritisches Denken nicht gleichzeitig zum Ergebnis kommen, dass, unter Berücksichtigung aller relevanten Fakten, Eltern, wenn sie den menschlichen Schwächen unterliegen, sich *nicht* vorrangig um ihre Kinder kümmern sollen. Im kritischen Denken gibt es keine unauflöslichen Konflikte (vgl. Hare 1992: 70f.).

Die Verankerung eines derartigen Vorrangprinzips ist eben ein tiefgehender Eingriff in die motivationalen Mechanismen des Menschen. Es muss eine Handlungsdisposition ausgebildet werden, die verlässlich gegenteilige Handlungsimpulse übertrumpft. Eine derart starke Handlungsdisposition wird aber auch in Situationen zum Tragen kommen, in der es aus konsequentialistischer Sicht besser wäre, wenn sie in dieser ganz speziellen Situation nicht vorhanden wäre. An dieser Stelle wird die schwächere Interpretation plausibel: Das kritische Denken kann noch immer zum Ergebnis kommen, dass es moralisch das Richtige ist, die beiden fremden Kinder zu retten, wenn man in dieser Situation keiner relevanten menschlichen Schwäche unterliegt. Das heißt, wenn man es beispielsweise schafft, aus dem Erkennen der moralisch besten Handlung genügend Motivationspotenzial aufzubauen, um

[11] Demnach ist es in der realen Welt unangemessen, der helfenden Person einen Vorwurf zu machen, wenn sie das eigene Kind gerettet hat, weil sie gemäß einer Disposition handelte, die andere – mit guten Gründen – in ihr erzeugt haben.

diese Handlung trotz gegenteiliger Neigungen durchzuführen.[12] Interpretiert man das Ergebnis des kritischen Denkens in der schwächeren Form, dann steht das Ergebnis des kritischen Denkens zudem nicht weiter in Konflikt mit dem Ergebnis des intuitiven Denkens. Es handelt sich bei der Kanusituation demnach um einen Scheinkonflikt.

Wenn das Ergebnis des kritischen Denkens in der konkreten Situation nur als eine „unter-idealisierten-Voraussetzungen"-Aussage zu verstehen ist und das Retten des eigenen Kindes angeblich unter Berücksichtigung der menschlichen Natur und der menschlichen Schwächen die richtige Handlung ist, wie kann dann das Retten der beiden fremden Kinder ebenfalls erlaubt sein? Das Retten der beiden fremden Kinder ist nichts, was dem Helfer in einem strengen Sinn unmöglich ist. Er hat die physischen Fähigkeiten, um zu diesen Kindern zu schwimmen und sie ans Land zu ziehen. Was ihn ggf. davon abhält, ist die Handlungsdisposition, sich vorrangig um sein eigenes Kind zu kümmern. Es ist jedoch möglich, die Grenzen, die uns durch die menschlichen Schwächen gegeben sind, ein Stück weit zu verschieben und somit unsere moralische Leistungsfähigkeit zu erhöhen: Wer den Willen zum moralischen Handeln hat, wird unweigerlich in Situationen geraten, in denen die moralischen Überzeugungen eine andere Handlung fordern als das Selbstinteresse. Insbesondere am Anfang wird es schwer sein, genügend Motivation für die moralisch richtige Handlung aufzubringen. Je öfter es dem Akteur jedoch gelingt, die moralisch richtige Handlung auszuführen, desto mehr wird er sich daran gewöhnen und sein Motivationsgefüge umbauen, so dass es ihm insgesamt leichter fällt, auch in Konfliktsituationen moralisch zu handeln. Auf diese Weise verschiebt er Stück für Stück die Grenzen, die ihn durch seine menschlichen Schwächen, in diesem Fall das Selbstinteresse, gegeben sind. Es sei zudem daran erinnert, dass sich die Prinzipien und Pflichten der intuitiven Ebene an der *gewöhnlichen* moralischen Leistungsfähigkeit orientieren, weil sie zu einem Zeitpunkt verankert werden, zu dem nicht klar ist, inwieweit die entsprechende Person tatsächlich den gewöhnlichen Schwächen unterliegen wird. Daher ist es nicht widersinnig, dass die erwachsene Person im Kanubeispiel zu denjenigen moralischen Akteuren gehört, die mehr leisten können, als es die Prinzipien und Pflichten der intuitiven Ebene verlangen.

### *6.2 Zweiter Einwand*

Gegen diesen Verteidigungsversuch lässt sich der erste Einwand in zwei Richtungen weiter spezifizieren: Erstens, wenn die erwachsene Person die Fähigkeit hat, sich über die menschlichen Schwächen hinwegzusetzen und die beiden fremden Kinder zu retten, dann ist es gemäß der konsequentialistischen Logik für den Akteur auch moralisch verpflichtend, die beiden fremden Kinder zu retten und damit verboten, das eigene Kind zu retten. Zweitens, wenn die erwachsene Person demgegenüber nicht die Fähigkeit hat, sich über die menschlichen Schwächen hinwegzusetzen, dann ist es verpflichtend, gemäß den Prinzipien der intuitiven Ebene zu handeln und demnach verboten, die beiden fremden Kinder zu retten.

Der erste Teil des Einwands läuft ins Leere, weil sich die Frage, wie es zu bewerten ist, wenn der Akteur trotzdem sein eigenes Kind rettet, nicht mehr sinnvoll stellen lässt: Wenn die beste Handlung in der Rettung der beiden fremden Kinder besteht und sich die Person dessen bewusst ist und sie außerdem dazu in der Lage ist, die beiden fremden Kinder zu retten, dann folgt daraus notwendig, dass sie auch die beiden fremden Kinder rettet, wenn sie ausreichend motiviert ist, die moralisch beste Handlung durchzuführen. Der Fall, dass die erwachsene Person unter diesen Bedingungen das eigene Kind rettet, kann nicht auftreten.

[12] Damit wird die Handlungsbeurteilung in der moralisch perfekten Welt offensichtlich. Dadurch, dass die intuitive Ebene in der moralisch perfekten Welt entfällt, fällt die starke Interpretation des kritischen Denkens mit der schwachen Interpretation zusammen. Jedoch gilt dieses Ergebnis nur für die moralisch perfekte Welt, daher braucht uns die Feststellung, dass wir in der moralisch perfekten Welt die beiden anderen Kinder retten müssen, nicht weiter zu beschäftigen.

Damit ist die Handlungscharakterisierung als verpflichtend – im Gegensatz zu erlaubt oder verboten – widersinnig, weil es eben die einzig mögliche Handlung in dieser Konstellation ist.

Das Retten des eigenen Kindes ist nur möglich, wenn die erwachsene Person nicht ausreichend motiviert ist, die moralisch beste Handlung auszuführen. Ist sie aber nicht ausreichend motiviert, die moralisch beste Handlung auszuführen, obwohl sie es kann und sie die beste Handlung erkannt hat, unterliegt sie eben doch der bereits angesprochenen Schwäche: Das Erkennen der moralisch besten Handlung führt nicht zwangsläufig zu einem Motivationspotenzial, das die Ausführung dieser Handlung notwendig nach sich zieht. In diesem Fall ist sie aber nicht fähig, sich über alle relevanten menschlichen Schwächen hinwegzusetzen und daher ist es ihr auch moralisch erlaubt, dem intuitiven Denken zu folgen und das eigene Kind zu retten.

Ist der zweite Teil des Einwands zutreffend? Angenommen im Kanuszenario rettet die erwachsene Person die beiden fremden Kinder, aber nicht aus moralischen Motiven, sondern weil sie sich als Held inszenieren will. Zweifelsohne unterliegt auch dieser Akteur den menschlichen Schwächen. Aber hat die Person eine verbotene Handlung ausgeführt? Auf Grund ihres egoistischen Motivs verdient sie vielleicht kein Lob für diese Tat, aber letztlich hat sie diejenige Handlung vollzogen, die das gesamte Wohlergehen maximiert hat. Eine derartige Handlung muss auch innerhalb der skizzierten konsequentialistischen Theorie erlaubt sein, selbst wenn sie aus den falschen Motiven geschieht. Ist diese Handlung aber erlaubt, dann kann das Retten des eigenen Kindes nicht mehr verpflichtend sein, wenn man den menschlichen Schwächen unterliegt.

### *6.3 Dritter Einwand*

Handelt es sich noch um eine konsequentialistische Theorie? Man könnte argumentieren, dass der Konsequentialismus zu Gunsten einer deontologischen Ethik, also einer Theorie, die die moralische Richtigkeit und Falschheit von Handlungen *nicht ausschließlich* aufgrund der (wahrscheinlichen) Handlungsfolgen beurteilt, fallen gelassen wurde. Mit der Einführung von erlaubten, aber nicht gebotenen Handlung sind Optionen ins Spiel gekommen, so könnte ergänzt werden, die das Prinzip, „den besten Zustand herbeizuführen“ (Schroth 2011: 146), aushebeln. Aber auch diese Schlussfolgerung greift zu kurz. Denn die vermeintlich deontologischen Pflichten der intuitiven Ebene müssen zunächst einen konsequentialistischen Filter passieren: sie müssen von der kritischen Ebene gerechtfertigt werden. Damit bleibt die Theorie eben doch konsequentialistisch. Die Entstehung von Optionen ist eine notwendige Folge der Verankerung des konsequentialistischen Prinzips im gewöhnlichen Menschen mit all seinen Schwächen.

### *6.4 Vierter Einwand*

Während die ersten drei Einwände zu zeigen versuchten, dass die Handlungscharakterisierung grundlegend falsch ist oder sich außerhalb des Rahmens konsequentialistischer Theorien bewegt, versucht der vierte Einwand zu zeigen, dass diese Charakterisierung zu Handlungserlaubnissen führen würde, die moralisch unplausibel sind: Wenn, wie oben gezeigt, das Halten eines Versprechens als besondere Pflicht zu verankern ist und es analog zum Kanubeispiel erlaubt ist, dieses Versprechen auch dann zu halten, wenn eine andere Handlung das Aggregationsergebnis verbessern würde, dann muss es auch erlaubt sein, bei einem Verkehrsunfall, bei dem man als einziger Retter zur Verfügung steht, nicht zu helfen, wenn dies die einzige Möglichkeit ist, um sein Versprechen zu halten. Dass es aber erlaubt sein soll, einem schwer verletzten Menschen nicht zu helfen, nur um ein Versprechen zu halten, widerspricht unseren moralischen Intuitionen und insbesondere der konsequentialistischen Grundidee.

Wie ist dieser Einwand zu bewerten? Zunächst ist anzumerken, dass bisher sehr wenig darüber gesagt wurde, wie der Inhalt der Versprechenspflicht im Detail ausformuliert ist. Warum sollte die verankerte Pflicht, seine Versprechen halten zu müssen, derart einfach strukturiert sein, dass sie keine Randbedingungen zulässt? Das wird dem Konzept des Versprechens nicht gerecht; Versprechen haben eine unterschiedliche Bindungskraft. Es macht einen relevanten Unterschied, ob ein erfahrener Dschungelführer verspricht, jemanden wieder aus dem Dschungel herauszuführen, in den er ihn gebracht hat, oder ob man einem Freund verspricht, ihn zu besuchen. Kriterien, um die Bindungskraft eines Versprechens zu ermitteln, sind beispielsweise die Kompensierbarkeit bei einem Nicht-Einhalten und der Grad der eigenen Schuldhaftigkeit für das Nicht-Einhalten. Diese unterschiedliche Gewichtung muss sich auch bei der Verankerung der Versprechenspflicht widerspiegeln. Da einfache Konflikte mit dem Versprechen relativ häufig vorkommen, stellt es uns vor keine allzu großen Probleme, hier ein gutes Gefühl zu entwickeln, ob es moralisch richtig ist, das Versprechen zu halten oder zu brechen.

Natürlich kann man damit nicht alle Schwierigkeiten aus dem Weg räumen. Vielleicht auch nicht die des Autounfalls. Letztlich drückt das Unfallbeispiel aber auch nicht mehr als die bereits bekannte Schwierigkeit aus, dass die Pflichten der intuitiven Ebene nur auf Situationen ausgelegt sein können, die relativ häufig auftreten. Eine Unfallsituation, in der es trotz einer spezifizierten Versprechenspflicht moralisch erlaubt ist, das Versprechen zu halten, wird jedoch äußerst selten vorkommen. Daher ist es auch kein großes Eingeständnis, dass es erlaubt sein *kann*, in einigen Situationen weiterzufahren.

Der Einwand ist aber aus einer anderen Perspektive sehr interessant. Denn mit diesem Einwand wird der Besondere-Pflichten-Einwand in die entgegengesetzte Richtung gedreht. Die Kritik zielt nicht mehr darauf, dass konsequentialistische Theorien keine besonderen Pflichten rechtfertigen können, sondern dass die gerechtfertigten quasi-besonderen Pflichten zu stark sind. Das zeigt sehr deutlich das Gesamtproblem des Besondere-Pflichten-Einwands: Es kann nicht einfach darum gehen, möglichst starke besondere Pflichten zu generieren, sondern diese Pflichten müssen innerhalb eines bestimmten Rahmens gerechtfertigt werden. Die hier diskutierte Theorie stellt einen plausiblen Rahmen bereit. Sie rechtfertigt quasi-besondere Pflichten, also eine gewisse Parteilichkeit, direkt aus der Unparteilichkeit und in Anbetracht der menschlichen Natur.

## 7. Zur vierten These

Wie einleitend erwähnt, kann der Besondere-Pflichten-Einwand noch immer in einer abgeschwächten Form vorgebracht werden. Hierzu wird behauptet, dass die gerechtfertigten quasi-besonderen Pflichten zu eng gefasst sind und es darüber hinaus relevante besondere Pflichten gibt, die innerhalb des hier besprochenen Konsequentialismus nicht generiert werden können.

Dieser schwächere Einwand trägt allerdings nichts aus. Wer das Prinzip der Unparteilichkeit akzeptiert, setzt sich zugleich, unabhängig davon, welche moralische Theorie er vertritt, einen sehr engen Rahmen für alle moralisch erlaubten Handlungen, insbesondere innerhalb einer moralisch perfekten Welt. Alle vermeintlich besonderen Pflichten, welche im Widerspruch zum Prinzip der Unparteilichkeit stehen, sind als moralisch irrelevante Intuitionen zu qualifizieren. Besondere Pflichten, die nicht im Widerspruch zum Prinzip der Unparteilichkeit stehen – quasi-besondere Pflichten –, können auch von der hier diskutierten konsequentialistischen Theorie generiert werden. Daher ist meine vierte Hauptthese, dass die generierten quasi-besonderen Pflichten – sowohl der moralisch perfekten Welt als auch der realen Welt – den Umfang an besonderen Pflichten abdecken, die innerhalb jeder

moralischen Theorie, die das Prinzip der Unparteilichkeit anerkennt, widerspruchsfrei gefordert werden können.

Eine Möglichkeit, diese These zu widerlegen, bestünde darin, zeigen zu können, dass es weitere relevante und widerspruchsfrei rechtfertigbare besondere Pflichten gibt, die nicht innerhalb des diskutierten Konsequentialismus generiert werden können. Es ist zwar an dieser Stelle nicht vollständig möglich, nachzuweisen, dass dies generell unmöglich ist, aber ich werde mit der folgenden Überlegung zeigen, dass dies unplausibel ist bzw. dass die generierten besonderen Pflichten nicht ausreichen, um den Besondere-Pflichten-Einwand als einen relevanten Einwand gegen den Konsequentialismus aufrechtzuerhalten.

Um den Besondere-Pflichten-Einwand zu retten, müssen die in Frage kommenden besonderen Pflichten fünf Bedingungen erfüllen. Die ersten beiden Bedingungen ergeben sich direkt aus der vierten These. Die dritte bis fünfte Bedingung ist notwendig, um den Besondere-Pflichten-Einwand plausibel zu halten.

(1) Es müssen besondere Pflichten sein, die kritisches Denken nicht auswählen würde.

(2) Die moralische Theorie, aus der die besonderen Pflichten abgeleitet werden, muss das Prinzip der Unparteilichkeit enthalten.

(3) Die besonderen Pflichten müssen selbst wünschenswert sein.

(4) Die besonderen Pflichten dürfen nicht in einem unauflöslichen Widerspruch zu anderen Pflichten der moralischen Theorie stehen, insbesondere nicht zum Prinzip der Unparteilichkeit.

(5) Die moralische Theorie, aus der die besonderen Pflichten abgeleitet werden, muss insgesamt plausibel vertretbar sein.

Die wesentliche Schwierigkeit dürfte darin bestehen, eine besondere Pflicht zu finden, die zugleich wünschenswert und mit dem Prinzip der Unparteilichkeit vereinbar ist und die dennoch nicht vom kritischen Denken als eine zu verankernde Pflicht ausgewählt werden würde.

## 8. Schlussbetrachtung

Wären Jeske und Fumerton von der Zurückweisung des Besondere-Pflichten-Einwands überzeugt? Die ernüchternde Antwort ist: vermutlich nein. Zur Verteidigung ihrer Thesen würden sie auf ihre Diskussion des Regelkonsequentialismus verweisen:

> Thus, it may be that the act of my saving my child has worse consequences than the act of my saving the other two children. However, my saving my own child is in accord with a moral rule (parents ought to save their own children before they save other children), and the consequences of everyone's or most everyone's following that rule has better consequences than their following some alternative rule. [...] We simply imagine a world in which the consequences of people's following the rule 'save your own children first' does not maximize value. (Jeske/Fumerton 1997: 149f.)

Unter diesen veränderten Bedingungen würden sie darauf insistieren, dass die besondere Beziehung der Eltern zu ihren Kindern noch immer besondere Pflichten generiert, nach denen es erlaubt bzw. verpflichtend ist, das eigene Kind zu retten (vgl. Jeske/Fumerton 1997: 151f.). Hält damit der Besondere-Pflichten-Einwand doch stand? Ich sehe nicht, warum das der Fall sein sollte. Zunächst ergeben sich erhebliche Zweifel an der grundsätzlichen Zulässigkeit, dass sich Intuitionen, die sich im Rahmen der realen Welt ausgebildet haben, auf fiktive Welten übertragen lassen. Fraglich ist zudem, ob die Intuition, sich vorrangig um

das eigene Kind kümmern zu müssen, in einer derartigen Situation überhaupt gestützt werden kann. Diese ist aus zwei Gründen zweifelhaft:

(1) Bei nicht-konsequentialistischen Theorien hängt die Richtigkeit einer Handlung zwar nicht *nur* von den Konsequenzen ab; für gewöhnlich hängt sie aber *auch* von den Konsequenzen ab. Im konkreten Fall sprechen diese aber für das Retten der beiden fremden Kinder und können somit die Intuition nicht stützen.

(2) Begründet wird die Intuition mit der besonderen Pflicht, die aus der besonderen Beziehung generiert wurde. Die Generierung der besonderen Pflicht allein reicht jedoch nicht aus, um die Bevorzugung des eigenen Kindes in dieser Situation zu erlauben. Sie steht in dem modifizierten Kanubeispiel im Konflikt mit einer allgemeinen Pflicht, wie zum Beispiel Menschen vor vermeidbaren Schaden zu bewahren. Gezeigt werden muss, warum die besondere Pflicht in dieser Situation die allgemeine Pflicht übertrumpft.

Entweder enthält die moralische Hintergrundtheorie, die den Pflichtenkonflikt auflösen soll, das Prinzip der Unparteilichkeit; dann ist umso fraglicher, wie die Auflösung des Konflikts in der imaginierten Welt widerspruchsfrei zu Gunsten der besonderen Pflicht geschehen soll. Oder die moralische Hintergrundtheorie enthält kein Prinzip der Unparteilichkeit, dann fehlt ihr aber ein wesentliches Prinzip moralischer Theorien und es wird fraglich, warum der Besondere-Pflichten-Einwand auf der Basis einer derartigen Theorie überhaupt relevant sein sollte. Erschwerend kommt hinzu, dass es sich beim modifizierten Szenario nur um ein theoretisches Beispiel für eine Welt handelt, die mit Annahmen arbeitet, die für die reale Welt nicht zutreffen. Demgegenüber konnte gezeigt werden, dass die diskutierte konsequentialistische Theorie in der realen Welt sehr wohl einige (quasi-)besondere Pflichten rechtfertigt.[13]

**Marcel Warmt**

Universität Kassel
marcel.warmt@googlemail.com

## Literatur

Birnbacher, D. 1995: „Handeln und Unterlassen im »Zwei-Ebenen-Model der Moral«, in C. Fehige, G. Meggle (Hrg.): *Zum moralischen Denken*, Band 2, Frankfurt am Main: Suhrkamp, 176-86.

Fehige, C., R. H. Frank 2010: „Feeling Our Way to the Common Good: Utilitarianism and the Moral Sentiments“, *The Monist* 93, 141-65.

Hare, R. M. 1992: *Moralisches Denken*. Frankfurt am Main: Suhrkamp.

Jeske, D. 2008: „Familien, Freunde und besondere Verpflichtungen“, in A. Honneth und B. Rössler (Hrg.): *Von Person zu Person. Zur Moralität persönlicher Beziehungen*, Frankfurt am Main: Suhrkamp, 215-53.

—, R. Fumerton 1997: „Relatives and Relativism“, in *Philosophical Studies* 87, 143–57.

[13] Für wertvolle Kommentare danke ich den TeilnehmerInnen an meinem Sektionsbeitrag beim 8. Internationalen Kongress der Gesellschaft für Analytische Philosophie, den TeilnehmerInnen des Saarbrücker Doktorandenkolloquiums zur Praktischen Philosophie im Oktober 2012, den TeilnehmerInnen am GAP-Doktorandenworkshop 2012 in Zürich sowie Philippe Brunozzi, Jens Schnitker und Walter Pfannkuche.

Koller, P. 1998: „Einwanderungspolitik im Kontext internationaler Gerechtigkeit“, in C. Chwaszcza, W. Kersting (Hrg.): *Politische Philosophie der internationalen Beziehungen*. Frankfurt am Main: Suhrkamp, 449- 66.

Mill, J.S. 2006: *Utilitarianism/ Der Utilitarismus*. Stuttgart: Reclam.

Nida-Rümelin, J. 1995: „Kann der Erzengel die Konsequentialismus-Kritik entkräften?“, in C. Fehige, G. Meggle (Hrg.): *Zum moralischen Denken*, Band 2, Frankfurt am Main: Suhrkamp, 42-52.

Rachels, J. 2008: „Eltern, Kinder und die Moral“, in A. Honneth und B. Rössler (Hrg.): *Von Person zu Person. Zur Moralität persönlicher Beziehungen*, Frankfurt am Main: Suhrkamp, 254-76.

Scanlon, T. M. 1990 : „Levels of Moral Thinking“, in D. Seanor und N. Fotion (Hrg.): *Hare and Critics – Essays on Moral Thinking*, Oxford: Oxford University Press, 129-46.

Schroth, J. 2011: „Könnte Nelson ein Konsequentialist gewesen sein?“, in A. Berger, G. Raupach-Strey und J. Schroth (Hrg.): *Leonard Nelson – ein früher Denker der Analytischen Philosophie? Ein Symposion zum 80. Todestag des Göttinger Philosophen*, Berlin: LIT Verlag, 129-48.

Singer P. 2004: „Outsiders: our obligations to those beyond our borders“, in D. K. Chatterjee (Hrg.): *The Ethics of Assistance. Morality and the Distant Needy*, Cambridge: Cambridge Universtiy Press, 11-32.

Spitz, R. 1974: *Vom Säugling zum Kleinkind. Naturgeschichte der Mutter-Kind-Beziehungen im ersten Lebensjahr*. Stuttgart: Klett.