Für M. und S.

Jens Kipper

A Two-Dimensionalist Guide to Conceptual Analysis

EPISTEMISCHE STUDIEN
Schriften zur Erkenntnis- und Wissenschaftstheorie

Herausgegeben von / Edited by

Michael Esfeld • Stephan Hartmann • Albert Newen

Band 25 / Volume 25

Jens Kipper

A Two-Dimensionalist Guide to Conceptual Analysis

ontos
verlag

Frankfurt I Paris I Lancaster I New Brunswick

Bibliographic information published by Deutsche Nationalbibliothek
The Deutsche Nastionalbibliothek lists this publication in the Deutsche Nationalbibliographie; detailed bibliographic data is available in the Internet at http://dnb.ddb.de

This book is a revised version of my dissertation which was written (and accepted as such) at the Faculty of Arts and Humanities at the University of Cologne

North and South America by
Transaction Books
Rutgers University
Piscataway, NJ 08854-8042
trans@transactionpub.com

United Kingdom, Ire, Iceland, Turkey, Malta, Portugal by
Gazelle Books Services Limited
White Cross Mills
Hightown
LANCASTER, LA1 4XS
sales@gazellebooks.co.uk

Livraison pour la France et la Belgique:
Librairie Philosophique J.Vrin
6, place de la Sorbonne ; F-75005 PARIS
Tel. +33 (0)1 43 54 03 47 ; Fax +33 (0)1 43 54 48 18
www.vrin.fr

P.O. Box 15 41, D-63133 Heusenstamm
www.ontosverlag.com

ISBN 978-3-86838-141-2

2012

Printed on acid-free paper
ISO-Norm 970-6
FSC-certified (Forest Stewardship Council)
This hardcover binding meets the International Library standard

Printed in Germany
by CPI buch bücher.de

Preface

I became interested in conceptual analysis when working on philosophical thought experiments. Even then it seemed to me that the most plausible account of our judgments about hypothetical cases construes them as resting on ordinary conceptual competence. Accordingly, the method of thought-experimentation can be understood as a type of conceptual analysis. If this is an adequate picture, then conceptual analysis is much more common in philosophical practice than is often acknowledged. Unfortunately, the viability of conceptual analysis as a philosophical method is highly disputed. It is therefore all the more important to set it on a solid theoretical footing.

In my view, the most serious challenge to conceptual analysis has been presented by externalism about linguistic meaning and mental content. Externalists offer an attractive account of meaning which, however, seems to leave little epistemic significance to our conceptual competence. In light of this, it is natural to think that conceptual analysis needs to be underpinned by an equally comprehensive and at least equally appealing semantic theory. This is where epistemic two-dimensionalism comes into the picture. Two-dimensionalism undergirds the method of conceptual analysis in general and our ability to evaluate hypothetical scenarios in particular. At the same time, I believe that it presents an independently plausible account of meaning which incorporates many of the advantages of externalist theories, while avoiding its main disadvantages. The main aims of this book are as follows: Firstly, to defend my interpretation of two-dimensionalism on independent grounds; secondly, to spell out the goals, the promises, but also the limitations of conceptual analysis on the basis of this semantic framework.

This book is a revised version of my dissertation which was accepted as such by the Faculty of Arts and Humanities at the University of Cologne. I would like to thank my supervisor, Thomas Grundmann, for providing me with the opportunity to write it and for his constant encouragement and

support. While working on the thesis, I was supported by the *Institut für Wissenschaft und Ethik* (*IWE*, Institute for Science and Ethics), where I worked in a Junior Research Group funded by the *Bundesministerium für Bildung und Forschung* (*BMBF*, Federal Ministry of Education and Research) on the moral implications of molecular medicine and brain research, headed by Thomas Heinemann; by the *Kölner Gymnasial- und Stiftungsfonds*; and by the *Deutsche Forschungsgemeinschaft* (*DFG*, German Research Foundation), which funded a project on two-dimensionalism and conceptual analysis, headed by Thomas Grundmann. Many thanks to everyone involved! I would also like to thank David Chalmers (and once again the *DFG*) for enabling me a very valuable stay at the Australian National University in Canberra.

I am indebted to the following people for discussions on philosophical issues related to those covered in this book: Alma Barner, Albert Casullo, David Chalmers, Lars Dänzer, Joachim Horvath, Hilary Kornblith, Stephan Kubicki, Kelvin McQueen, Laura Schroeter, Daniele Sgaravatti, Ernest Sosa, Marius Thomann, Clas Weber, and Andrea Wille. Many thanks also to Alma Barner, Kelvin McQueen, Lisa Tambornino, and Andrea Wille for proof-reading, help with formatting, and other editorial stuff.

Introduction **1**
1 What is conceptual analysis and what is the problem? **9**
1.1 What is conceptual analysis? 9
1.2 From Frege to Kripke and Putnam 13
2 Two-dimensionalism and the necessary a posteriori **21**
2.1 Two-dimensionalism 21
2.1.1 Primary and secondary intensions 21
2.1.2 Metaphysical plenitude and two-fold world dependence 25
2.1.3 Scrutability and canonical descriptions 30
2.1.4 Two-dimensionalism and Jackson's descriptivism 34
2.1.5 Two notions of apriority 38
2.2 Modal illusions according to Kripke and according to two-dimensionalism 42
2.2.1 Kripke's two models of modal error 44
2.2.2 Doubts about the accounts of modal error 46
2.2.2.1 Doubts about the epistemic counterpart model 46
2.2.2.2 Doubts about the reference fixer model 53
2.3 Summary and outlook: What has been shown and what is yet to be shown 58
3 The challenge from the epistemic arguments **65**
3.1 Primary intensions and the epistemic arguments 68
3.1.1 The primary intensions of natural kind terms 72
3.1.1.1 Vagueness 80
3.1.1.2 Intersubjective variation and the individuation of concepts . 82
3.1.2 Semantic deference and the primary intensions of names 84
3.1.2.1 The argument from Ignorance and Error 86
3.1.2.2 Deferential concepts and the alleged problem of circularity 87
3.1.2.3 Deferential concepts and apriority 92

3.1.2.4 Two methods for detecting deferential concepts and two problems for two-dimensionalism 94
3.2 Linguistic meaning, mental content, and two-dimensionalism 98
4 Primary intensions, defining the subject, and communication 109
4.1 Defining the subject 110
4.1.1 A case for the epistemic thesis 115
4.1.1.1 From (CJ) to (CJ+) – Schroeter's improv model 117
4.1.1.2 From (CJ+) to (CJ++) 133
4.1.2 The failure of the semantic thesis 137
4.2 Two-dimensional communication 141
4.2.1 The semantic thesis 144
4.2.1.1 Communication involving indexical expressions 146
4.2.1.2 Communication involving proper names 151
4.2.1.3 Communication involving natural kind terms 155
4.2.2 The epistemic thesis 157
4.2.2.1 The importance of shared primary intensions 158
4.2.2.2 How primary intensions help to promote co-reference even when they are not shared 163
4.2.2.2.1 Proper names 164
4.2.2.2.2 Natural kind terms 166
5 Epistemic transparency and epistemic opacity 169
5.1 Arguments for ubiquitous opacity 170
5.1.1 Millikan 171
5.1.2 Putnam 173
5.1.3 Kornblith 175
5.2 Revealing opacity 182
5.2.1 Revealing opacity via the function of a term 183
5.2.2 Revealing opacity via considerations about hypothetical cases 185
5.2.3 Can opacity be determined a priori? 187

5.3 The value of opaque terms in conceptual analysis 194
5.3.1 Discovering essences 195
5.3.2 Variation in primary intensions 196
6 Scrutability, primary intensions, and conceptual analysis 201
6.1 Scrutability and primary intensions 202
6.1.1 From descriptivism to the scrutability thesis 205
6.1.2 A case for (CJ) 211
6.1.2.1 Argument from metaphysical plenitude 212
6.1.2.2 Arguments from the scrutability of specific kinds of facts 213
6.1.2.3 Arguments from the absence of clear counterexamples 215
6.1.2.4 Arguments from the epistemic indispensability of scrutability 218
6.2 Semantic idealizations and epistemic reality 222
6.2.1 Are primary intensions too coarse-grained? 223
6.2.2 Scrutability for real subjects 227
6.2.2.1 The two-level model 233
6.2.2.2 Approaching ideal judgments 242
7 The trouble with definitions and the aims of conceptual analysis 249
7.1 The aims of conceptual analysis 251
7.1.1 Definitions – complete, partial, and absent 251
7.1.2 Reductive explanations 255
7.1.3 The Canberra Plan 262
7.1.3.1 Ramsey sentences, primary intensions and unique reference 266
7.1.3.2 The scope of the Canberra Plan 268
7.1.3.3 The practicability of the Canberra Plan 271
7.2 The trouble with definitions 274
7.2.1 Adequacy conditions for definitions 274
7.2.2 Objections to the eligibility of definitions 278

7.2.2.1 Objections from the relation between definiendum and everything else 278
7.2.2.2 Objections from the format of concepts 282
7.2.3 The absence of successful definitions and some reasons for optimism 285
8 Concluding remarks 291
References 297

Introduction

Conceptual analysis has served as a pivotal method through much of the history of philosophy. And even though it does not always go under that label, conceptual analysis is still ubiquitous in today's philosophical practice. This is despite the fact that a number of serious objections against its tenability have been raised. Among these, externalist theories about meaning and mental content presumably present the most influential challenge to the method's utility. Some of the insights of these theories seem highly compelling, yet they are hard to square with the idea that we can gain philosophical knowledge through an analysis of our concepts.

Epistemic two-dimensionalism, which is a comparatively recent modal theory of meaning, incorporates many of the insights of externalist accounts, while leaving room for an important component of meaning which is a priori accessible to a speaker. Importantly, epistemic two-dimensionalism thereby promises to restore an intimate connection between apriority and metaphysical modality. The two-dimensionalist framework thus seems ideally suited to provide the theoretical basis for conceptual analysis. Indeed, its two most prominent proponents, David Chalmers and Frank Jackson, have argued that conceptual analysis is a viable philosophical method. Jackson even claims that it is indispensable in any kind of inquiry, in philosophy and elsewhere.

It is fair to say, however, that epistemic two-dimensionalism has met with widespread skepticism. In my view, the skepticism is unjustified. Two-dimensionalism does provide a potent account of meaning and content, in terms of a unifying framework which can account for a great number of phenomena associated with language and thought. I take it that there are some reasons why this framework has nevertheless not been more widely accepted: Since two-dimensionalism is still a young theory, its motivation and many of its implications are not yet well understood. There are still a large number of open questions to be answered, both concerning the positive reasons to endorse epistemic two-dimensionalism and how various objections against it can be dealt with.

As part of a defense of the method of conceptual analysis, it will thus be all the more important to shed light on these issues. Accordingly, the first part of this book will mainly turn on the two-dimensionalist framework. More specifically, my goals in the first part (chapters 1–4) will be the following ones: I will expound the theory and elaborate its implications, examine whether it is able to incorporate externalist insights and how it can deal with objections, and I will develop and investigate a number of arguments in favor of epistemic two-dimensionalism.
The development and establishment of two-dimensionalism will serve as a crucial step in my defense of conceptual analysis. However, the utility of conceptual analysis as a philosophical method depends on many other factors apart from the adequacy of the two-dimensionalist framework. In the second part of the book (chapters 5–7), I will identify these factors in order to investigate the use of conceptual analysis in philosophical practice.

Let me outline the structure of the book in more detail:
In **chapter 1**, I will briefly clarify what I mean by conceptual analysis. I will assume a fairly liberal understanding, according to which for a judgment to be the result of conceptual analysis, it only needs to be based on conceptual competence to a significant extent. I will, however, maintain that conceptual analysis must at least involve an important step which is a priori.
Subsequently, I will give a rough sketch of some of the most important semantic theories of the twentieth (and late nineteenth) century, namely those of Gottlob Frege, Rudolf Carnap, Hilary Putnam and Saul Kripke. This brief historical excursion will help to elucidate the foundations and the motivations of the two-dimensionalist framework, since two-dimensionalism stands in the tradition of the theories of Frege and Carnap while being designed to accommodate Kripke's and Putnam's critique of theories of this (internalist) kind.

In **chapter 2**, I will outline the two-dimensionalist framework in detail. I will first spell out its basic ideas and its key theses and explain the way it

incorporates externalist insights into an account which may nevertheless be called Fregean in nature. Its crucial claim is that linguistic expressions are associated with two intensions, called the primary and the secondary intension. The former of these intensions is a priori accessible to a speaker; the worlds involved in it reflect epistemic possibilities. An expression's secondary intension corresponds to intensions as they are conceived by Kripke and can thus be dependent on features which are external to a speaker. According to two-dimensionalism, secondary intensions are (roughly speaking) determined by the corresponding primary intensions relative to a given context.

The thesis that expressions are associated with primary intensions is intimately linked with the scrutability thesis, which states that the grasp of a primary intension is supposed to bestow the ability to determine the extension of an expression with respect to any hypothetical scenario a priori.

Another key component of two-dimensionalism is the thesis of metaphysical plenitude, according to which there is a metaphysical possibility corresponding to every epistemic possibility. In the second part of chapter 2, I will therefore deal with the question of how this thesis can be squared with the existence of a posteriori necessities. I will do so on the basis of a discussion of Kripke's own two models of modal illusions. This approach draws its justification from the fact that Kripke's models are in fact very close in spirit to the two-dimensionalist account of the necessary a posteriori. I will identify the similarities as well as the differences between the accounts. I will then discuss the plausibility of Kripke's theory. With respect to those aspects of Kripke's models which have attracted criticism, I will discuss how the two-dimensionalist account can deal with the objections. It will transpire that there are no decisive objections to Kripke's models and that the two-dimensionalist framework can account for the standard cases of a posteriori necessities.

Finally, I will recapitulate what the two-dimensionalist account does show and what is yet to be shown in a defense of conceptual analysis.

In chapters 3 and 4, I will defend the central two-dimensionalist claim that linguistic expressions are associated with primary intensions. The main aim of **chapter 3** will be to argue that natural kind terms and proper names have primary intensions. This is particularly important since these kinds of terms have been the main target of the so-called epistemic arguments which are supposed to show that such terms have no a priori associations whatsoever.
In the first part of chapter 3, I will concern myself with natural kind terms. Drawing on proposals due to David Lewis, Chalmers and Jackson, I will construe the natural kind term 'water' as a theoretical role term, such that it is a priori that water satisfies the associated theoretical role at least approximately.
In the second part of the chapter, I will try to give a rough approximation of the primary intensions of proper names. Following a proposal of Peter Strawson, Chalmers and Jackson hold that the primary intensions of names should be understood deferentially, i.e. as involving a reference to the use of the term by other speakers. I will argue that this understanding is plausible from a theoretical perspective and in line with our judgments about Kripke's hypothetical cases. Moreover, the deferential understanding of specific kinds of concepts also promises to solve the more general problem of incomplete understanding raised by Putnam and Tyler Burge. It will transpire, however, that the notion of deference thereby raises some other problems. I will argue that in order to account for these problems, primary intensions should first and foremost be taken to represent the content of the thought expressed by an utterance. The final part of the chapter will be dedicated to a discussion of the consequences of this proposal to conceptual analysis.

In **chapter 4**, I will present and discuss a couple of theses due to Jackson about the role of primary intensions in communication and in determining the subject matter. In the first part of the chapter, I will discuss Jackson's semantic thesis that primary intensions define the subject and his epistemic thesis that they are what enables us to determine our subject matter, beginning with the latter thesis. Since Jackson's case for the epistemic

thesis which rests on rather general considerations seems hard to evaluate, I will develop my own version of the thesis. That thesis is based on the idea that we are able to pass judgments in response to empirical evidence. I will argue from there to the conclusion that primary intensions are required to account for this ability. The starting point of the argument will be a comparably weak version of the scrutability thesis which only concerns judgments about the actual world and which does not require apriority.
After this, I will argue that Jackson's semantic thesis that primary intensions define the subject does not apply to all cases and is therefore false. It will nevertheless turn out that even in such cases, constancy of the subject matter can only be ensured by taking the associated primary intensions into account.
The second part of chapter 4 will illuminate the role of primary intensions in communication. Again, Jackson puts forth a semantic and an epistemic thesis: According to the semantic thesis, primary intensions are what is communicated in a conversation; the epistemic thesis states that communication would not be possible without primary intensions. I will discuss and eventually reject the semantic thesis. Since this result threatens the epistemic thesis as well, I will propose two alternative ways to defend the epistemic thesis of communication. The first of these proposals provides an argument for a restricted version of the semantic thesis. The second proposal is independent of the semantic thesis: It aims to show that primary intensions facilitate communication even in those cases where they are not transmitted from speaker to hearer.

Chapter 5 will deal with questions surrounding epistemically transparent and epistemically opaque terms: According to the analyses of natural kind terms and proper names which I propose in chapter 3, those terms are epistemically opaque, i.e. their secondary intensions and thus their metaphysical application conditions are not a priori accessible. The existence of such terms is potentially problematic for the prospects of conceptual analysis. In the first part of the chapter, I will discuss attempts which have been made to undermine the utility of conceptual analysis by

arguing that epistemic opacity is ubiquitous. I will reject these arguments on the basis of the fact that considerations about the function of many terms and about our judgments about hypothetical scenarios involving them are best explained by the assumption that these terms are epistemically transparent.
In the second part of chapter 5, I will consider in more detail how one can determine whether an expression is opaque or transparent. It will eventually emerge that transparency and opacity can be determined a priori.
The final part of the chapter will be devoted to the potential use of opaque expressions for conceptual analysis. I will show that while such expressions are undeniably less suitable for some classical analytic projects, an analysis of their primary intensions can nevertheless be valuable for a number of purposes.

Chapter 6 will be dedicated to various issues concerning scrutability. In the first part of the chapter, I will first have a closer look at the thesis and its relation to the central two-dimensionalist claim that linguistic expressions are associated with primary intensions. After that, I will discuss a number of arguments for and against a (comparably weak) version of the scrutability thesis. I will argue that the considerations in favor of the thesis weigh more heavily than those against it. Since the version of the scrutability thesis defended here is precisely the one which served as the premise in my argument for primary intensions from our ability to determine the subject matter in chapter 4, these considerations will complete my case for primary intensions.
In the second part of chapter 6, I will discuss the idealizations involved in the scrutability thesis and more generally how the truth of the thesis bears on the feasibility of conceptual analysis in philosophical practice. I will argue that while there is undoubtedly a huge gap between the idealized rationality invoked in the scrutability thesis and those conditions present in our epistemic reality, there are no reasons to be particularly skeptical towards our ability to make the judgments in question in a reliable way.

In **chapter 7**, I will identify potential goals of conceptual analysis and the preconditions for realizing these goals. In the first part of the chapter, I will outline a couple of applications of conceptual analysis, such as its use (i) in a search for a category's essence or just for either necessary or sufficient conditions for membership in the category, (ii) in reductive explanations and (iii) in the Canberra Plan. It will transpire that although conceptual analysis can be fruitful even in the absence of explicit analyses, such analyses are nevertheless important for many of the purposes discussed. For this reason, I will then investigate the prospects of the project of providing such explicit analyses. I will argue that there are no principled reasons to think that adequate definitions cannot be had. At the same time, I will identify a number of practical obstacles to that project, which may explain why successful definitions are so hard to come by. In conclusion, I will give some reasons to consider the project of providing definitions as a progressing one.

1 What is conceptual analysis and what is the problem?

1.1 What is conceptual analysis?

In this book, I am going to defend the viability of conceptual analysis as a philosophical method. It therefore seems appropriate to say at least a few words about what I mean by conceptual analysis.

It is usually held that conceptual analysis is essentially a priori. I am actually not sure whether one should consider apriority as a nonnegotiable requirement. However, since it fits well into the general project with which I will be concerned, I can accept the apriority condition for the purposes of this work.[1] Aside from that, I am inclined to adopt a very liberal understanding of conceptual analysis: Any way of trying to gain knowledge – philosophical or otherwise – which is based on conceptual competence will qualify. I do not even want to claim that conceptual analysis has to be based on conceptual competence alone. If it turns out that a priori faculties such as logic and imagination are not part of our conceptual competence, yet are necessary to make the relevant judgments as well, then this will be fine for my purposes.

Let me also note that the term 'conceptual analysis' is used in two slightly different senses. Sometimes, it is used to denote the process of analyzing concepts, while at other times it stands for what is typically considered as the intended result of such an analysis – an explicit analysis or a definition. Throughout this work, I will use the term 'conceptual analysis' to denote the process, whether it aims at an explicit analysis or not. When I am

[1] There are two things to note here, though. Firstly, I will argue in chapter 3 that conceptual analysis can be understood as a two-step process, the second of which is empirical. Accordingly, the apriority requirement only applies to the first step, which is essentially based on conceptual competence. Furthermore, conceptual analysis can be a part of a broader epistemic enterprise which delivers empirical results (cf. also my discussion of the aims of conceptual analysis in chapter 7).

concerned with the second sense of ‘conceptual analysis’, I will speak of ‘(explicit) analyses’ or ‘definitions’.

In my view, the importance of conceptual analysis in philosophical practice is illustrated particularly vividly by philosophers’ reliance on thought experiments. Very often, philosophical theories are tested by checking whether they are compatible with our judgments about hypothetical cases. Far from everyone believes, however, that the evaluation of hypothetical scenarios should be understood as a way of doing conceptual analysis. A number of alternative proposals have been made: Hilary Kornblith argues that we evaluate hypothetical scenarios on the basis of empirical background information (cf. Kornblith 1998). According to Timothy Williamson, judgments about hypothetical cases rely on our everyday ability to evaluate subjunctive conditionals (cf. Williamson 2007, ch. 6). He agrees with Kornblith that these judgments are empirically justified.[2] On the other side, there are philosophers who think that the judgments in question are a priori and who invoke a special faculty, such as rational intuition,[3] to account for this fact (cf. e.g. BonJour 1998). An extreme example of a view of this sort is held by James Brown (cf. Brown 1991).[4] He believes that thought experiments provide us a privileged Platonic insight into the laws of nature.
In my view, none of these explanations of our (purported) ability to evaluate hypothetical scenarios is entirely satisfactory. It does not seem very well motivated, for instance, to assume that we have a special faculty

[2] I am simplifying Williamson’s position a bit here. He believes that some judgments about subjunctive conditionals, and thus about metaphysical modality, are a priori, and many are neither clearly a priori nor clearly a posteriori (cf. Williamson 2007, 165ff.). I think it is fair to say, however, that on his view judgments about hypothetical scenarios in the context of thought experiments will generally come out as a posteriori.
[3] I should note that it is possible to hold that rational intuitions can ultimately be traced back to conceptual competence (cf. Bealer 1998). I do not have any quarrels with such a view.
[4] Notice, however, that he is mainly concerned with thought experiments in science.

which allows us to evaluate hypothetical cases,[5] in particular since the origin and the underlying mechanisms of this alleged faculty are quite mysterious. Kornblith and Williamson, on the other hand, do not postulate any special faculty, which should be considered a definite advantage. However, their accounts are highly revisionary with respect to their understanding of philosophical method which has traditionally been construed as being, at least to a significant extent, a priori. Furthermore, I have serious doubts that our judgments about hypothetical cases, in particular about remote ones, could be considered reliable if they depended on empirical information. And finally, as I will argue in some detail in chapter 5, our modal judgments exhibit a number of characteristics which are best explained by regarding them as a priori.
Construing our evaluations of hypothetical scenarios as instances of conceptual analysis is therefore much more in line with these characteristics, and also with a traditional understanding of philosophical method. Apart from that, one need not thereby postulate the existence of a special a priori faculty, either: On this understanding, our judgments about hypothetical cases are just based on an everyday ability, namely on conceptual competence. In the following chapters, I will say a lot more about the connection between conceptual competence and our ability to evaluate hypothetical scenarios. I will also outline in detail how such an approach, within a two-dimensionalist theory of meaning, promises to provide a general account of modal epistemology according to which we have a priori access to the domain of metaphysical possibilities.

Unfortunately, though, the reputation of the method of conceptual analysis is far from pristine. Many people think that there are decisive objections to its viability. These objections can be divided into two categories, corresponding to the two senses associated with the term ‘conceptual analysis’ outlined above:

[5] On BonJour’s view, rational intuition is required to pass all other kinds of judgments as well, however.

Firstly, there are objections to the idea that it is possible to define philosophically relevant terms. But as I mentioned above, conceptual analysis need not aim at definitions. Therefore, this kind of objection is not suitable for a general attack on the tenability of conceptual analysis as a philosophical method. I will nevertheless address the objection that definitions are not to be had in chapter 7.
Secondly, there are objections to the idea that conceptual competence can be a source of substantial philosophical knowledge. These kinds of objections obviously pose a more principled threat to conceptual analysis as a philosophical method. The arguments of Saul Kripke and Hilary Putnam (cf. Kripke 1980; Putnam 1962, 1970, 1975) which are of this kind are surely among the most influential reasons for philosophers to reject conceptual analysis. The gist of these arguments is that the meaning of an expression and also, more specifically, its reference are not determined by a subject's internal states. Consequently, Kripke and Putnam claim that we do not have a priori access to the application conditions of the expressions we use. Their alternative semantic account emphasizes the importance of environmental and social features for the determination of meaning.[6]
One of the aims of this book is to show that the externalists' attack on conceptual analysis can be parried. What I think their considerations do show, however, is that conceptual analysis needs to be placed on a solid footing in the form of a systematic semantic theory. Such a theory should inter alia give an account of conceptual competence and of the way in which reference and meaning are determined which is at least compatible with the view that conceptual analysis is a way to gain philosophical insights. And as I hope to demonstrate in the following chapters, two-dimensionalism is ideally suited to satisfy these desiderata.

In the following, I will give a brief sketch of Gottlob Frege's and Rudolf Carnap's theories of meaning, and of Kripke's and Putnam's own accounts and their critique of internalist theories. The description of Frege's and Carnap's theories will provide the background for Kripke's and Putnam's

[6] For the relevance of the latter kind of features cf. also Burge 1979.

accounts, which will be outlined subsequently, in the following two respects: It will present the kind of internalist theories which they oppose and it will motivate the intimate connection between meaning and modality on which their arguments rely. Taken together, these considerations will in turn provide the background for the two-dimensionalist account of meaning which I will outline in the following chapter.

1.2 From Frege to Kripke and Putnam

Frege introduces the notion of sense (*Sinn*) in the context of a problem concerning identity statements (in *Über Sinn und Bedeutung*, cf. Frege 1892/2002): What does an identity statement of the form 'a = b', such as 'the morning star = the evening star' express? If the terms involved were only associated with a referent (*Bedeutung*), then such a statement, if correct, could only say that a certain object (in this case, the planet Venus) is identical with itself. But this is hardly plausible since statements of the form 'a = b', unlike those of the form 'a = a', are typically of cognitive value. After all, a competent speaker need not know that the morning star and the evening star are actually the same celestial body. From this, Frege concludes that a term does not only have a referent, but also a sense. A sense is primarily a mode of presentation of the corresponding referent. By introducing senses, Frege can explain why it can be a genuine insight to realize that morning star and evening star are identical. Figuratively speaking, when one looks at the same object twice but from different perspectives, one need not be aware that one saw the same object on both occasions (even if one's memory works perfectly).

Frege holds that while there can be many senses corresponding to one referent, i.e. many ways the referent is presented to us, to each sense there corresponds only one referent. Furthermore, Frege argues that whole sentences also have a sense and a referent: The sense of a declarative sentence is a thought; its referent is a truth-value. But how are senses to be individuated? From Frege's considerations about identity statements mentioned above, one can derive a criterion of identity for singular as well

as for general terms: Two terms ‘a’ and ‘b’ have the same sense if ‘a = b’ is cognitively insignificant. Consequently, identity conditions for the senses of sentences, i.e. thoughts, are also tied to cognitive significance. In *Über Sinn und Bedeutung*, Frege says that the two sentences ‘The morning star is a body illuminated by the sun’ and ‘The evening star is a body illuminated by the sun’ do not express the same thought since someone could simultaneously consider one of these sentences to be true and the other one to be false (cf. Frege 1892/2002, 29). Let me thus say that when two sentences S_1 and S_2 express the same thought, then one cannot believe S_1 without believing S_2 and vice versa.[7] At least in one sense, these identity conditions for senses can be considered as criteria for synonymy.

Carnap replaces Frege’s distinction between sense and referent by the notions of intension and extension (cf. Carnap 1947/1956).[8] On Carnap’s account, the intension of a singular term is what he called an ‘individual concept’; its extension is the denoted object. The intension of a predicative expression (a ‘predicator’) is a property, its extension the class of entities having that property. Finally, the intension of a sentence is a proposition and its extension a truth-value. Up to here, this account does not seem to differ too much from Frege’s. The crucial feature of Carnap’s semantics is that he ties intension to modality, in the following way: First he defines L-truth (logical truth) as truth in all state-descriptions. Since a state-description is supposed to be an explication of the notion of a possible world, an L-truth is a necessary truth. Then Carnap states that two expressions have the same intension if and only if they are L-equivalent, i.e. if and only if they have the same extension with respect to all possible worlds. Today, an intension is usually defined as a function from possible

[7] In my wording, this is a necessary condition for the identity of thoughts. I think it can be argued that it is also a sufficient condition, but this will not matter for my purposes here.

[8] Carnap actually believed that Frege’s notion of referent faces more serious problems than his notion of sense (cf. Carnap 1947/1956, 129ff.). Nevertheless, for my purposes it will be more important to highlight the differences between senses and intensions.

worlds to extensions. Although this is not quite the way Carnap puts it, it is at least compatible with his account just described.

Intensions are less fine-grained than senses: For instance, since '$2^7 = 128$' is necessarily true, '2^7' and '128' have the same intension. But obviously, it can be of cognitive value to be told that $2^7 = 128$ and thus, the two expressions do not have the same sense.[9] This illustrates that intensions are not connected with cognitive significance, but rather with apriority, since Carnap believed that all necessary truths are analytic and thus can be known a priori. Although this suggests that intensions are more remote from an intuitive notion of meaning than senses, Carnap's account does have some advantages over Frege's: It is not altogether clear what exactly it takes for a statement to be cognitively significant. Moreover, cognitive significance seems to be highly subject-relative: '$2^7 = 128$' is plausibly cognitively significant for some subjects, but not for others. And even if it was possible to give a more precise and objective account of cognitive significance and thus of sense, the notion would still be too fine-grained for the normative purposes which a semantic theory should arguably serve as well.[10]

In defining intensions, Carnap is thus able to provide more precise and more objective identity conditions than Frege. But what makes it sensible to connect meaning with modality in the first place?

Firstly, there are general considerations concerning language and information which speak in favor of such a picture. Sentences (and thoughts) represent states of affairs, i.e. they carry information about the world. Information can be defined as the exclusion of alternatives: The more information a signal carries, the more alternatives it excludes. (Note that the unit in which information is commonly measured is a Bit, where one Bit stands for a binary alternative.) If you are told that, say, the person who stole your car was female, this will provide you with less information

[9] Carnap is aware of this and suggests that synonymy might rather be connected with 'intensional isomorphism' which corresponds roughly to sameness in intension on the level of an expression's constituents (cf. Carnap 1947/1956, 56).

[10] I will say more about these issues in chapter 6.

than if you are told that it was a six foot tall female, because the latter proposition rules out more potential suspects. Arguably, such alternatives can simply be construed as possibilities. Thus, if one believes, as many do, that a theory of meaning is supposed to account for the informational or representational content of linguistic expressions, it seems consistent to do this by means of a possible worlds framework.

Secondly, a possible worlds account can preserve many of the merits of Frege's theory compared to a purely extensional theory of meaning. There are expressions which have no extension, but which nevertheless seem to have a meaning, such as 'unicorn'. Frege can account for this by insisting that the term does have a sense. It also has an intension, which can be represented by the set of possible worlds where there are unicorns – since, although unicorns do not exist, they could have possibly existed.[11] Another problem for an extensional semantics is posed by expressions which have the same extension but seem to differ in meaning, such as the above mentioned 'morning star' and 'evening star' or 'Joachim Sauer's second wife' and 'the first female Chancellor of Germany'. Frege would hold that the two expressions differ in sense, as witnessed by the fact that the statement 'Joachim Sauer's second wife is the first female Chancellor of Germany' is (or at least can be) of cognitive value.[12] Clearly, the expressions also differ in intension, since Joachim Sauer could have never got married. Then there is the problem of so-called intensional contexts, such as belief sentences: If meaning is just extension, how can someone believe that the first female Chancellor of Germany is important without believing that Joachim Sauer's second wife is important? Once again, this problem can be solved if one acknowledges that there are two different beliefs involved because they differ in sense, or intension.

---

[11] Kripke is famously skeptical regarding the possibility of there being unicorns (cf. Kripke 1980, 157f.). I will ignore this complication here.

[12] The problem is more pressing for a purely extensional account when *simple* co-referring expressions are involved. For the purposes of illustration, my example will do as well.

The broadly Fregean account of meaning just sketched thus seems very attractive. However, many philosophers think that it was severely shaken, if not refuted, by Kripke's and Putnam's arguments. The most influential of these arguments rely on modal considerations. In the remaining part of this chapter, I will focus on these. Against this background, it will be easier to grasp the ideas underlying two-dimensionalism, which will be outlined in the following chapter. Since Frege himself did not say anything about the modal implications of his theory, it is not in all cases clear that his account is threatened by the arguments which will be discussed below. But these arguments clearly target a number of theories in the tradition of Frege, including Carnap's. Kripke and Putnam also invoked non-modal arguments against internalist theories of meaning. Those arguments potentially undermine any kind of broadly Fregean account. These so-called epistemic arguments will be discussed in chapter 3.

In *Naming and Necessity* (1980), Kripke argued specifically against the description theory of reference for proper names, which also stands directly in the tradition of Frege. According to descriptivism, the reference of an expression is determined by a description which a speaker associates with that expression. This description is also supposed to give the expression's meaning. But in his so-called modal arguments, Kripke pointed out that names and the definite descriptions which speakers could associate with them are not modally equivalent. Take the name 'Aristotle'. One speaker could think of Aristotle as the teacher of Alexander the Great, another as the last great philosopher of antiquity. However, Aristotle could have died early or spent his life as a shepherd, in which case he would not have satisfied any of these descriptions. This shows that with respect to these possible worlds, 'Aristotle' does not refer to the same person as 'the teacher of Aristotle' or 'the last great philosopher of antiquity'.

Putnam reaches a similar conclusion with respect to natural kind terms, in his famous 'Twin Earth' thought experiment (cf. Putnam 1975): Suppose that there is a remote planet, Twin Earth, which is very similar to Earth. The liquid in Twin Earth's rivers and lakes resembles our water in all of its superficial properties. However, this liquid has a different molecular

structure, which we abbreviate as XYZ. According to Putnam, the liquid on Twin Earth would not be water. Then he invites us to imagine two speakers in 1750, one of them from Earth and one from Twin Earth. They know nothing to distinguish H_2O from XYZ, but still, when they both say 'water', the Earthling's utterance refers to H_2O, and the Twin Earthling's utterance to XYZ. Accordingly, if there is a sense connected with the term 'water' which is grasped by the speakers, then by all appearances, it cannot determine the reference.

Furthermore, since the intension is supposed to pick out the extension of a term in any given world, the intensions of both proper names such as 'Aristotle' and of natural kind terms such as 'water' are not accessible to a speaker. And therefore, given the intimate connection between meaning and modality (which the notion of intension is supposed to capture) nothing internal to a speaker can determine the reference of an expression. The underlying reason is that proper names and natural kind terms are, in Kripke's terminology, rigid designators, i.e. they pick out the same individual or kind in every possible world.[13] But speakers typically only associate contingent properties with these individuals or kinds. One of the most remarkable consequences of this insight is the fact that there are necessary truths which can only be known a posteriori. Typical examples are identity statements involving two rigid designators, for example 'Hesperus = Phosphorus', 'water = H_2O', or 'heat = mean molecular kinetic energy'.

Kripke and Putnam conclude that reference is determined by features external to the subject, in particular by causal relations which need not be accessible to a speaker. In light of this, the idea that we can gain genuine philosophical insights by way of pondering on our concepts appears highly dubious. And indeed, as I mentioned at the outset these arguments for semantic externalism were taken by many as the primary reason to reject conceptual analysis.

[13] It is not clear whether it is theoretically fruitful to apply the notion of rigid designation to general terms (cf. e.g. Soames 2002). I will ignore these complications here, though.

One of the most attractive features of two-dimensional semantics is that it integrates many insights of Putnam and even more so of Kripke into a systematic semantic account which can still be called broadly Fregean. It therefore offers hope concerning the prospects of conceptual analysis as a philosophical method even in light of this critique. How two-dimensional semantics tries to accomplish this feat will be outlined in the following chapter.

2 Two-dimensionalism and the necessary a posteriori

As the name suggests, two-dimensional semantic theories posit two dimensions of meaning. There are many versions of two-dimensional semantics, which differ concerning the scope of the theory, the nature of the two dimensions and how they are related. Some of these versions, for instance, are only applicable to a narrow range of linguistic expressions and/or do not treat the first dimension as a full-fledged semantic value. None of these accounts is suitable to (re-)establish the viability of conceptual analysis as a philosophical method. But there is one version of a two-dimensional theory of meaning which seems ideal for this purpose: According to what David Chalmers calls 'epistemic two-dimensionalism', every linguistic expression is connected with a semantic value which is a priori accessible to a speaker. In this book, I will therefore focus on this version of two-dimensional semantics (which I will henceforth simply call 'two-dimensionalism'), which is advocated by Chalmers (cf. Chalmers 2002b, 2004, 2006) and Frank Jackson (cf. Jackson 1998a, 1998b, 2004).[14] The following presentation of two-dimensionalism will draw heavily on their work.

2.1 Two-dimensionalism

2.1.1 Primary and secondary intensions

According to two-dimensionalism, every expression which is a candidate for having an extension is associated with two intensions, which correspond to two different ways of considering a possible world: One can consider it as actual, for example by asking questions like 'What if we find out / What if it turns out that the world is like this?'. Or one can consider it

[14] For a comprehensive discussion of other two-dimensional accounts cf. Chalmers 2006.

as counterfactual, by asking questions like ‘What if the world had been like this?’. Accordingly, the primary intension (or, in Jackson’s terminology, the A-intension) of an expression is a function from worlds considered as actual to extensions. Its secondary intension (or C-intension) is a function from worlds considered as counterfactual to extensions.

This second way of considering a possible world corresponds to a common understanding of various Twin Earth scenarios. As applied to Putnam’s original thought experiment, one asks: Given that the liquid in our rivers and lakes is H_2O, is a scenario where the rivers and lakes contain XYZ one in which they contain water? That is, the character of the actual world is taken to be fixed, and we are then invited to regard the hypothetical scenario in question as counterfactual. If we consider a world this way then no matter what kind of scenario we conceive of, we will always find that water is (nothing else than) H_2O. This suggests that secondary intensions are connected with metaphysical necessity, in the following way:

(2D1) A sentence S has a necessary secondary intension iff S is metaphysically necessary.

The secondary intension of an expression is thus simply its post-Kripkean intension: The secondary intension of ‘water’ picks out H_2O in all worlds, which implies that the secondary intension of ‘water = H_2O’ is true with respect to all worlds. But since it is not a priori that water is H_2O, secondary intensions need not be accessible a priori.

Primary intensions work quite differently. Let us suspend our empirical knowledge about the molecular structure of water for a moment and imagine a situation where we are just about to make a chemical analysis of various samples of water. At least from this point of view, it could turn out that water is XYZ. This indicates that if we consider a Twin Earth scenario not as counterfactual, but rather as actual we get a different result. The scenario corresponds to the epistemic possibility that water is XYZ. Here, ‘epistemic possibility’ has to be understood in a broad sense – it is compatible with everything we can know a priori. Primary intensions are thus tied to apriority:

(2D2) A sentence S has a necessary primary intension iff S is a priori.

However, it may still be a bit unclear what it means to say that the world could turn out to be a certain way and whether this notion is really connected with apriority. For example, when the schoolboy is asked to calculate 2^7, then from his point of view, it could turn out to be 64, or 256, or whatever. The same is true more generally for complex mathematical statements: Before we have determined their truth-value, they could turn out either way. What is thus required is an idealized notion of epistemic possibility, where something could turn out to be the case if it is not ruled out even by ideal rational reflection. According to Chalmers, a sentence S is true with respect to a world considered as actual, and thus epistemically possible, if and only if there is a D such that D epistemically necessitates S. Here, D stands for a canonical description of the world in question in a specific kind of vocabulary – I will say more about the characteristics of such a description later on. The notion of epistemic necessitation again involves an idealization. It can roughly be taken to mean that given D, ideal rational reasoning would lead a thinker to conclude S (or alternatively, that given D, a thinker *should* conclude S).[15] In this case, in Chalmers' terms the world which corresponds to D verifies S.

The two-dimensional account of meaning can be represented in a matrix. Here is the matrix for 'water is H_2O':

[15] Chalmers discusses this issue in much more detail, for instance in Chalmers 2006, sections 3.3 and 3.9.

'water is H_2O'	w_1	w_2	w_3
w_1 (WS: H_2O)	T	T	T
w_2 (WS: XYZ)	F	F	F
w_3 (WS: ABC)	F	F	F

(Figure 1)

The worlds on the left are worlds considered as actual; the worlds on the top are worlds considered as counterfactual. 'WS' stands for 'watery stuff' (this is again Chalmers' term, cf. e.g. Chalmers 1996, 57), which can be taken as an abbreviation of a description like 'the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape'. One can see that in w_1, this liquid's molecular structure is H_2O. w_1 can thus be taken to be the actual world. w_2 is like the Twin Earth world: Here, the molecular structure of the liquid which satisfies the description is XYZ; etc. Accordingly, the top row represents the secondary intension of 'water is H_2O', or maybe rather the set of extensions of the expression with respect to different worlds considered as counterfactual. 'Water is H_2O' thus has a necessary secondary intension, which is in line with Putnam's insight that water is necessarily H_2O. Obviously, to determine the secondary intension of the expression one has to have empirical knowledge. To stay in the picture, one has to know in which row one is located, i.e. one must know which of those worlds on the left really is the actual world.

As I said before, two-dimensionalists claim that there is another semantic value – another intension – which is accessible a priori. The basic idea is that we have (implicit) *conditional* a priori knowledge of the following kind: We know that if w_1 is the actual world, i.e. if the watery stuff in our

world is H_2O, then water is H_2O. We can also say that if w_2 – the Twin Earth world – is actual, then water is XYZ; etc. This putative a priori knowledge is mirrored in the diagonal from the top left to the bottom right of the matrix, which thus represents the primary intension. One can see that the primary intension of 'water is H_2O' is contingent, i.e. its extension varies from world to world. This reflects the fact that it is not a priori that water is H_2O – it is epistemically possible that water is XYZ, or ABC, or whatever.

From (2D1) and (2D2), one can straightforwardly infer that all necessary a posteriori truths have this two-dimensional structure. Thus:

(2D3) A sentence S is necessary a posteriori iff S has a contingent primary intension and a necessary secondary intension.

Likewise, one can infer the following for contingent a priori truths:

(2D4) A sentence S is contingent a priori iff S has a necessary primary intension and a contingent secondary intension.

2.1.2 Metaphysical plenitude and two-fold world dependence

Two-dimensionalism thus posits an a priori accessible dimension of meaning even for a term like 'water' which is among the prime examples invoked by semantic externalists and whose involvement gives rise to necessary a posteriori truths. Although the post-Kripkean (secondary) intensions of 'water' and of 'H_2O' pick out the same substance in every possible world, there is still an intuitive sense in which they do not have the same meaning. This intuitive difference in meaning is straightforwardly captured by two-dimensionalism, exploiting the fact that 'water' and 'H_2O' do not pick out the same substance with *epistemic* necessity. Thereby, two-dimensionalism establishes a semantic value in the tradition of Frege's sense.

However, even if that much is granted, if one wants to draw any conclusions concerning the viability of conceptual analysis one has to face an obvious objection. So far, I have always talked of the scenarios in the

first dimension, i.e. those which constitute the primary intension, as worlds. But calling them 'worlds', so the objection goes, is already misleading. We saw for instance that the primary intension of 'water' assigns to some of those so-called worlds XYZ as extension, or ABC, or … The lesson we should have learned from Putnam and Kripke is precisely that there are no worlds where water is XYZ.

One can of course construe the worlds in the first dimension as purely epistemic possibilities, bearing no deeper connection with metaphysical possibility. In fact, in more recent writings Chalmers does propose such a version of two-dimensionalism (cf. Chalmers 2006, section 3.4.2; Chalmers 2011). For some purposes, such an account may be useful – for example to model the cognitive value of a linguistic expression for the subject, i.e. from her subjective point of view. But it is questionable whether it can serve as a basis for doing conceptual analysis, as usually understood. I take it that the aim of conceptual analysis is to gain insight into what is *really*, i.e. metaphysically, possible or necessary. If for instance conceptual analysis can only reveal that it is epistemically possible that water is not XYZ (which is just another way of saying that it is not a priori that water is not XYZ), then this hardly teaches us anything about water – after all, two-dimensionalists usually agree that water could not have really been XYZ.

To illustrate this problem: Kornblith, as a reliabilist who is also one of the most vigorous critics of conceptual analysis (cf. Kornblith 1998, 2002, 2007a) can agree with an epistemic internalist that it is not a priori that knowledge is reliably produced true belief. It may for instance be epistemically possible that a reliably produced true belief is not knowledge or even that some version of an internalist account of justification is true. But still, Kornblith will insist that knowledge is necessarily reliably produced true belief. He says that as an epistemologist, he is not interested in our concepts of justification and knowledge, but in justification and knowledge themselves. Therefore, the mere epistemic possibility of his account being wrong does not concern him.

These considerations indicate that while the existence of a semantic value which is (a priori) accessible to a speaker is plausibly a necessary condition

for the viability of conceptual analysis, it is not a sufficient one. If the possibilities which make up the first dimension are construed as merely epistemic possibilities, then two-dimensionalism can hardly help to vindicate conceptual analysis as a valuable philosophical method. Thus, if one's aim is to defend conceptual analysis, one seems committed to holding that the scenarios involved in an expression's primary intension correspond to genuine possible worlds. Chalmers postulates such a correspondence in the following principle:

Metaphysical plenitude: For all S, if S is epistemically possible, there is a centered metaphysically possible world that verifies S. (Chalmers 2006, 82)

At first glance, this thesis stands in direct conflict with the conceded fact that sentences like 'water is XYZ' express epistemic possibilities, but not metaphysical possibilities. However, things are not that simple, as is witnessed by the fact that Kripke himself may quite plausibly be taken to endorse something at least roughly like this principle. I will discuss this issue in more detail in 2.2; for now I will just outline the basic idea. Contrary to what his considerations were taken to imply by many, Kripke argues in *Naming and Necessity* that whenever we conceive of something, what we conceive corresponds to some metaphysical possibility. It is just that sometimes, we have not conceived quite what we think we have (cf. Kripke 1980, 142ff.). In cases where rigid designators are involved, we are prone to fall prey to what has been called a 'proposition confusion' (cf. Stoljar 2006). That is to say, what we have conceived is not really, say, that water is XYZ, but something else. So what is this something else? Let us go back to the two-dimensional matrix of 'water is H_2O': There, w_2 was taken to be a world where the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape is XYZ. This surely represents a metaphysical possibility. Putnam's Twin Earth scenario itself describes precisely such a world, and it has rarely been suspected to be impossible.[16] Similarly, Kripke argues that when it seems to us that

[16] Except for the fact that no Twin-Earthling would be an exact duplicate of a person from Earth, since inhabitants of Twin Earth would largely consist of XYZ.

Hesperus could not have been Phosphorus, we really conceive of a world in which the brightest object visible in the morning sky is not identical to the brightest object visible in the evening sky – which could have surely been the case. The upshot is that the principle of metaphysical plenitude is compatible with Putnam's and Kripke's discoveries. The epistemic possibility that water is not XYZ corresponds to a genuine metaphysical possibility. It is just that this possibility is misdescribed by 'water is XYZ'.
Metaphysical plenitude raises another kind of problem in connection with indexicality. Which possible world verifies the epistemic possibility that I am famous, or that it is cold here, or that it is now 10 p.m.? It is well-known that even a complete objective characterization of a possible world is insufficient to settle all questions involving indexical expressions (cf. Castañeda 1967; Lewis 1979; Perry 1979). Therefore, the possible worlds in the first dimension should be construed as centered worlds. The notion of a centered world dates back to W.V.O. Quine (cf. Quine 1969). A centered world is simply a world with a marked individual at a time. 'I am famous' is thus verified by a centered world if the individual at the center of the world is famous; 'It is cold here' is verified if the individual at the center is located at a cold place; etc.

On the current reading of two-dimensionalism, the worlds involved in both the secondary and primary intension should be understood as metaphysically possible worlds. Both intensions thus range over the same space of possibilities, the only difference being that the worlds in the first dimension have a marked center. Accordingly, the difference between primary and secondary intensions is just the one mentioned at the beginning of this chapter: In the primary intensions, the worlds are considered as actual; in the secondary intensions, they are considered as counterfactual. This may still appear a bit puzzling: At least in some cases, the two intensions assign different extensions to the same possible world. But how is this possible, i.e. why should the extension of an expression with respect to a world be dependent so to speak on the perspective from which one considers the world? One way to explain this is by pointing out a

peculiar feature of names and natural kind terms, i.e. those terms which give rise to necessary a posteriori truths. I already mentioned that determining the extension of such terms with respect to a (counterfactual) world requires empirical knowledge. This shows that such a term's extension is dependent not only on the character of the world to be evaluated, but also on the actual world. In the case of names, this is explained by the fact that they are rigid designators: In every world, a name picks out the individual which it picks out in the actual world. Something similar is true for natural kind terms: In every world, they pick out the kind which they pick out in the actual world, regardless of the superficial properties of that kind in the world considered.

It may be worthwhile to have a closer look at the semantics of these terms in order to figure out what exactly it is about the actual world that determines their extension across all possibilities.[17] In the case of names, Kripke argued that their reference depends on causal chains. More precisely, he said that all that matters is the *actual* causal chain which connects *our* use of the name with its bearer. Whenever we use a name, we thus refer to the individual at the beginning of that chain. Natural kind terms seem to function quite similarly. They refer to the kind with which we are actually acquainted across all worlds. For this reason, the term 'water' does not refer to XYZ, even if we consider Twin Earth not just as a counterfactual world, but rather – like in Putnam's original scenario – as a remote planet in our galaxy. If this is correct, then names and natural kind terms behave similarly to indexicals: their extension with respect to a world is dependent on the actual referent which in turn depends on its relation to us, or more precisely to the speaker. In fact, Putnam himself argued in *The Meaning of 'Meaning'* that 'water' is an indexical (cf. Putnam 1975, 233f.; cf. also Haas-Spohn 1997). This suggests that the centering of the worlds in the first dimension is not only required to evaluate indexical expressions, but also for names and natural kinds. Imagine a world like the one described in Putnam's Twin Earth thought experiment. There are two

[17] I will try to spell out the primary intensions of names and natural kind terms in more detail in chapter 3.

'watery stuffs' on two different planets. In order to determine the extension of the term 'water', one needs more than an objective characterization of the world: One additionally has to know which of these substances we are acquainted with. Another thing to note in this context is that it is surely no coincidence that two-dimensional accounts have often been used to describe the semantics of indexical (or otherwise context-sensitive) expressions (cf. e.g. Kaplan 1989): In such accounts, the two-dimensional matrices are supposed to capture how content, i.e. the horizontal, varies with the context of use.

Two-dimensional semantics can straightforwardly capture the twofold world-dependence of specific kinds of terms. The two-dimensional structure reflects the fact that, as was just shown, the extension of an expression is dependent both on the world to be evaluated and on the actual world. In fact, the notion of a counterfactual world already contains this idea: A world can only be counterfactual relative to another – the actual – world. In the two-dimensional framework it is possible to hypothetically consider any world as actual. This can be useful since, although of course only one world is actual, without the relevant empirical information we do not know which of all the possible worlds we inhabit.

After what has just been said one can see why considering the same world in different ways can lead one to assign different extensions to it: It is because some terms make an implicit reference to the actual world, or to be more precise to our relation to the referent.

2.1.3 Scrutability and canonical descriptions

In my discussion of the two-dimensional matrix of 'water is H_2O', I mentioned that according to two-dimensionalism speakers have a specific kind of conditional knowledge: Although we do not know a priori that water is H_2O, we do know a priori that if the drinkable colorless liquid in rivers and lakes which sometimes falls from the sky in drop shape is H_2O, then water is H_2O. According to Chalmers and Jackson, this conditional

ability to identify the extension of 'water' given specific empirical information can be generalized:

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension. (Chalmers & Jackson 2001, 323)

In the background of this thesis is a very important idea which Chalmers calls the 'scrutability of truth and reference' (which may thus just be called the 'scrutability of extension') (cf. e.g. Chalmers 2002a). The idea is roughly that once we are given complete information about the character of the actual world, we are in a position to determine the referents of all of our expressions and the truth-values of all sentences.[18] Without any further specifications, this thesis does not sound very bold. If I am told for example that 'water' refers to H_2O, or that 'water is XYZ' is false, it is easy to determine the extension of 'water' or of 'water is XYZ'. To give the thesis its bite, one therefore has to say more about what 'sufficient' or 'complete' information amounts to. I already mentioned in passing that possible worlds can be characterized by the canonical description D. There are two basic constraints on the vocabulary used in D. Firstly, in order to avoid triviality it must be a limited vocabulary. And secondly, it must not contain terms like 'water' or 'Hesperus'. The reason for this qualification is that a sentence like 'the oceans are filled with water' may metaphysically determine the nature of the substance in the oceans; but it is epistemically compatible with there being H_2O, or XYZ, or … Consequently, the terms used in D must be such that their secondary application conditions are a priori accessible; let me call such terms 'epistemically transparent'. A term can only be epistemically transparent if its primary and its secondary

[18] Mainly due to potential problems regarding indeterminacies of reference, Chalmers is more confident in endorsing the scrutability of truth, though (cf. e.g. Chalmers 2002a, 174f.).

intension coincide;[19] let me call those terms whose two intensions are equivalent 'semantically neutral'.[20]

The scrutability thesis thus states that given a complete description of the actual world in a limited and semantically neutral vocabulary – let me call this description $D_@$ –, one is in a position to know all truths. Before I look at a proposal as to what such a vocabulary could look like, it still has to be clarified what it means to be in a position to know all truths. Once again, the notion involves an idealization: No actual subject would be able to grasp a complete description of our world, let alone draw all the required inferences from it. One can try to spell out what it means to be in a position to know something with the help of the notion of epistemic necessitation, which was introduced above: A given body of evidence E puts a subject in a position to know a true sentence S if E epistemically necessitates S, i.e. if given E and given ideal rational reflection, a subject would conclude S. The scrutability thesis can now be phrased as saying that for all S, if S is true, $D_@$ epistemically necessitates S.

In *Conceptual Analysis and Reductive Explanation* (2001), Chalmers and Jackson make a quite ambitious claim as to what the vocabulary used in $D_@$ might look like. They claim that a conjunction of the following kinds of information is sufficient to know all truths: microphysical information, information about phenomenal states, indexical information, and a concluding clause, which basically says that the world contains nothing beyond what has been stated in the description. Chalmers and Jackson call this conjunction PQTI (for **P**hysics, **Q**ualia, **T**hat's all, **I**ndexicals). It should be noted, however, that neither two-dimensionalism in general nor the scrutability thesis in particular are committed to the thesis that PQTI epistemically necessitates all truths. One could add a lot more (explicit)

[19] As I will argue in chapter 5, there is no such entailment in the opposite direction – there can thus be semantically neutral expressions which are not epistemically transparent.

[20] Here I deviate slightly from Chalmers' usage of the term, whose 'semantic neutrality' is rather supposed to capture what I call epistemic transparency (cf. his discussion in Chalmers 2006, 86ff.).

information to $D_@$ and still have, depending on one's purposes, a reasonably narrow (nontrivial) 'scrutability base'. There is also no compelling reason to assume that there is some privileged vocabulary which has to be used in $D_@$ – there may thus be several kinds of descriptions which satisfy the scrutability thesis.

Chalmers and Jackson hold that the scrutability thesis is not only true with respect to the actual world, but also with respect to other worlds which we can hypothetically consider as actual. Given the scrutability thesis, this seems to be a reasonable assumption: If the entailment from D to some S is a priori anyway, then why should it matter whether D describes the actual world or some other possible world?[21] This idea again goes nicely with the thesis that mastering a term or understanding a complex expression, i.e. grasping its primary intension, is connected with a specific kind of conditional a priori knowledge, as mentioned above: We are not only able to judge that since the watery stuff on Earth is H_2O, water is H_2O. We can also say that if the watery stuff on Earth had turned out to be XYZ, water would be XYZ, etc. That is to say that we have a priori knowledge of a kind that enables us to determine the extension of an expression with respect to different worlds considered as actual, if provided with sufficient nontrivial information about that world.

One minor complication is that if the relevant world is sufficiently remote from ours, then we need a different kind of vocabulary. PQTI is insufficient if a world contains (additional) nonphysical or nonphenomenal ingredients. Another thing to note is that it is most probably not a good idea to hold that there is a sensible scrutability statement with respect to all possible worlds. Why should there for instance not be worlds where only fundamental properties are instantiated – or even worlds with just one fundamental property? In this case, any description of the world would either have to explicitly describe this fundamental property, thus rendering the scrutability thesis trivial, or rather pointlessly somehow encode the information.

[21] Laura Schroeter nevertheless rejects this seemingly innocuous step. Her reasons for this will be discussed in chapter 4.

I just gave an outline of the two-dimensionalist framework, its key theses and its motivation. Before I turn to a more thorough discussion of the Kripkean and the two-dimensionalist accounts of the necessary a posteriori in 2.2, let me address two issues related to the two-dimensionalist framework which still deserve clarification: Firstly, the relation between two-dimensionalism and Jackson's descriptivism, and secondly, the notion of apriority implicit in the two-dimensionalist account.

2.1.4 Two-dimensionalism and Jackson's descriptivism

Jackson explicitly states that he adheres to a two-dimensional account of meaning (cf. e.g. Jackson 2004). However, elsewhere he says that he is committed to descriptivism (cf. e.g. Jackson 1998b). The combination of these views may seem surprising. In the following, I will try to spell out how they go together.

It is important to note that Jackson does not hold that the reference of an expression is determined by an associated description, where a description is understood as a linguistic expression. He understands descriptivism as the thesis that reference is determined by associated *properties*. But what exactly does this claim amount to? For a start, what counts as a genuine property is a hotly debated metaphysical question. But I think it is obvious that one should understand 'property' in a more liberal sense here, since there is no reason to expect that we associate words only with particularly 'natural' or fundamental properties. Jackson's take on this issue thus seems reasonable: On his view, what is required for a term to be associated with a property is just for the term to be associated with some kind of pattern, however grue-like (cf. Jackson 2007, 140; Jackson 1998b, 202). Given these considerations, let me propose to understand a property here simply as a set of extensions across possible worlds.[22] Taken this way, to say that an expression is associated with a property is equivalent to holding that it

[22] This is of course not to propose a general reductive account of properties.

has an intension, i.e. a function from possible worlds to extensions. Unlike in traditional descriptivist accounts, this thesis does not imply that the associated properties have to be expressible by a linguistic description. In fact, Jackson does argue that it should be possible to put the associated properties into words,[23] but I think this view should not be considered as an essential part of his descriptivism.

A descriptivist obviously wants to claim more than just that every linguistic expression has an intension. What is required is that the relevant property is associated with the expression by competent speakers. But it still has to be clarified what it means for a speaker to associate an expression with a property. According to Jackson, a speaker need not be able to articulate the associated properties. In most cases, her knowledge of these properties is only implicit. Jackson likens this to a card player's knowledge of the card game's rules or a competent speaker's knowledge of the rules of grammar (cf. Jackson 1998b, 211f.; Jackson 2004, 272f.; Jackson, Mason & Stich 2009). The player need not be able to spell out all the rules of the game, just like the average speaker would be at loss to give a detailed description of the grammar of her language. Nevertheless, they do follow the relevant rules. Take our average speaker: She is capable of forming grammatically correct sentences and of distinguishing grammatical expressions from ungrammatical ones. If we transfer this to the case of associated properties, then implicit knowledge would amount to being able to tell whether the expression applies to a particular situation or not, i.e. to determine its extension with respect to that situation.[24]

If one generalizes this idea and abstracts away from a subject's cognitive limitations, the resulting thesis is already quite close to Chalmers and Jackson's account of concept possession quoted above, according to which

[23] Cf. e.g. Jackson 2005. Cf. also my discussion in 7.2.

[24] Jackson holds that the competent speaker cannot only classify a given sentence as grammatical or ungrammatical, but that she can also give a reason for this; and apparently, he thinks that implicit knowledge of the associated properties is connected with a similar ability (cf. Jackson 1998b; Jackson 2004, 272). I will leave the question aside what exactly this ability amounts to.

possession of a concept enables a speaker to determine the concept's extension with respect to a hypothetical scenario. But there is still one more step to go: Chalmers and Jackson's thesis is clearly about worlds considered as actual, they speak about the speaker being provided hypothetical information about the actual world. For the descriptivist claim that speakers have implicit knowledge of the associated properties to have any chance of being correct, one would have to make the same restriction. This is because as we have seen, to determine the extension of an expression with respect to a world considered as counterfactual, one often requires empirical information which need not be accessible to the speaker. One could define two kinds of (associated) properties here: A primary (or primarily associated) property is a set of extensions across worlds considered as actual; a secondary (or secondarily associated) property is a set of extensions across worlds considered as counterfactual. Given metaphysical plenitude any primary property will be identical with some secondary property. Thus, these definitions should not be taken to mark a genuine metaphysical distinction, but rather as being purely epistemically and semantically motivated.

We can now formulate two theses concerning knowledge of associated properties, which should be considered as true by definition:

(KAP1) If a speaker has knowledge of the primary property associated with an expression, then she can determine its primary extension with respect to every possible world (or alternatively, she can determine its extension with respect to every world considered as actual), given a canonical description of the world and ideal rational reflection.

(KAP2) If a speaker has knowledge of the secondary property associated with an expression, then she can determine its secondary extension with respect to every possible world (she can determine its extension with respect to every possible world considered as counterfactual), given a canonical description of the world and ideal rational reflection.

Since speakers are often unable to determine a term's extension with respect to a world considered as counterfactual and thus do not know the secondary property associated with an expression, the descriptivist's thesis should be that a competent speaker knows the associated primary properties of the terms she uses. In fact, in *Why We Need A-Intensions*, Jackson himself invokes the notion of associated properties in connection with A-intensions, i.e. his version of primary intensions (cf. Jackson 2004, 264). What does this imply for Jackson's thesis that reference is determined by speaker associations? First of all, as a thesis about reference with respect to worlds considered as actual, it is already implied by two-dimensionalism. The same is true for reference in our world, since in the two-dimensionalist framework, the primary and the secondary extension of an expression with respect to the actual world are always identical. The thesis that speaker associations determine reference with respect to worlds considered as counterfactual seems clearly false, however, given the existence of necessary a posteriori truths. But one could hold a slightly weaker thesis, namely that speaker associations determine the reference of an expression with respect to a world considered as counterfactual in a context, or given a specified actual world with a marked center. Once again, this thesis is straightforwardly implied by two-dimensionalism.

To sum up: On its most plausible reading, Jackson's claim that competent speakers know the property associated with a given linguistic expression is equivalent to the thesis that they grasp its primary intension. His account thus offers a natural way to understand mastery of a term or possession of a concept, which involves the ability to determine the term's or concept's extension with respect to a given hypothetical scenario. On this interpretation, Jackson's variety of descriptivism should not be taken to exceed two-dimensionalism. Rather, it provides another way of framing some familiar two-dimensionalist theses about meaning, reference, and concept-possession.

2.1.5 Two notions of apriority

According to two-dimensionalism, a sentence is a priori if and only if it has a necessary primary intension, i.e. if and only if it is true with respect to every epistemic possibility. On the face of it, connecting apriority with epistemic necessity seems plausible: If a sentence is epistemically necessary, then it is in principle possible to determine its truth by going through all epistemic possibilities in one's mind. If, however, a sentence is not epistemically necessary, then its truth is dependent on contingent features of the world. It seems reasonable to think that it can thus only be known empirically. However, that latter claim is at least debatable. Let me illustrate this by means of a hypothetical case:
Suppose that a piece of chocolate which I put into my safe just a few hours ago is gone. The safe shows no signs of a violent break-in. I know that my room-mate, who likes chocolate a lot, is the only person who was in the room at that time. She is also the only person aside from me who knows the combination of the safe's lock. When I ask her, she tells me outright that she opened the safe and ate the chocolate. I thereupon come to believe that my room-mate took my chocolate. It is very plausible that (absent defeaters) this belief amounts to knowledge. Now take all the evidence I have in that hypothetical case, all my knowledge relevant to the judgment and, if necessary, also other relevant empirical background facts to build the antecedent of a conditional: 'If there is no chocolate in safe A at time T and there was chocolate in the safe two hours before T and the safe was subsequently locked and … and the laws of nature are such-and such …'. The consequent of the conditional is my judgment about the fate of the chocolate, i.e., that it was taken by my room-mate. If a subject is presented with that conditional, thinks carefully about it and concludes that it is correct, then her justification for the resulting belief is at least as strong as mine for my belief that my room-mate took the chocolate. But that subject's justification, unlike mine, is not based on empirical evidence: All the information relevant for justifying the conclusion is already in the antecedent of the conditional. Accordingly, it is natural to say that the

subject knows the conditional a priori.[25] At the same time, it is plausible that the conditional is not epistemically necessary. For example, the prevailing laws of nature may make it extremely unlikely that the chocolate suddenly dematerialized, but they do not exclude this possibility. Consequently, the conditional in question is knowable a priori, even though it is not epistemically necessary.

This result seems to raise a problem for two-dimensionalism which postulates an intimate connection between apriority and epistemic necessity, as seen above. An obvious way to solve that problem is to require stronger justification conditions for a priori knowledge. This would mean that the conditional I mentioned above would not be knowable a priori. Chalmers does indeed construe apriority as requiring 'conclusive' justification (cf. e.g. Chalmers 2006, 98). I do not like this idea, however. It is hard to see a reason to require stricter justification conditions for a priori knowledge (cf. also Casullo 2003, 33ff.; BonJour 1998, 110ff.). Moreover, since such an understanding makes it unnecessarily hard to acquire a priori knowledge, it could be detrimental to the project of defending the epistemic fruitfulness of conceptual analysis.

Let me therefore propose to distinguish two notions of apriority: The first one is equivalent to epistemic necessity, i.e. to having a necessary primary intension. One might call this notion strong apriority.[26] On the second notion, which one might thus call 'weak apriority' or 'standard apriority', a sentence (thought/proposition) is a priori if it can be known on the basis of justification which is independent of sense experience, where the relevant justification conditions are just as strong as those we impose on empirical knowledge. It is plausible, barring skepticism about for instance complex mathematical or logical truths, that if a sentence is strongly a priori, it is also standardly a priori, though the reverse does not generally hold. Both of these notions of apriority can be useful, depending on one's purposes. In

[25] We will encounter a similar move in an argument for a priori scrutability in 4.1.

[26] This notion is related, but not equivalent, to Hartry Field's 'strong aprioricity' which is tied to indefeasibility by empirical evidence (cf. Field 1996).

the following, I will briefly discuss which roles they could play within the two-dimensionalist framework.

Above I explained that according to two-dimensionalism, the grasp of an expression's primary intension bestows a subject with conditional a priori knowledge, i.e. the ability to determine the extension of the expression with respect to every world considered as actual a priori. Now it might be that the grasp of a term often, or even usually, enables a subject to determine the expression's extension conclusively, because the description of the scenario necessitates the extension. But I do not want to consider this as a requirement. I therefore hold that for the two-dimensionalist claim to be correct, it suffices that we are able to determine the extension of an expression with respect to a scenario in a standardly a priori way. Notice that this entails that conclusive justifiability and strong a priority can come apart, since it is possible that a sentence is epistemically necessary without being conclusively justifiable. For even if the truth of a sentence with respect to every scenario can be determined in a standardly a priori way, the justification we can have for at least some of these judgments may still be inconclusive. Let me therefore introduce the notion of super-strong apriority: A sentence is super-strongly a priori if and only if it is knowable with conclusive a priori justification.[27] It is plausible that just like strong apriority entails standard apriority, super-strong apriority entails strong apriority.

I just argued that two-dimensionalism should only require that the extension of an expression with respect to worlds considered as actual is determinable in a standardly a priori way. However, for other purposes, the notion of strong apriority is plausibly more relevant. It seems reasonable, for example, that for a feature F to be an a priori associated property of 'A' in Jackson's sense, F must be correlated with A with epistemic necessity, i.e. 'A → F' must be strongly a priori.

[27] This is the notion which may be equivalent to Field's notion of strong aprioricity (cf. Field 1996), cf. the preceding footnote.

For the majority of this book, the distinction between standard and strong apriority will not be important. I will mention it whenever it does become relevant.

So far, we have seen that two-dimensionalism promises to reconcile Kripke's and Putnam's insights with a broadly Fregean picture of meaning. But in itself, the existence of the two-dimensional framework does not suffice to establish such a picture, much less does it show that conceptual analysis is or can be a valuable philosophical method. I will address the questions of what exactly the two-dimensional framework shows and of what is still to be shown in 2.3. However, I presented it as a key motivation for adopting this framework that it can account for the Kripkean a posteriori necessities. So I think this is one thing which two-dimensionalism really needs to accomplish in order for the whole project to offer any prospects of success: It must be able to explain these a posteriori necessities in a way which leaves room for an a priori dimension of meaning and thereby establishes a connection between epistemic and metaphysical modality. I will therefore now turn to the question of whether a two-dimensionalist account manages to accomplish this task. It may seem curious that I will take Kripke's theory of modal error as the starting point for addressing this question. However, this approach draws its justification from the fact that the two-dimensionalist account can not only be understood as a reaction to Kripke's arguments, but also as a derivative of his modal epistemology. My general aims will be these: Firstly, I will carve out the similarities as well as the differences between the Kripkean and the two-dimensionalist explanation of a posteriori necessities. Secondly, I will point out what Kripke's account is and what it is not committed to, as compared with two-dimensionalism. And thirdly, at those points where Kripke's account faces objections, I will examine how two-dimensionalism can deal with them.

2.2 Modal illusions according to Kripke and according to two-dimensionalism

Due to the connection between meaning and modality, any account which tries to defend the tenability of conceptual analysis as an a priori method has to establish some kind of a priori access to what is possible and what is necessary. I outlined above how an intimate relation between modality and the a priori can be posited within a two-dimensional framework. In this respect, two-dimensionalism seems to be very much at odds with Putnam's and Kripke's views. Indeed, the lesson which Putnam drew from the existence of necessary a posteriori truths was to reject any kind of a priori access to modality. Kripke has often been assumed to take a very similar stance, not least by Putnam himself. But actually, as was briefly sketched above, many aspects of the two-dimensionalist explanation of the necessary a posteriori can already be found in Kripke's *Identity and Necessity* (1971) and in *Naming and Necessity* (1980).[28] In the following, I will outline Kripke's account of so-called 'modal illusions' in more detail and highlight its similarities with two-dimensionalism. I will then discuss some objections which have been raised and some which could be raised against Kripke's view. Insofar as these objections prove critical, I will examine how two-dimensionalism can fare against them.

Take a sentence like 'Hesperus = Phosphorus'. Since proper names are rigid designators, the sentence expresses a necessary truth. But still, it appears contingent. It at least seems conceivable that Hesperus is not identical with Phosphorus. Likewise, it seems as though it could have turned out that Hesperus is not Phosphorus. After all, for a long period of time in human history people believed that Hesperus and Phosphorus are not the same celestial body. Unlike many other semantic externalists, Kripke does not rest content with merely calling attention to the existence of such illusions of contingency. He tries to give an explanation for them. The first thing he points out is that sentences like 'Hesperus ≠ Phosphorus',

[28] In a later paper, Putnam admitted that he had earlier taken Kripke's position to be closer to his own than it actually is (cf. Putnam 1990).

though being metaphysically impossible, are nevertheless epistemically possible. This mere fact may account for the appearance of possibility. It is also an important part of the motivation for endorsing two-dimensionalism – after all, primary intensions are built precisely on the notion of epistemic possibility. However, saying that it is epistemically possible that Hesperus is not Phosphorus can just be taken as another way of stating that it is not (strongly) a priori that Hesperus is Phosphorus. As was pointed out above, this is of not much help to the proponent of conceptual analysis unless there is some kind of connection between epistemic and metaphysical modality.

Kripke's next step is to ask what it could mean to say that although it could have turned out that Hesperus is not Phosphorus, it could not have been the case. If it turns out or had turned out that Hesperus is not Phosphorus, would it not be / have been the case that Hesperus is not Phosphorus? Kripke accepts this line of reasoning and states that for every (metaphysical) impossibility I, it cannot turn out that I (cf. Kripke 1980, 141). He argues that when we say that it could have turned out that Hesperus is not Phosphorus, we are just speaking loosely. Something similar can be said about conceivability: When it seems conceivable that Hesperus is not Phosphorus, this is mere seeming – we do not really conceive of a situation where Hesperus is not identical with Phosphorus.

Kripke's way of putting things is disputable. Chalmers, for example, uses 'could turn out' purely epistemically throughout his writings. The same is true for Kripke's use of 'conceiving': Most writers have taken conceiving (and related notions) to be tied to apriority by definition (cf. e.g. Yablo 1993; Tidman 1994; Menzies 1998). However, these are just terminological issues. The important point is that according to Kripke, when a necessary falsehood appears to be conceivable, what we have really conceived is something else. And when we say, for some necessary falsehood I, that it could turn out that I, what could have really been the case is some I*. So the natural question to ask is: What is this I*, i.e. what is the genuine possibility which we really conceive?

2.2.1 Kripke's two models of modal error

Kripke offers two different models for explaining (or explaining away) these modal illusions. Yablo calls them the 'epistemic counterpart model' and the 'reference fixer model' (cf. Yablo 2006). Kripke illustrates the first of these models by means of his example of the wooden table (cf. Kripke 1980, 142): We may intuitively believe that this table could, contrary to appearances, turn out to be made of ice. But since facts about the origin and the constitution of a material object are necessary, this is impossible. That is, if the table is made of wood, it is essentially made of wood and it could thus not turn out / have turned out to be made of something else. However, it could have really been the case that there was

> *a table* looking and feeling just like this one and placed in this very position in the room, which was in fact made of ice. In other words, I (or some conscious being) could have been *qualitatively in the same epistemic situation* that in fact obtains, I could have the same sensory evidence that I in fact have, about *a table* which was made of ice. (Kripke 1980, 142)

That is to say, although there is a possible state of affairs involved, it is not about this particular table but rather about an 'epistemic counterpart' of it, that is a table which looks and feels just like it. It is certainly possible for a table made of ice to be an epistemic counterpart of a given wooden table. Moreover, there is at least some initial plausibility to the view that what is really conceived is just some table – after all, we cannot even distinguish that particular table from an epistemic counterpart. Thus, Kripke holds that when we seem to conceive of an impossible situation where this very table is made of ice, what we conceive is something possible: Namely that some table which looks like this wooden one is made of ice (cf. Kripke 1971, 160f. – though he talks about a lectern there).

Kripke introduces his second model by drawing on the already familiar example of the identity statement 'Hesperus = Phosphorus'. In his critique of descriptivism, he noted that although an associated description cannot give the meaning of a proper name, it may still be used to fix its reference. Take for example 'the brightest object in the evening sky'. The reason why this definite description cannot give the meaning of 'Hesperus' is that the

expressions are not modally equivalent: As was pointed out above, names refer rigidly, unlike the associated descriptions. But this does not mean that the description cannot fix the reference in the actual world, given that Hesperus really is the brightest object in the evening sky. Now if we replace the rigid designators in the metaphysically impossible sentence 'Hesperus ≠ Phosphorus' by corresponding reference-fixing descriptions, we get a sentence like 'the brightest object in the evening sky ≠ the brightest object in the morning sky' which is still false, but contingently so. This observation provides the basis for the reference fixer model. For any identity statement involving two rigid designators '$R_1 = R_2$', the reference of R_1 and R_2 can also be fixed by the descriptions D_1 and D_2, yielding the contingent '$D_1 = D_2$' (cf. Kripke 1980, 143f.). It is our confusing the former with the latter sentence which produces the illusion that 'Hesperus = Phosphorus' and the like are contingent. It should be noted that to create a contingent statement, it suffices to replace one of the rigid designators by a reference-fixing description – provided that the property by which the reference is fixed is not an essential property of the referent.

One can see some obvious similarities between Kripke's models and the two-dimensionalist treatment of necessary a posteriori truth. Both views concede that there are epistemically possible sentences which are metaphysically impossible, but argue that in each such case, there is a genuine metaphysical possibility nearby. Kripke's remarks thus suggest that he is committed to a principle quite similar to the thesis of metaphysical plenitude – the thesis that for every epistemic possibility there is a corresponding (centered) metaphysically possible world. And just like in the two-dimensionalist account, Kripke's explanation can be taken to imply that the seeming impossibilities are simply misdescribed: In the case of the wooden table, we mistake a situation where an epistemic counterpart of the table is made of ice for one where the table itself is. In the case of Hesperus and Phosphorus, the diagnosed misdescription is evident. We confuse the metaphysical possibility 'the brightest object in the evening sky ≠ the brightest object in the morning sky' with the impossibility 'Hesperus ≠ Phosphorus'.

2.2.2 Doubts about the accounts of modal error

2.2.2.1 Doubts about the epistemic counterpart model

Kripke's explanation of the modal illusions has been disputed. Stephen Yablo, for instance, thinks that although Kripke's models can account for some of the cases which he discusses, they cannot account for all of them (cf. Yablo 2006). The biggest part of Yablo's criticism is directed against the epistemic counterpart model. He begins by pointing out that being an epistemic counterpart of the wooden table can mean either of two things. It can mean that there is a counterfactual situation in which a table looks and feels to some individual just like the wooden table actually looks and feels to me. Or it can mean that there is a counterfactual table which looks and feels the same *to us* or *to me* as the wooden table. The first kind of scenario is certainly possible. If the individual in the counterfactual scenario has a neural (or whatever kind of) architecture which is sufficiently different from ours, any kind of table or non-table can look and feel to her like the wooden table does to us. But as Yablo remarks, this hardly explains the modal illusion. It is not very plausible that what we really imagine when we think we imagine this (wooden) table to be made of ice is, say, an 'ordinary' table made of ice which due to a counterfactual observer's extraordinary neural architecture looks to her like the wooden table looks to us. The proper reading of the epistemic counterpart model should thus be that there has to be a counterfactual table which would be an epistemic counterpart of the wooden table for us, i.e. including our actual neural architecture. Still, the case seems unproblematic on this reading, too. It is plausibly possible that a cleverly prepared table made of ice looks and feels just like a wooden one. However, Yablo tries to argue that the epistemic counterpart model fails with respect to another one of Kripke's examples, or at least a specific version of it.

An important set of a posteriori necessities which Kripke discusses are theoretical identifications, like 'heat = mean molecular kinetic energy'. Here, the epistemic counterpart model should be able to explain the intuition that heat could have turned out to be something else and point up

the possibility which is really conceived. Yablo grants that there could have possibly been an epistemic counterpart of heat, i.e. some phenomenon other than mean molecular kinetic energy which causes the same sensations in us. But let us consider a slightly more specific scenario. It seems likewise that heat (here understood as the opposite of cold) could have turned out to be low mean molecular kinetic energy, instead of high mean molecular kinetic energy (here, 'high' and 'low' are highly relative, of course). But the seeming possibility of 'heat = low mean molecular kinetic energy' cannot be accounted for by the epistemic counterpart model, according to Yablo. Kripke does note that it is contingent that mean molecular kinetic energy is felt as heat: Had our neural architecture been different, mean molecular kinetic energy could have caused very different sensations in us (cf. Kripke 1980, 133). This seems irrelevant, however. After what has been shown above, the required kind of epistemic counterpart of mean molecular kinetic energy would have to be felt as heat for us, with the neural architecture we have. But Yablo argues that it is impossible for mean molecular kinetic energy to be felt differently than it is felt, if we keep our actual neural architecture fixed. It is thus also impossible for low mean molecular kinetic energy to be felt as heat. If so, Kripke's model fails: There is no genuinely possible state of affairs to be found which is really conceived. It seems like the epistemic possibility of heat turning out to be low mean molecular kinetic energy corresponds to no metaphysical possibility.

If Yablo's argument succeeds, this is bad news not only for Kripke's view, but possibly also for the thesis of metaphysical plenitude, which is supposed to play a key role in the defense of conceptual analysis. In the following, I will therefore discuss in some detail how Yablo's challenge can be met. I will first identify two principled ways in which his 'fool's cold' case could be rejected and point out how Yablo would have to respond to these objections. Against this background, it will be possible to show straightforwardly that Yablo's critique does not threaten the two-dimensionalist account of modal epistemology. After that, I will argue that Yablo's critique does not provide a refutation of Kripke's model, either.

For a start, one could wonder whether 'fool's heat' – low mean molecular kinetic energy which is felt as heat – is not possible after all. The conceivability of fool's heat could even be used as a premise in a kind of 'inverted spectra' argument against physicalism. The acceptance of such a premise would hardly be a problem for either Kripke or Chalmers, both of whom in fact put forth quite similar arguments against physicalism. Yablo could respond by insisting that not only the actual physical basis of our neural architecture has to be held fixed in the counterfactual scenario, but also, say, the possibly partly non-physical laws which govern our brain processes. Thus understood, fool's heat is surely impossible. It is not clear that this is the most plausible way to interpret the fool's heat scenario. But I grant that it can be understood this way. Moreover, on this version fool's heat still seems conceivable. So Yablo's argument still poses a challenge.

Another way to attack the argument is to deny that fool's heat is conceivable in the relevant sense. Recall the student in class who is supposed to calculate 2^7. Before she switches on her pocket-calculator, in one sense the correct answer could turn out to be 256 – unless she has thought sufficiently hard about it. But given the idealized notion of epistemic possibility, '$2^7 = 256$' is not really epistemically possible, i.e. it cannot really turn out that way. Chalmers explicitly argues that only the idealized sense of epistemic possibility is of relevance here. This thought is also present in Kripke's writings: Kripke says that the modal illusion that Hesperus could have turned out to be distinct from Phosphorus goes deeper than the (putative) illusion that the four color theorem could turn out to be false (cf. Kripke 1980, 103).[29] So maybe 'heat = low mean molecular kinetic energy' only seems to be epistemically possible and thus is not conceivable given ideal rational reasoning.

There is a second way in which non-ideal epistemic conditions could be responsible for the appearance that fool's heat is possible: We simply do not know enough about our neural architecture to see that it is not

[29] By the time Kripke *Naming and Necessity* was published, the theorem had not yet been proven, so he did not know if it would turn out to be true.

compatible with low mean molecular kinetic energy being felt as heat. We certainly do not know anything about non-physical laws of nature which are involved in the emergence of phenomenal states, if there are such laws. One could thus even argue that we do not really understand the proposition we are supposed to conceive. Let me therefore try to spell out its exact content. The relevant hypothetical situation is one where our actual neural architecture is to remain fixed and in which low mean molecular kinetic energy is felt as heat. Since we do not know what our actual neural architecture is like, we cannot specify this aspect of the situation descriptively. The content of the proposition thus has to be just this: 'Our neural architecture is as it actually is and low mean molecular kinetic energy is felt as heat.' Presumably, this is unproblematic on Yablo's view. He remarks that modal judgments involving explicit reference to the actual world are quite common. One example he gives is 'There could have been less ivory-billed woodpeckers than there actually are' (cf. Yablo 2006, 332). Of course, this judgment is based on the belief that the ivory-billed woodpecker has not yet gone extinct. If Yablo is wrong about this, then so is his modal belief.

The considerations just made will help to answer the question of whether the alleged conceivability of fool's heat is capable of undermining the two-dimensionalist explanation of modal error: The first thing to note is that on Yablo's account, modal judgments can explicitly rely on empirical information. This is most obvious in his wood-pecker example. Accordingly, conceivability thus construed cannot be an a priori source of knowledge. Secondly, and relatedly, recall how hypothetical scenarios are evaluated in two-dimensionalism. The scenarios are given via a complete canonical description in a vocabulary which must not include terms whose evaluation is implicitly dependent on characteristics of the actual world. On this account, any description of a scenario containing the term 'actual' is obviously not suitable. This suggests that the modal intuition that fool's heat is possible, if it is spelled out the way it is by Yablo is irrelevant for the two-dimensionalist account. In a canonical description, the term

'actual' would have to be replaced by a complete specification of the subject's neural architecture.

One might object that this response is too simple. Given that 'actually'-involving modal judgments are quite common, as Yablo argues, is it not just ad hoc to exclude this kind of judgment? After all, they seem to be a source of modal error and if Kripke's account and two-dimensionalism do not deal with this type of modal error, so much the worse for them. However, I think this objection is ill-founded. Consider for instance the illusion that water could have failed to be H_2O. As was elaborated above, two-dimensionalism can be taken to explain this illusion, or respectively the fact that 'water is H_2O' is both metaphysically necessary and epistemically contingent, by pointing out that the term 'water' makes an implicit reference to the actual world. The solution to this problem has two parts. The first part involves the thesis that in each of these cases, there is a genuine possibility which corresponds to the impossibility. This possibility is revealed by focusing on the primary intension of the relevant expression. The second part of the solution is to prevent any possible misdescription of the conceived possibility by avoiding the use of terms whose secondary intensions differ from their primary intensions. Thus, the implicit dependence of certain terms on the character of the actual world is identified as the source of modal error, and the exclusion of those terms in a canonical description is an important part of the strategy to avoid such error. Accordingly, if one encounters a modal illusion and then tries to replace the (supposedly) misleading description of the conceived situation, it would not make much sense to replace it by one which makes *explicit* reference to the actual world.

Of course, this still leaves open the question what is really conceived when we deem fool's heat to be possible. It is certainly not what might be suggested by what I said about the requirements of a canonical description. That is, it is not what we get by replacing 'actual' in 'our neural architecture is as it actually is and low mean molecular kinetic energy is felt as heat' by a complete specification of that architecture. For firstly, such a situation would presumably no longer be conceivable given ideal

rational reflection. Secondly, it is not what we do or even something we can imagine, given our cognitive limitations and our empirical ignorance. What we actually imagine, I think, is something very unspecific. The details of course depend on how sketchy the conceiver's knowledge of the human brain is. In my case, it would probably be something like 'It is possible that some person has a brain weighting a bit more than one kilogram, consisting inter alia of a couple of billion neurons and even more glial cells, and experiences low mean molecular kinetic energy as heat.[30] This modal judgment is quite obviously correct, and I actually think this is sufficient to explain away the seeming possibility of fool's heat.

I have just shown that the modal illusion that low mean molecular kinetic energy could have been felt as heat does not threaten the thesis of metaphysical plenitude and the two-dimensionalist explanation of the necessary a posteriori in general. It is still worth checking whether it nevertheless threatens Kripke's epistemic counterpart model. There are at least two ways for Kripke to reject Yablo's conclusion. One is to deny that fool's heat is conceivable in the relevant sense. That is to say, once the relevant scenario is spelled out in sufficient detail, any appearance of possibility would be just due to the subject's cognitive limitations. The modal illusion that fool's heat is possible would then be no different in kind from the illusion that the four color theorem could have turned out to be false. Another way to deny the conclusion is to reject Yablo's interpretation of the fool's heat scenario. It is at least not completely clear that what is really required for there to be an epistemic counterpart of heat is that there is a phenomenon which is felt as heat for us as we actually are, up to every detail in our neural architecture. This need not imply that the possibility of a creature with a completely alien perceptual system experiencing low mean molecular kinetic energy as heat is sufficient to establish the possibility of fool's heat. One could also take an intermediary position and say for instance that what is required is that low mean molecular kinetic

[30] This description is still not very accurate since I do not know exactly what neurons, glial cells or molecular kinetic energy are.

energy is possibly felt as heat given all we currently know about our neural architecture.[31] It is not clear which of these lines Kripke should take, and of course it's even less clear which of them he would take. But in any case it is safe to say that the example of fool's heat does not provide an outright refutation of the epistemic counterpart model. However, I think there is another lesson to be learned here, which is not directly related to this particular example. If an epistemic counterpart is understood the way Yablo understands it, then in addition to the relevant object or property, there would also have to be a counterpart of the conceiver present in the scenario. Taken as a general requirement, this seems problematic for two reasons: Firstly, it makes the epistemic counterpart model's value as a heuristic device problematic. It is too demanding to require a subject to imagine a duplicate of herself, or to judge how she would experience some object or property given her neural architecture. And secondly, it has to be possible to talk about hypothetical scenarios where no observer is present. It seems sensible to ask for example if the wooden table would still look the same if all life in the universe had been extinguished. So the fools' cold scenario might be interpreted as a special case such that a counterpart of the subject has to be present in the scenario conceived. But it seems wrong to take the epistemic counterpart model to posit not only that the genuinely possible (and genuinely conceived) scenario has to contain a counterpart of what is conceived, but also a counterpart of the conceiver. In general, an epistemic counterpart of an object or property should thus be understood in a weaker sense: An epistemic counterpart of for example a table is an object which looks, feels, smells, … like the original table, without further qualification.

[31] The latter proposal would be in line with an understanding of a qualitatively identical situation as one in which all the evidence we have, sensory or not, is the same.

2.2.2.2 Doubts about the reference fixer model

Now let me turn to Kripke's reference fixer model. In many cases, this strategy for explaining away modal illusions seems to work just fine. Just replace the rigid designators flanking the identity sign in a sentence like 'Hesperus = Phosphorus' by contingent reference-fixing descriptions and you get a contingent sentence. The same seems to apply to 'heat = mean molecular kinetic energy'. If one replaces 'heat' by something like 'the phenomenon which causes heat sensations', the resulting sentence is plausibly contingent. However, if we understand the latter case the way Yablo proposes, this immediately raises some problems. As was just discussed, Yablo holds that many of our modal judgments involve reference to the actual world. I do not want to deal with the question of how specific examples should be interpreted. I will grant that there are modal judgments which refer to features of the actual world and just deal with the general question of whether and how such cases can be accounted for by the reference fixer model.

So, to stick with the current example, what if the question we are concerned with is not just whether molecular kinetic energy could have failed to cause heat sensations, but whether it could have failed to cause such sensations in us as we actually are? One possible solution proposed by Yablo himself amounts to 'rigidifying' the relevant description, in our case yielding 'the actual cause of heat sensations'. But in his view, this would not work either (cf. Yablo 2006, 338). He thinks that the reference fixing description would have to be a 'piece of language'. However, a token of 'the actual cause of heat sensation' uttered in some counterfactual scenario does not refer to the actual world, but to the counterfactual scenario itself, so this does not help. And even if there was a description which referred to the actual world no matter which world it was uttered in, then such a description would not be understandable in a counterfactual scenario.

I do not think this objection is decisive. The main thing it teaches us is that reference-fixing descriptions should just be understood as (linguistic) types associated with expressions by speakers. One should thus not require that a

token of them has to be present in a counterfactual scenario. After all, rigid designators can also be evaluated with respect to possible worlds which do not contain any token of the expression – as in 'Aristotle could have never existed'. There is no reason why Kripke should be committed to the existence of a token of the relevant description in the counterfactual scenario.
However, it is easy to see that invoking rigidified reference-fixing descriptions does not help anyway, since the expression which is supposed to replace the original one is not contingent, either. Thus, for example 'the phenomenon which actually causes heat sensations ≠ mean molecular kinetic energy' is necessarily false and does not help to explain away the modal illusion. It is thus crucial to take care that the reference-fixing description which replaces the rigid designator is really contingent. Accordingly, the solution to the problems raised by Yablo's examples does not lie in a modification of the reference fixers. Rather, one should proceed like in the case of the epistemic counterpart model discussed above: Any explicit reference to features of the actual world has to be obliterated and replaced by a description which explicitly specifies these characteristics. There is no apparent reason why this solution should not be available here as well.

The reference fixer model can only be successful if the statement which replaces the original one plausibly represents a situation which is really imagined by the subject. I.e., it does not suffice that for a given rigid designator there exists a reference-fixing description which turns a false possibility statement like 'heat could have failed to be mean molecular kinetic energy' into a true one like 'the phenomenon which causes heat sensations could have failed to be mean molecular kinetic energy'. The reference-fixing description also has to be accessible to the subject and it has to be strongly associated with the term in question. With this in mind, the reference fixer model seems even closer in spirit to two-dimensionalism than the epistemic counterpart model. Where two-dimensionalism holds that our evaluation of an expression with respect to a possible world is

guided by our grasp of a primary intension, Kripke's model seems to imply that we evaluate these worlds by relying on reference-fixing descriptions which we associate with the relevant expressions. The extent of these similarities is quite surprising, given the general picture of meaning Kripke is otherwise committed to. Precisely this serves as the starting point of the critique of George Bealer (cf. Bealer 2006) and Christian Nimtz (cf. Nimtz 2007). They argue that the reference fixer model is not compatible with Kripke's anti-descriptivism. And indeed, this model seems to commit Kripke to holding that speakers have access to descriptions which fix the reference of the terms they use.[32] But this is at odds with Kripke's own so-called arguments from Ignorance and Error: Kripke himself pointed out that speakers frequently do not know anything which could determine the reference of a given term. Nevertheless, they do refer when they use it. Take Alvin who uses the name 'Gell-Mann'. The only thing he believes about Gell-Mann may be that he is a famous physicist. He therefore does not know anything to distinguish Gell-Mann from any other famous physicist like, say, Richard Feynman. Or take Batu who does have more specific beliefs about Gell-Mann. She believes that Gell-Mann is the famous physicist who developed the theory of quantum electrodynamics. But as it happens, she is wrong: It was Feynman who developed quantum electrodynamics. Nothing that Batu associates with the name can thus determine its reference, either. For her belief is still a false belief about Gell-Mann, not a true one about Feynman. Thus, when she utters the name 'Gell-Mann', she refers to Gell-Mann, just like Alvin. If this is correct, then it seems that speakers need to know nothing which determines the reference of the terms they use. But then there is no basis for the reference fixer model. In cases of Ignorance or Error, there is no suitable reference-fixing description to be found which can replace the relevant rigid designator, in order to explain away the modal illusion.

[32] This does not seem to be Bealer's line of argument, however. In fact, I think that his critique of Kripke at this point relies on a failure to distinguish between associated descriptions which give the meaning and those which merely fix the reference of an expression.

There are two possible conclusions which could be drawn from this. One could either hold that since Kripke himself is committed to saying that speakers have access to reference-fixing descriptions, maybe there is some room for a moderate version of descriptivism after all. This is in effect Nimtz' proposal. Or one could conclude that Kripke's reference fixer model is flawed, since it is based on an idea which has been refuted by Kripke himself.

In fact, I think both of these conclusions are premature. Kripke does not claim that the reference fixer model can be applied to all cases of modal error. So he could just say that the model applies in those cases where speakers do have access to an appropriate reference- fixing description. But things are not quite as simple for Kripke who is committed to a strong thesis concerning the scope of his models. For in his argument against materialism, he argues that since the seeming possibility of pain without C-fiber activity cannot be explained away by one of his models, there could have been pain without C-fiber activity (cf. Kripke 1980, 144ff.). Thus, for his argument to work, it must be possible to explain away all kinds of seeming modal error. We saw above that it is not altogether clear whether the epistemic counterpart model can be applied to all kinds of cases. Now if Kripke's second model is only of limited range as well, the ambition to explain away all kinds of modal illusions may not be satisfiable. Fortunately, as Alma Barner points out, the range of the reference fixer model can be greatly extended with a small modification (cf. Barner ms.). One should not require that a rigid designator is replaced by a description which suffices to determine the term's reference. It suffices if it is replaced by some description which represents what the speaker actually does associate with the term. Take for example the modal illusion that Richard Feynman could have been Murray Gell-Mann's brother. In the case of Alvin, we get 'Some famous physicist could have been the brother of some (other) famous physicist'; in the case of Batu it could be 'The author of 'Ivanhoe' could have been the brother of the inventor of quantum electrodynamics'. Both of these sentences clearly express possibilities. But do they really explain the modal illusion? It might be argued that especially

in the first case, the allegedly imagined possibility is much too unspecific and the truth of the modal judgment is thus trivial. I actually think this rather speaks in favor of the current proposal, though. If what a speaker associates with a term is very unspecific, it would be odd if the resulting scenario was much more detailed. For the model to be psychologically adequate, the replacing description has to reflect the subject's state of knowledge, or ignorance.[33]

It is plausible that the modified 'reference fixer model', which now has to be considered misnamed is applicable to all kinds of cases of modal error.[34] This requires only that a speaker associates something with a given term. These associations do not even have to be semantic. They must only be sufficiently strongly connected with the term to make it plausible that they represent what the subject has in mind when she conceives of a corresponding situation. Thus understood, the model is compatible with Kripke's other theoretical commitments. At the same time, this clearly distinguishes his view from two-dimensionalism which posits that what a speaker associates with a term enables her to determine the term's reference with respect to every possible world considered as actual. Since two-dimensionalism cannot appeal to Kripke's model of modal error in this respect, it has to deal with the arguments from Ignorance and Error. I will discuss the question of whether the mastery of a term, or the possession of a concept, really amounts to grasping its primary intension in detail in chapter 3.

[33] It has been objected that in some cases, Kripke's model is inadequate because it does not provide a *de re* possibility, for instance by Janine Jones (cf. Jones 2004). A complaint to that effect could be raised here as well – it is not clear which philosophers the judgment of Alvin is about. However, I do not see why his judgment has to be about particular philosophers. I would again consider the fact that the description is just as unspecific as the speaker's ideas of Gell-Mann and Feynman as a virtue.

[34] Whether it can be used to explain away all kinds of modal error is a separate issue, though (cf. also 2.3).

2.3 Summary and outlook: What has been shown and what is yet to be shown

I think it has become clear that two-dimensionalism offers some very useful tools for a defense of conceptual analysis. But of course, the existence of the two-dimensional framework as such does not show that conceptual analysis can play a substantial role in philosophical inquiry. In the remaining part of this chapter, I will sum up what has been shown so far, and examine what is yet to be shown. This will also set the agenda for the following chapters. Some of the theses connected with two-dimensionalism go beyond what is required to offer conceptual analysis a theoretical foundation. I will thus also identify those parts of the two-dimensionalist account which seem dispensable with respect to my aims in this book.

A very important virtue of two-dimensionalism is that it offers a way to account for the Kripkean a posteriori necessities which still leaves room for an a priori dimension of meaning in the tradition of Frege. In the foregoing section, it transpired that two-dimensionalism does indeed offer a convincing explanation for the necessary a posteriori and the accompanying modal illusions, provided that there really are primary intensions associated with the relevant terms. However, the existence of primary intensions has not yet been established. Moreover, we saw that although Kripke's own theory of modal illusions is compatible with the existence of such a semantic value, it by no means presupposes it. The existence of primary intensions is obviously crucial for a defense of conceptual analysis. In the following two chapters, I will therefore try to dismantle a number of arguments against primary intensions and provide some positive reasons for positing them.

There is another question related to the modal illusions which is still open. Even if two-dimensionalism is able to explain the typical Kripkean a posteriori necessities, this does not imply that it can explain all of them.

That is to say, maybe metaphysical plenitude is not true after all – there could still be epistemic possibilities to which no metaphysical possibility corresponds. The question of whether there are such so-called 'strong necessities' has been most extensively discussed in the context of the debate on physicalism in the philosophy of mind. I am not planning to discuss these questions here in any detail – I will address the issue briefly in chapter 6, though. I think that the main motivation for denying that we have a priori access to modalities stems from Kripke's and Putnam's examples. Since examples of this kind are so common, they also pose a considerable threat to the prospects of conceptual analysis. Thus, if it is conceded that two-dimensionalism can explain away these typical cases, then any critique of conceptual analysis which invokes the necessary a posteriori no longer has much force. For even if there are local exceptions to metaphysical plenitude, this does not yet undermine the existence of a sufficiently reliable route from epistemic to metaphysical possibilities.

Another central component of two-dimensionalism is the scrutability of truth. The scrutability thesis could also play an important role in establishing the viability of conceptual analysis, which becomes apparent from the fact that, as was already pointed out in chapter 1, conceptual analysis is often done via thought experiments: From the description of a hypothetical scenario, we are supposed to judge whether the case described is a case of knowledge, or a good action, or a free choice, etc. I.e., we are supposed to determine the extension of a term with respect to a particular hypothetical scenario. The scrutability thesis can be taken to provide the theoretical foundation for our ability to judge such cases. Grasping a term's primary intension just means to be able to determine the term's extension when given hypothetical information about the world. Thus, two-dimensionalism provides a very straightforward rationale for the reliability of our judgments about hypothetical cases, basing it on our mastery of the relevant terms, or respectively on our possession of the relevant concepts. If successful, this would be an important achievement in itself, given the

ubiquity of the method of thought-experimentation in philosophical practice.

Against this background, it is quite remarkable that the most influential of Kripke's, Putnam's, and Tyler Burge's (cf. e.g. Burge 1979) arguments which are supposed to refute any Fregean account of meaning or content are themselves based on thought experiments – for example Kripke's 'Gödel and Schmidt', Putnam's 'Twin Earth', and Burge's 'arthritis' scenario. This might suggest another way to try and sustain the existence of primary intensions: One could argue that these arguments presuppose our ability to correctly evaluate hypothetical cases – and thus that they implicitly rely on the grasp of primary intensions. In fact, at a number of places Chalmers and Jackson seem to suggest precisely this (cf. e.g. Chalmers 2002b, 169; Jackson 1998b, 213). So do the arguments against semantic internalism really presuppose an internalist dimension of meaning? In my view, there is something to say in favor of this suspicion in the case of Putnam's 'Twin Earth'. Putnam gives a description of a hypothetical scenario with respect to which we are supposed to determine the extension of 'water'. One might think that our judgment that the liquid on Twin Earth is not water is based on our grasp of the primary intension of 'water' plus our empirical knowledge that the liquid in our lakes and seas is H_2O. However, it does not have to be understood this way. Firstly, the thought experiment does not really require us to determine the extension of water from a qualitative description of the scenario. We only have to infer that the liquid on Twin Earth is *not* water. To do this, it would suffice if for example we merely grasped some necessary condition for being water which XYZ fails to meet. And secondly, the inference in question – from the liquid on Earth being H_2O to the liquid on Twin Earth not being water – can still be taken to be based on empirical considerations, say, considerations from the history of science.

So what about Kripke's thought experiments? Take the afore-mentioned hypothetical case about Schmidt and Gödel: Kripke argues that it could turn out that it was not Gödel who discovered the incompleteness of arithmetic, but rather a man called 'Schmidt'. Gödel may have stolen it

from Schmidt and published it under his own name (cf. Kripke 1980, 83f.). For now, it need not concern us what the thought experiment is supposed to show.[35] The important questions here are how it works and what it presupposes. Chalmers argues in effect that Kripke's thought experiment conforms completely to the scrutability thesis (cf. Chalmers 2002b, 169): We are given a description of the hypothetical scenario and judge what the names 'Gödel' and 'Schmidt' refer to. Chalmers likens this to Gettier cases in the context of the analysis of knowledge. There, we are given a description of a scenario which we judge to be a case of justified true belief without knowledge. Thus, if Chalmers is correct, then far from providing a refutation of conceptual analysis, Kripke's thought experiment itself has to be considered as an instance of conceptual analysis. But again, things are not so simple. In fact, it would be odd if Kripke's thought experiment really worked as described by Chalmers. After all, Kripke does not tire to point out that possible worlds are not given to us qualitatively. In his view, we are not provided with a description of a world and then have to infer what is the case in that world. Rather, we just stipulate what is the case in a possible world (cf. Kripke 1980, 42ff.). On closer inspection, one can see that his 'Gödel/Schmidt' scenario is completely in line with this account: He does not provide a purely qualitative description of two persons from which we are supposed to judge who of them is Gödel and who is Schmidt. Rather, it is stipulated that it is Schmidt who discovered the incompleteness of arithmetic in the scenario, and Gödel who then stole and published it (cf. also Byrne & Pryor 2006, 50). So if these things are simply stipulated, then how do we know that things could really turn out to be as described in the scenario? I think the work in Kripke's argument is done by the following two intuitions: firstly, an intuition to the effect that discovering something is a contingent feature of a person; and secondly, the intuition that the names in question, or names in general, are used to refer to the same person with respect to every possible world. There is thus an important difference between Kripke's thought experiment and a typical Gettier scenario: In the latter case, it is not part of the description of the scenario that it is a case of

[35] I will return to this case in chapter 3.

non-knowledge. Therefore, it can be taken to rely on a conditional ability to determine the extension of our terms as expressed by the scrutability thesis, unlike Kripke's Gödel/Schmidt case. This does not mean that Kripke's account is in principle incompatible with the idea that we can infer the extension of a term with respect to a possible world from a qualitative description of that world. Even less has he shown that this cannot be done. But it is just false to say that he is committed to such a view.

Jackson often stresses that the 'method of cases' cannot refute two-dimensionalism even in principle (cf. e.g. Jackson 1998b, 213). For whatever one's verdict about a hypothetical scenario is, two-dimensionalism can take this judgment to be guided by our grasp of the relevant expressions' primary intensions. This may well be so.[36] But still, neither Putnam's nor Kripke's reliance on thought experiments has to commit them to (implicitly) assuming an internalist semantic value. This observation once more highlights that what is required is an independent defense of primary intensions, i.e. one which does not exclusively rely on the existence of the two-dimensional framework.

Since the scrutability thesis is quite ambitious, one may wonder if a proponent of conceptual analysis has to be committed to it. I pointed out before that there are many ways to construe the vocabulary used in the canonical description D, so the scrutability thesis need not be based on anything like PQTI. Moreover, the tenability of conceptual analysis surely does not depend on whether *every* truth is scrutable from the scrutability base. It has been argued for instance that there are true mathematical statements which are nevertheless not discoverable even by ideal rational reflection. In this case, they would not be epistemically necessitated by D, even though they are metaphysically entailed. But it is hardly plausible to derive an argument against conceptual analysis from the undecidability of certain mathematical statements. It is also questionable whether conceptual analysis has to assume that there is one basic vocabulary from which the

[36] It still has to be discussed how two-dimensionalism is able to handle the so-called epistemic arguments, which are also often based on hypothetical cases, though. This will be done in the following chapter.

other truths can be inferred. No actual cognizer is able to grasp a complete description of a world anyway, so for many practical purposes, this requirement seems irrelevant. In order to pursue conceptual analysis via the construction and evaluation of hypothetical cases, it is sufficient that we are able to determine the extension of an expression if given a description of a situation which does not explicitly use that term. Here, the described situation will typically only represent a tiny part of a possible world. And it is not even clear whether something like this can be done for each of our expressions, since there might be primitive expressions which defy any kind of analysis. Still, even though conceptual analysis need not be based on the scrutability thesis, this does not mean that an adherent of conceptual analysis should abandon the thesis. If correct, it may provide the theoretical foundation for conceptual analysis. And if it is combined with the idea that the grasp of a primary intension manifests itself in the respective ability to determine a term's extension with respect to a hypothetical scenario, then it can also help to give a reason for considering the intuitions elicited by thought experiments reliable. For these reasons, various versions of the scrutability thesis will play an important role throughout this thesis.

The scrutability thesis does not require that all terms are definable with the help of those in the scrutability base. This becomes apparent when one realizes that the point of many thought experiments is to undermine a proposed analysis of a specific term. Many have argued that it is impossible to give explicit analyses of most, or at least many philosophically interesting terms. These people often base their claim precisely on the fact that innumerable proposed analyses have been refuted by counterexamples. This arguably shows that it is possible to evaluate the invoked hypothetical scenarios even in the absence of any definition (cf. Chalmers & Jackson 2001, 320ff.). Given this, the fact that the primary intension, or the associated property, of a term does not have to be expressible by a description has to be considered a virtue of the theory. However, it may also be a reason to worry. If it turns out that it is impossible to extract any kind of analysis from a term's primary intension, then it becomes

questionable if primary intensions are of any use for conceptual analysis. The more general issue here is this: Even if it can be successfully argued that linguistic expressions are connected with primary intensions, it has not yet been shown that it is possible to gain any substantial philosophical insights via conceptual analysis. The second central aim of the following chapters – besides from the defense of the existence of primary intensions – will thus be to examine the practical epistemic value of conceptual analysis. There are two main issues to be addressed here: Firstly, how substantial are the a priori implications connected with a term – i.e., how much can be gained from the analysis of primary intensions? This question will mainly be discussed in chapter 5. And secondly, how is conceptual analysis to proceed – i.e., what could its method look like, and what are its aims? These questions will be the subject of chapter 6 and especially of chapter 7.

3 The challenge from the epistemic arguments

In this and in the following chapter, I will defend the two-dimensionalist thesis that every linguistic expression is associated with a primary intension. In the preceding chapters, I mainly discussed modal arguments which were supposed to show that the meaning of a name or natural kind term cannot be given by speaker associations, since nothing that a speaker associates with such a term is modally equivalent with it. We saw that while these arguments do undermine traditional brands of descriptivism, they have no force against two-dimensionalism. But there are prominent arguments of a different kind, the so-called epistemic arguments, which still need to be discussed.[37] Just like the modal arguments, the epistemic arguments most clearly apply to proper names and natural kind terms. And indeed, presumably most philosophers think there is nothing to be gained from an analysis of our concepts of tigers, gold or Feynman. Accordingly, if it can be successfully argued that names and natural kind terms have primary intensions, then a reasonably strong case has been made for the general thesis. But how important is the claim that *every* linguistic expression has a primary intension for a defense of conceptual analysis? As I have mentioned before, if the thesis turns out to be viable, this will be good news for the prospects of conceptual analysis. Nevertheless, it might be argued that the two-dimensionalist thesis is much stronger than required. There are basically two ways to weaken the thesis that every expression has a primary intension without denying conceptual analysis any significance in philosophical inquiry. The first is to say that although our terms do have conceptual associations which are accessible to a speaker, these associations do not amount to anything like primary intensions. One could hold for instance that some necessary conditions for a term's applicability

[37] The arguments from Ignorance and Error (see below) are called 'semantic arguments' by many, who thus distinguish a third kind of argument against descriptivism.

are a priori accessible, but no sufficient ones are. I will not discuss this idea any further here, though, since it implies a serious restriction of the aims and ambitions of conceptual analysis as it has traditionally been conceived.[38] This verdict does not necessarily apply to the second way of weakening the two-dimensionalist thesis in question, which is to concede that names and natural kind terms do not have primary intensions, but insist that other kinds of terms do. Such a position would be in line with a popular view according to which the semantics of names and natural kind terms is very different from that of at least many other kinds of terms (cf. e.g. Schwartz 1980; Salmon 1981, 66; Devitt & Sterelny 1999, especially ch. 5). A proponent of such a view could further argue that names and natural kind terms are not of primary philosophical interest anyway. It would thus be sufficient if terms like 'free will', (morally) 'good' and 'bad', 'justification', 'personal identity', 'consciousness', etc. were associated with primary intensions.

I grant that there is something to this kind of reasoning. Still, there are a number of reasons for a proponent of conceptual analysis to argue that all kinds of terms have primary intensions. Firstly, most notably Frank Jackson holds that conceptual analysis does not only have a role to play within philosophy. He argues that it is at least implicitly involved even in otherwise clearly empirical investigations, for example in the natural sciences (cf. Jackson 1998a).[39] To grant that some terms do not have primary intensions would thus already mean to restrict the potential scope of conceptual analysis. Secondly, it should be considered as an important theoretical virtue of two-dimensionalism that it offers a unified account of meaning. In a case where two theories are equally compatible with the data, the one which is simpler will generally be preferable. I thus think that it would be more problematic to base conceptual analysis on a 'mixed' semantic theory, especially if one deals with critics who invoke unitary externalist approaches. Thirdly, for a number of philosophically relevant

[38] I will say a lot more about the aims of conceptual analysis in chapter 7.

[39] Jackson's account of the role of conceptual analysis will be examined in more detail in chapters 4 and 7.

terms, it has been argued that they do refer to natural kinds – examples are ‘knowledge’ (cf. Kornblith 2002), ‘person’ (cf. Wiggins 1976) and ‘belief’ (cf. Lycan 1988, 31f.). If this turns out to be so for at least some philosophical terms, then to grant that natural kind terms do not have primary intensions means to diminish the role of conceptual analysis in philosophical inquiry. Fourthly, there are some prominent arguments in favor of semantic externalism which, although originally invoked in the context of names and natural kind terms, plausibly apply to other kinds of terms as well. An example for such a type of argument is provided by the argument from Ignorance and Error, which will be discussed below. If one accepts these arguments in the cases of names and natural kind terms, then it is hard to avoid the same conclusion for philosophical terms. It would thus obviously be nice to have a general strategy to resist these arguments and to defend the existence of primary intensions.

Here is a basic outline of this chapter: I will start by sketching two epistemic arguments. In a nutshell, the first of these arguments, which I will call the argument from empirical defeasibility, aims to establish that many terms have no a priori associations; the second one – the so-called argument from Ignorance and Error – seeks to show that speaker associations often cannot determine a term’s reference. After that, I will turn to a discussion of the implications of the former of these arguments for the semantics of natural kind terms, mainly by reference to the natural kind term ‘water’. My aim is to rebut the epistemic arguments by identifying properties which are a priori associated with ‘water’. Very roughly, I will argue, drawing on proposals due to David Lewis, Chalmers, and Jackson, that the primary intension of ‘water’ should be understood as involving a theory about a substance.
Subsequently, I will discuss an application of the epistemic arguments to proper names. Seizing on a proposal of Peter Strawson, Chalmers and Jackson argue that the properties associated with names are to be construed deferentially, i.e. as involving a reference to the usage of the name by other speakers. It will transpire that the appeal to deference can be understood as

a general strategy for dealing with the argument from Ignorance and Error. I will defend the notion of deference against some objections, but also point out that it does raise a couple of serious problems. My solution to the problems concerning deferential concepts will entail that two-dimensionalism is to be understood as an account of the contents of the thoughts expressed by utterances. I will conclude the chapter by discussing the consequences of this understanding of two-dimensionalism for conceptual analysis.

3.1 Primary intensions and the epistemic arguments

Let me start with the case of proper names. According to Kripke's argument from empirical defeasibility, everything we believe about the bearer of a name could, in the epistemic sense, turn out to be mistaken. It might perhaps seem obvious to someone that Gödel was the discoverer of the incompleteness of arithmetic or that Aristotle was a philosopher. But we can certainly envisage circumstances which would force us to abandon these beliefs (cf. Kripke 1980, 83f.). To elucidate the difference between these considerations and the modal arguments: What is important here is not that the beliefs in question concern contingent properties of Gödel and Aristotle. Rather, it is the fact that they are empirically justified and empirically defeasible. This is even true for beliefs we might have about necessary properties – say, that Kurt Gödel's parents were Rudolf and Marianne Gödel, or that Aristotle was human. The same kind of reasoning has been applied to natural kind terms: Nothing we associate with gold or cats seems to be a priori, in the sense of epistemically necessary: It could for example be discovered that gold is neither yellow nor is a metal (cf. Kripke 1980, 39), or that cats are not animals (cf. Putnam 1970).
The idea that all of our beliefs about some individual or kind are revisable in the light of empirical findings has some obvious affinities with Quine's view that no sentence is immune to revision and that there is thus no sensible distinction between the analytic and the synthetic (cf. Quine 1951). Of course, Quine's theses go further than the epistemic arguments currently

under discussion, which are (for the time being) restricted to specific kinds of terms. But if one succeeded in rebutting them by showing that there are epistemically necessary connections between our concepts, one would have also made a case against Quine's view.

The argument just presented aims to show that the terms in question do not have any a priori associations. There is another, related argument which also threatens to undermine the idea that names and natural kind terms have primary intensions – the argument from Ignorance and Error. The upshot of this argument is that frequently, speaker associations cannot determine the reference of the relevant terms. In some cases (in cases of Ignorance), what the speaker associates with a term is too unspecific to determine a particular referent; in other cases (in cases of Error) all or most of what she associates with a term is false. Let me again begin with the case of names. Kripke argues that many speakers know nothing to distinguish Gell-Mann from Feynman. They may know that Feynman is a famous physicist and that Gell-Mann is a famous physicist, but nothing else. Still, when they say 'Gell-Mann', they refer to Gell-Mann and when they say 'Feynman', to Feynman (cf. Kripke 1980, 81). Similarly, there are people whose sole belief about Einstein is that he invented the atomic bomb. But when they utter the term 'Einstein', their utterance does not refer to Oppenheimer or Szilard, but to Einstein. Again, the same seems to apply to natural kind terms. Hilary Putnam for example says that he is unable to tell apart elms from beeches in any way (cf. Putnam 1975, 226). Someone else's only belief about elms may be that they are New Zealand's most common deciduous tree. But still, so the argument goes, both refer to elms when they use the relevant term. The conclusion is that speaker associations cannot determine the reference of these terms in the actual world, much less so with respect to every other world considered as actual, which once more seems to show that these terms do not have primary intensions.

Let me try to spell out more precisely how the two arguments just outlined conflict with the two-dimensionalist thesis that every expression is associated with a primary intension: Primary intensions are supposed to be *a priori* accessible associations which *determine an expression's extension*

as well as its extension across all other possible worlds (considered as actual). But according to the first epistemic argument, i.e. the argument from empirical defeasibility, everything a speaker associates with names or natural kind terms is empirically defeasible. Consequently, these terms have no a priori associations whatsoever. The second argument, i.e. the argument from Ignorance and Error, aims to show that what speakers associate with an expression is often either false or too indeterminate to pick out a unique referent. Consequently, speaker associations often cannot determine a term's extension. The conclusions of both kinds of arguments are thus clearly incompatible with two-dimensionalism.

Nevertheless, Chalmers and Jackson are not impressed by the epistemic arguments. I mentioned before that Jackson thinks that (his brand of) descriptivism cannot be refuted by the 'method of cases' (cf. Jackson 1998b, 213) – of which the epistemic arguments surely are instances. Chalmers specifically says that the epistemic arguments have no force against two-dimensionalism (cf. Chalmers 2002b, 173). The reason for their optimism becomes apparent when one considers the way in which these arguments are supposed to establish their conclusions. The epistemic arguments usually proceed as follows: First, a potential belief of a subject about a given individual or kind is identified. Then it is either shown that this belief is not a priori or that it cannot determine the reference of the relevant term. From this, it is concluded that nothing that a speaker associates is a priori or determines the reference. However, there are a number of reasons why an argument with this structure seems ill-suited to attack two-dimensionalism: For a start, in such an argument only a specific belief is picked out of which it is then argued that it is not a priori or that it cannot determine the term's reference. But in itself, this only shows that the belief in question does not express the term's primary intension, not that the term has none. Furthermore, it is a mistake to think that primary intensions have to mirror a speaker's most salient beliefs about the referent or category in question. Accordingly, they need not correspond to what an average speaker would spontaneously say if asked what she associates with a given term. In fact, the associated properties do not have to be readily

accessible to a speaker at all, which is one of the main reasons why conceptual analysis is often so hard to do. Two-dimensionalism is not even committed to holding that a term's primary intension is articulable in principle. All that is required is that a subject has the conditional ability to determine the term's extension with respect to a scenario if she is provided with a canonical description of that scenario. So a typical two-dimensionalist response to an epistemic argument is to say that those speaker beliefs of which it is shown in a thought experiment that they cannot determine a term's reference or that they are not a priori simply do not express the associated primary intension. As I mentioned in the preceding chapter, Chalmers and Jackson even go one step further and suggest that the epistemic arguments themselves presuppose that we do have the ability to determine a term's reference with respect to hypothetical scenarios (cf. Chalmers & Jackson 2001, 326f.). When for example an argument is brought in which aims to show that a particular belief is not a priori, then that argument itself proceeds by giving a description of a scenario from which we are supposed to infer that the belief is false with respect to the scenario.

If it could indeed be shown that the epistemic arguments themselves presuppose the ability expressed by the scrutability thesis, then we would be warranted to conclude that the epistemic arguments have no force at all against two-dimensionalism. But I also already noticed in the previous chapter that this is far from clear: In Kripke's famous 'Gödel' case, for example, it is just stipulated that it was Gödel who did such and such, it is not read off from a description of the scenario. In the case of the argument from Ignorance and Error, it is even less clear how the relevant thought experiments are supposed to conform to the scrutability schema: At least in many of the relevant cases, we as external and supposedly knowledgeable observers judge what a term used by an ignorant or errant speaker is supposed to refer to. But the fact that we can determine the term's reference with respect to that scenario in no way implies that she herself could do so.

Generally speaking, I agree with Chalmers and Jackson that the epistemic arguments do not refute the idea that names and natural kind terms have

primary intensions. But unlike them, I nevertheless think that they raise a serious challenge. For, if a speaker can meaningfully use a term without associating anything substantial with it, and if we are often at a loss to identify any a priori connections of a term, what reason do we have then to suppose that primary intensions nevertheless exist? Why should we for instance assume that Putnam would be able to determine the extension of 'elm' with respect to a given scenario? I thus believe that it would be an important component of a defense of primary intensions to at least give an idea as to what the primary intensions of names and natural kind terms might be like.[40] This is what I will do in the following two sections. I will not, however, try to give a precise definition for any of these terms. I only want to say enough to make it plausible that names and natural kind terms do have primary intensions.

3.1.1 The primary intensions of natural kind terms

By far the most paradigmatic of all natural kind terms, at least judging from its presence in philosophical discourse, is 'water'. It will thus also serve as my prime example. Given the – literally – vital role water plays in our everyday life, it is safe to assume that the average speaker's (true) beliefs about it are easily sufficient to fix the term's reference. One should therefore expect the argument from Ignorance and Error to have little force here. But it would be unwise to assume that this kind of argument fails to apply to natural kind terms in general. For it is plausible that many speakers' beliefs connected with terms such as 'lepton', 'chromosome' or 'dung beetle' are much less distinct. The argument from Ignorance and Error will be discussed in detail in the next section, as it has traditionally been conceived to be most pertinent in the case of names, and not without reason. There, the question to what extent such an argument is applicable to other kinds of terms will also be addressed. In the case of 'water', the

[40] The fact that Stephen Laurence and Eric Margolis complain that Jackson says curiously little about the specifics of the primary intensions of particular concepts may yield some further support to this thought (cf. Laurence & Margolis 2003, 262).

argument from universal defeasibility is more relevant: It has often been questioned whether anything we associate with water is a priori. The point is familiar from the writings of Putnam, who argued that natural kind terms are only associated with stereotypes, and that even the most central of these stereotypes are empirically defeasible (cf. Putnam 1975). In the following, I will thus try to identify properties which could be a priori associated with natural kind terms.

How should one go about to do this? Remember that grasp of a term's primary intension is connected with a subject's ability to determine the term's extension with respect to a hypothetical scenario, if provided with a canonical description of that scenario. So we should just try to find out what it is that guides our judgments with respect to these scenarios. If we find some property such that there is no scenario where water does not have this property, then we have identified an a priori associated property – and thus a genuinely associated property in Jackson's sense, as explicated above. Of course the method in itself is anything but unusual. It is just the way conceptual analysis typically proceeds, on the basis of thought experiments. It should be noted, however, that when the concept of free action, or justification, or probability are analyzed, the aim is usually to discern (metaphysically) necessary features of these phenomena. But in the case of 'water', we should not expect to discover its essential properties by way of conceptual analysis in the first place, since we already know that the primary intension of 'water' differs from its secondary intension. It is thus crucial that the scenarios to be evaluated are considered as actual.

One might wonder whether the attempt to show that a term has a primary intension by considering our judgments about hypothetical cases begs the question, since it presupposes that these judgments themselves are a priori. This thought is mistaken, however, for the following reason: The current line of reasoning aims to rebut arguments to the effect that for any putative a priori connection one can give counterexamples – i.e. hypothetical circumstances of which we are supposed to judge that if they were to occur, the connection would fail to hold. Even if, as was argued above, this kind of objection need not itself follow the scrutability schema, a response to it

which invokes hypothetical scenarios as well can hardly be accused of begging the question for doing so. There are certainly other, and in a way more radical arguments against the view that there are a priori connections between our terms or concepts. Especially in chapters 4 and 6, I will therefore also pursue the question of whether there are more general considerations which speak in favor of the existence of primary intensions.

From previous discussions, we already learned that the term 'water' picks out the kind which it actually picks out in every possible world (considered as counterfactual). Any attempt to express the term's primary intension should thus make use of some 'rigidifying' expression such as 'actual'.[41] Another lesson which can be drawn from Putnam's Twin Earth scenario is that we are not willing to call XYZ water because it is not the liquid in *our* rivers and lakes, i.e. because we are not acquainted with it. It thus seems plausible to assume that some causal connection to us is an epistemically necessary condition for being water.[42] But even so, actuality and the existence of an acquaintance relation are clearly insufficient to uniquely pick out water. One will therefore have to identify other a priori implications of the term 'water'.
Two-dimensionalists often say that water is 'the actual watery stuff' or something to that effect.[43] But what exactly does 'watery stuff' stand for? Chalmers' original formulation is "the dominant clear, drinkable liquid in the oceans and lakes" (Chalmers 1996, 57). The phrase 'watery stuff' thus simply stands for a description of some of water's most glaring superficial properties. We are now in a position to give a first approximation of the primary intension of 'water': 'the dominant clear, drinkable liquid in the

[41] This only becomes relevant when one considers a given scenario as counterfactual, though.
[42] This requirement is discussed for example in Jackson 1998a, 38f.
[43] Chalmers & Jackson 2001 put it thus: "the watery stuff in our environment" (341). Here, the reference to our environment roughly corresponds to what I expressed by saying that our acquaintance with it is a necessary condition. I will not go into these subtleties here.

oceans and lakes, of our actual acquaintance'.[44] There are some obvious objections to this proposal, though. It is true that the enumerated properties plausibly correspond to what an average speaker associates with water. But these clearly seem to be a posteriori associations. We can easily imagine water which is, say, quite murky. And even worse, not even actual water is always clear. Likewise, there is a lot of water with which we are not acquainted, which is solid, not drinkable or does not flow in rivers and lakes.[45] So it seems that as it stands, the above proposal does not even get off the ground.

Apparently it is not a good idea to understand the description 'watery stuff' as a universally quantified conjunction such as 'For all x, if x is a sample of water, then x is clear and a liquid and drinkable and …'. So one might try to adjust this formulation by including the conditions under which water has the properties in question. For instance, water is clear and also drinkable when it is free from certain impurities; it is liquid when the temperature/pressure ratio is within such and such a range; etc. But for a number of reasons, this would not work, either. To point out just one problem, many of these conditions are simply not known to the average speaker, not even implicitly, and so they cannot be among the associated properties. I think the following observation provides the key to a better solution: People naturally assume that the liquid they mostly drink from bottles (and typically when it is clear) and that which flows in our rivers and lakes and that which falls from the sky in drop shape etc. is the same substance. The term 'water' is supposed to denote that substance. Adding some arsenic or mud to it, for instance, may change some of its superficial properties, but it won't make it a different substance. So the description 'the dominant clear, drinkable liquid in the oceans and lakes, of our actual acquaintance' can be seen as a theory about a particular substance, or as it

[44] In fact, I think it would be more precise to say 'the dominant clear, drinkable liquid in the oceans and lakes, of *my* actual acquaintance', since it is plausible that only the relevant subject's acquaintance with the substance is relevant.

[45] This objection is raised by Laurence & Margolis 2003, 262f.

is sometimes put the theoretical role played by water.[46] Accordingly, this theory only says that the substance plays the theoretical role in question, not every sample of it. By the same token, it should be taken to imply that we are acquainted with the substance, not with every sample of it.

Let me test the current proposal by means of the following conditional: 'If a substance stands in a causal connection to us which is clear, drinkable, …, then it is water'. So can a substance have all these properties and still fail to be water? This seems indeed implausible.[47] Likewise, a substance which does not have any of these properties would hardly be water. This permits us to formulate a second conditional: 'If something does not stand in a causal connection to us, is not clear, not drinkable, …, then it is not water.' It would not provide a good objection against these theses to point out that if H_2O had never been watery or if we had been acquainted with XYZ, the watery stuff of our acquaintance would not have been water. For remember that we are dealing with issues concerning apriority here, and thus we have to consider the worlds in question as actual. With respect to this way of considering scenarios, I already argued in the preceding chapter that if the watery substance in our rivers and lakes had turned out to be XYZ, then plausibly, water would have turned out to be XYZ. Note that the reference to a particular substance is crucial here: There might be samples of water which do not have any of the specified properties, and there might even be substances other than water samples of which have all of the properties specified by the water role.

Unfortunately, even if one conjoins the two conditionals just brought in, one still has not thereby given necessary and sufficient conditions for a substance to count as water. This is because they are silent concerning the question what one should say in cases where the theoretical role is partly satisfied. But let me set this problem aside for a second. In any case, the conditionals do suggest that there is a conceptual connection between a

[46] Jackson says in a number of places that the Twin Earth scenario contributed to eliciting our 'folk theory' of water (cf. Jackson 1998a, 38). The idea to treat 'water' as a theoretical role term is also at least implicit in Lewis' remarks in Lewis 1994.

[47] Wolfgang Schwarz also argues for this thesis (cf. Schwarz 2009, ch. 11).

substance's being water and its satisfaction of a specific theoretical role, which corresponds to what Jackson calls the 'folk theory' of water. Let me thus say that, roughly, for something to be water, it has to be (a sample of) a substance which satisfies the folk theory of water.

Plausibly, our folk theory is much richer than what is listed in the description given above: One might add that water does not smell, that it boils at 100° Celsius, expands when it freezes, falls from the sky in crystal shape when the ambient temperature is low, etc. We would then have a comprehensive cluster of properties which define our concept of water.

According to the current understanding, 'water' is a theoretical role term. With one peculiar feature, though: It rigidly denotes the kind which it actually picks out. One function of the term is thus to secure reference to the substance which actually satisfies the theoretical role in (metaphysically) modal contexts, not to hold fixed the theoretical role itself. Among other things, this insight provides us with one more way to illustrate why the primary intension of 'water' diverges from its secondary intension: We have a priori access only to the theoretical role played by the substance, not the nature of the satisfier of the role itself.

However, there are still a couple of residual questions. One concerns the criteria for two given samples of matter to count as samples of the same substance. An obvious proposal is to say that the categories are determined by nature. But since according to two-dimensionalism, reference ultimately relies on speaker associations, this would have to be specified by the term's primary intension. So one could, like Lewis does, hold that it is a part of our 'folk theory' that water is a natural kind (cf. Lewis 1994, 424). Maybe it is – but is it really a priori that water is a natural kind? If it had turned out, say, that what flows in our rivers and lakes etc. is actually a mixture of many different chemical substances, would it really have turned out that there is no water? Hardly, I think. People once surely believed that air is a natural kind – like water, it was even considered as an element from Antiquity to early Modern history. But when it turned out that what we breathe is not at all a natural kind, but rather a mixture of many different gases, this did not imply that there is no air – just like Lavoisier did not

discover that there is no water, but only that water is not an element. One could in fact raise the same worry with respect to most, if not all of the features which our folk theory attributes to water. Could it not have turned out that water is never clear and that we are all just subject to some kind of illusion? Or that the liquid which rains from the sky is actually not water, but only becomes water when it merges with the substance in our rivers and lakes due to some chemical process? Even worse, it is very unlikely that there is a substance which perfectly satisfies our folk theory of water – we surely got some properties of the liquid around us wrong –, but it would be most unfortunate if one had to conclude that there is no water.

Recall, however, that the question of what one should say in a case where the theoretical role in question is satisfied partially was left open above. In view of the problems just identified, one might thus suggest that less-than-complete satisfaction is also sufficient for a substance to count as water. Such a proposal would be in line with Lewis' view, who argues in *How to Define Theoretical Terms* that for a theoretical term to refer, the corresponding role does not have to be satisfied perfectly (cf. Lewis 1970, 432). As applied to 'water', this would mean that a substance only has to have a sufficient number of the properties in the folk theory in order for the term to apply to it. If one adopted such a view, one could adhere to the idea that there is a conceptual (a priori) connection between water and, say, freezing at 32° Fahrenheit without it being strictly a priori true, i.e. epistemically necessary, that water freezes at 32° Fahrenheit.[48]

The current account of the semantics of natural kind terms like 'water' crucially relies on the idea that for something to fall under the term 'water', it only has to satisfy the associated theoretical role to some sufficient degree. One might object that such a move seems ad hoc. It may make sense for pragmatic reasons to say that if the theory only roughly applies to some kind, we should still take it to refer to that kind – but is this really a conceptual truth?

[48] It is of course a difficult question which degree of satisfaction of the theoretical role is required. I will turn to this issue in 3.1.1.1.

I agree with the objection in that I think that it is questionable whether Lewis' proposal yields a plausible general condition for the reference of theoretical terms: There might be theoretical terms which only refer if the associated theoretical role is satisfied perfectly; that possibility should at least be left open. However, in the case of 'water', requiring only partial satisfaction of the folk theory is consistent with our judgments about hypothetical cases: We are inclined to call a substance which satisfies most, though not all of the properties in the cluster, 'water'. And generally speaking, if one considers the primary intension of 'water' just as expressing a theory about a particular substance in the way proposed above, then this inclination seems perfectly reasonable. For, the idea behind that proposal was roughly this: We introduce the term in question when we encounter various samples of a liquid with certain properties and hypothesize that these are all samples of the same liquid, i.e. of a particular substance. Since (at least initially) we do not know anything about the essence of this substance, the theory associated with the term primarily serves to determine which substance it is that we talk about – it serves so to speak to fix the term's reference.[49] For that purpose, it would be impractical if we took it to be required that the substance has all of the properties we invoke to do this.

Let me add that one should not assume that all of the properties in the cluster are equally important, so there has to be some kind of internal weighting. In fact, while I think there is a lot to be said for the account at hand, I am not sure that it is a good idea to hold that all of the features which are involved in our folk theory of water are relevant to the primary intension of 'water'. I believe it is plausible to assume that much of what folk theory postulates would actually be given zero weight.[50] On the other hand, it could well be that there are properties in the cluster which

[49] This is in fact a simplification. I will discuss the function(s) of epistemically opaque terms like 'water' in some detail in chapter 5.

[50] Of course this only means zero conceptual weight, concerning what guides our evaluation of hypothetical scenarios. As empirical criteria, these properties can still be very important in our epistemic practice.

constitute strictly epistemically necessary conditions for being water. I already mentioned that the existence of an acquaintance relation could be epistemically necessary. But given that these subtleties are not my main concern and that it would take up a lot of space to settle them, I will not argue for any of these theses here.
I just expounded how the primary intension of the natural kind 'term' water is to be construed. But before it can be concluded that the epistemic arguments are ineffective against this construal, there are two issues which still need to be discussed, namely vagueness and intersubjective variation in primary intensions.

3.1.1.1 Vagueness

I just suggested that the properties in the cluster which constitute our concept of water are most likely not all equally important. In principle, we do have a way to get a grip on their relative importance if we wanted to settle this question. We could just rely on the familiar method of testing our responses to various hypothetical scenarios to see to what extent they are guided by some property or other. It would still be unwise to expect that such a procedure could yield particularly precise ratios; not only because it is hard to determine them, but mainly because there probably are none to be determined. Nearly all of our terms are infected with vagueness. In the case of the term 'water', whose extension is dependent on a considerable number of features of varying degrees of importance, it is safe to assume that the degree of vagueness will be quite significant. Now it seems to me that a particular kind of critique which is sometimes raised against the thesis that natural kind terms have primary intensions actually builds on this phenomenon. Nimtz for example denies that we are able to determine the extension of 'water' with respect to every world considered as actual (cf. Nimtz 2004). One scenario he invokes is such that there is a transparent, tasteless liquid which fills the rivers and lakes and which can be used to extinguish fire, but which never freezes or evaporates and can be walked on (cf. Nimtz 2004, 139). Would this liquid be water? Nimtz argues

that we cannot tell and moreover, since it is possible to contrive many more of these undecidable cases, the two-dimensionalist thesis that speakers have a conditional ability to determine the extensions of their terms if given a qualitative description of a scenario is false. One can certainly argue over particular cases – although the scenario is somewhat underspecified (does the liquid deviate in any other ways from our folk theory of water, is it walkable because its surface tension is higher, or its density, or because the density of humans is lower, …?), I for one am inclined to think that in this case it would turn out that we can walk on water. Moreover, I think it is possible to construe cases where the substance in question satisfies nearly everything we attribute to water and which we would definitely consider as water; and other cases where a substance satisfies almost nothing of it and would thus clearly not be water. But still, like I said before I agree that it should be possible to devise a large number of cases concerning which we would be hard pressed to say whether the described substance is water or not and where indeed a judgment in either direction would be arbitrary. But this fact need not amount to an objection against two-dimensionalism if it traces back to vagueness. This is because, plausibly, a hypothetical scenario which represents a borderline case is just one in respect to which it is undecided whether the term applies or not. Two-dimensionalism can and should only require that we are able to determine a term's extension with respect to worlds where it has a determinate extension. If all vagueness is due to semantic indecision, then this should be exactly mirrored by the indecision in our judgments. Of course, if the epistemic theory of vagueness is right and (seeming) borderline cases are actually semantically (and/or metaphysically) determinate though indeterminable for us (cf. e.g. Williamson 1994, Sorensen 2001), and if moreover we would not even have epistemic access to the boundaries of our terms if endowed with complete empirical knowledge and ideal rational capacities, then vagueness poses a problem to the two-dimensionalist framework. But while I concede that much, I do not think that such a theory offers a particularly plausible account of vagueness.[51]

[51] Jackson 2002 develops a critique of the epistemic theory of vagueness that is based

3.1.1.2 Intersubjective variation and the individuation of concepts

Chalmers and Jackson hold that primary and secondary intensions are attributed to expression tokens rather than to expression types (cf. e.g. Chalmers 2006, 64). In this way, they allow that there is intersubjective variation in the properties associated with a linguistic expression. Laurence and Margolis argue that this would mean that a person with a slightly deviant primary intension does not possess the relevant concept, which they deem implausible (cf. Laurence & Margolis 2003, 262f.). By the same reasoning, one could conclude that people who do not associate the same primary intension with an expression cannot be said to share the concept in question. This may seem unfortunate, especially if it turns out to be a frequent phenomenon. So how common should we expect intersubjective variation of primary intensions to be in the case of natural kind terms? Chalmers seems to think that it is at least not obvious that one will find such a kind of variation (cf. Chalmers 2002b, 174), but this will in fact be hard to avoid. If one analyzes terms like 'water' as cluster terms as has just been proposed, then this becomes particularly clear. There is a considerable number of properties which define the theoretical role associated with the term. It is quite unlikely that two speakers' primary intensions comprise exactly the same properties weighted in exactly the same way.[52]
It is not clear whether intersubjective variation in itself is a problem for two-dimensionalism, however. A two-dimensionalist is not committed to holding that concepts are always individuated via their primary intensions. It is true that for many kinds of terms, there are even good independent reasons to hold that they are individuated not just via their extension (in order for instance to distinguish co-referential concepts like 'phlogiston' and 'ether') or their secondary intension (in order for instance to distinguish co-intensional concepts like 'Hesperus' and 'Phosphorus'), so it

---

on ideas concerning the role of language in communication which will be discussed in the following chapter.

[52] I will argue in 4.2 that with respect to at least some natural kind terms, one should expect to find very substantial variation in the associated primary intensions.

makes sense to draw on primary intensions in the individuation of concepts. But firstly, one does not have to require that concepts are generally sensitive to differences in primary intensions; and secondly, even concerning those concepts which one deems to be so sensitive, there are a number of ways in which primary intensions can be involved in their individuation. One could for instance hold that some kinds of concepts, such as maybe natural kind concepts, should be individuated via their secondary intensions only. An alternative would be to say that for two people to share a natural kind concept, the secondary intensions have to be identical and the primary intensions sufficiently similar to a degree which would still have to be specified. I think that this latter proposal has its merits at least in the case of 'water'. Various options seem reasonable, depending on the kinds of concepts concerned or even on one's explanatory purposes. In any case, two-dimensionalism is not a theory about the individuation of concepts and therefore, it does not entail any of these theses.

There is a related issue, though, which cannot be avoided as easily: When primary intensions can vary wildly among speakers within a linguistic community, this raises the question which role primary intensions can actually play in a theory of linguistic meaning. I will defer a discussion of this issue to section 3.2.

To sum up – how does the account of the meaning of 'water' just presented fare against epistemic arguments according to which natural kind terms do not have primary intensions? Two-dimensionalists have pointed out that those attempts to give a descriptivist characterization of the meaning of terms such as 'water' and the like which were traditionally discussed were essentially incomplete: Firstly, they did not allow for rigidification, thus making them vulnerable to the modal argument. Secondly, they did not account for the role played by causal relations. Accordingly, one should just include these features into the term's primary intension. But even then it is not trivial to provide sufficient conditions for the applicability of the term – i.e. conditions which are a priori and which are sufficient to

determine the term's extension with respect to every world considered as actual. To solve this problem, I proposed to consider 'water' as a term whose primary intension comprises a theory about a particular substance.
It is plausible that at least some of the properties which make up our 'folk theory' of water are in some way conceptually connected with 'water'. One reason to assume this is that it seems inconceivable that a substance has all of these features, yet fails to be water. Likewise, it seems inconceivable that a substance fails to have any of these features and is still water. Another reason is that adding or removing some of the properties in question in a hypothetical scenario does make a difference in our judgments about these scenarios. An important virtue of the proposed analysis is that it can uphold such a conceptual connection, while being compatible with the observation that all or nearly all of the relevant properties are empirically defeasible: If a substance fails to have one of these properties, it can still be considered as water if it has enough of the other properties in the cluster.
Furthermore, the fact that it is possible to construct cases which are not determinately evaluable does not speak against the existence of primary intensions if, as seems plausible, it can be attributed to vagueness and thus to semantic indeterminateness. And finally, an argument from intersubjective variation did not prove to be fatal, either. I conclude that the epistemic arguments do not provide any principled obstacles to assuming that natural kind terms such as 'water' do have primary intensions.

3.1.2 Semantic deference and the primary intensions of names

One might try to apply an account akin to the one I just gave of natural kind terms to proper names: One could start by identifying a 'theoretical role' associated with a person, maybe including some remarkable things the person has done or other particularly salient features of hers. In a second step, one would then rigidify the resulting description and maybe add some additional constraint concerning our acquaintance with that person or alternatively the existence of a more indirect causal relation. But this

strategy is unlikely to succeed. In the case of names, Kripke's epistemic arguments against descriptivism seem very convincing. He did not only show that for any property of a person which comes to mind, it could turn out that the person does not have this property; but also that it could even be discovered that the person has none of the properties generally associated with her. It could well be for example, that none of the things the Bible attributes to the prophet Jonah is true, i.e. that God ordered him to go to Niniveh, that Jonah resisted and was swallowed by a whale when trying to flee on a ship, that he was vomited up by the fish and then went to Niniveh to preach after all, that he became angry when God decided to spare the city, etc. Or alternatively put: Even if we found out that no-one did any of the things just listed, the name 'Jonah' would still not have to be empty. Accordingly, it seems that nothing we associate with a name is a priori, not even the disjunction of properties commonly attributed.[53] This observation already suggests that if one nevertheless wants to argue that names have a priori associations, these associated properties will have to be quite different from what speakers superficially associate with a person.

As I mentioned before, there is another kind of argument, the argument from Ignorance and Error, which militates against the idea that proper names have primary intensions: In many cases, speakers seem to associate nothing or close to nothing with a name (in cases of Ignorance), or most of what they associate with a name – or even everything – is false (in cases of Error). If one aims to defend two-dimensionalism against the epistemic arguments in the case of names, one will therefore have to argue for two things: Firstly, one has to show that speakers do generally associate something with a name which suffices to fix the name's reference. Secondly, one must make a case that these associations are associated

[53] Frederick Kroon nevertheless argues that at least *some* names, and in particular names such as 'Jonah' which are only known to us from a specific story in which they are embedded, do function like theoretical role terms (cf. Kroon 1983). In 4.2, I will argue that names may work like theoretical role terms in cases where we are closely familiar with the referents.

properties in the proper sense, i.e. a priori associated properties. Let me thus take the argument from Ignorance and Error as a starting point.

3.1.2.1 The argument from Ignorance and Error

The argument from Ignorance and Error is particularly pressing because its potential scope plausibly extends far beyond proper names. Take for example Tyler Burge's famous 'arthritis' thought experiment which is designed to show that we should be externalists about mental content (cf. Burge 1979): Imagine a person who truly believes that arthritis is a disease, that she has it in her wrists and in her fingers, that stiffening joints is one of the disease's typical symptoms, etc. In addition, she thinks that she has recently developed arthritis in her thigh. When this person visits her doctor, she is surprised to hear that this is not possible, because arthritis is an inflammation of joints. But she accepts what her doctor tells her and discards her belief. The person in this case has a severely mistaken view of what arthritis is and still she (at least intuitively) possesses the concept of arthritis. Another example is provided by Putnam's afore-mentioned ignorance concerning elms and beeches: Although he is unable to distinguish elms from beeches in any way, Putnam nevertheless refers exclusively to beeches when he says 'beech'. One can easily give many more examples which show that this phenomenon does not only occur in the case of names or natural kind terms (like 'elm' and 'lepton'), but also many other kinds of terms – examples are 'arthritis', 'MRI' and 'transistor'.

Externalists argue that cases like these show that reference cannot be determined by features intrinsic to the subject. The second part of Burge's thought experiment makes this point even clearer: Imagine a complete intrinsic duplicate of our patient, who lives in a different linguistic community. When she goes to her doctor to report her worry, it emerges that she is actually correct. In her linguistic community, 'arthritis' labels a disease which can affect joints and muscles; we may call this disease

‘tharthritis’. Consequently, reference does not supervene on a subject’s intrinsic properties.
So how is reference determined according to the externalists? On Putnam’s view, there is a division of linguistic labor (cf. Putnam 1975). Since there are experts in my linguistic community who can distinguish elms from beeches, I need not have the ability to do so myself and can still meaningfully use the relevant terms. Burge’s account is quite similar. He invokes the notion of semantic deference: Because of her incomplete understanding of the term, the patient defers to her doctor’s usage because she considers him an expert. The reason why the patient’s concept differs from that of her duplicate is thus that the word ‘arthritis’ is used differently by the relevant experts in their linguistic communities.

3.1.2.2 Deferential concepts and the alleged problem of circularity

Chalmers and Jackson make very similar proposals to deal with the problem just outlined. Jackson points out that Putnam does after all know something to distinguish elms from beeches: He knows that elms are called ‘elms’ and that beeches are called ‘beeches’ by the experts in his linguistic community (cf. Jackson 1998b, 209). Jackson’s proposal is then to consider this as being among the properties associated by Putnam. Likewise, in his discussion of Burge’s thought experiment Chalmers argues that the patient’s intention to defer is represented in her concept’s primary intension (cf. Chalmers 2002c, 2003). On a first approximation, one could thus describe the primary intension which Putnam putatively associates with ‘elm’ by ‘the tree called ‘elm’ by the experts in my linguistic community’; the primary intension associated by the patient in Burge’s hypothetical scenario with ‘arthritis’ could be described analogously. The idea to transfer this account to the case of proper names suggests itself. Owing to the insight that no facts about a person’s famous deeds are relevant for the determination of the reference of a name, but rather the existence of a causal chain leading from the introduction of the name to my use of it, the

primary intension of a name like ‘Gödel’ could then be something like ‘the individual called ‘Gödel’ by those from whom I acquired the name’.
Chalmers and Jackson thus regard the use of the term by other members in the linguistic community not as a content-determining factor external to the speaker, but as a part of her concept’s content. In fact, the general idea is anything but new. A similar proposal was already made by Strawson in *Individuals* (cf. Strawson 1959, 182 fn.), but it seems to have been widely rejected. One of the main objections which have been raised against accounts of this kind is that they involve a circle. This is most likely due to Kripke’s discussion in lecture two of *Naming and Necessity*. At the beginning of the lecture, after having listed six descriptivist theses, he adds the following clause:

> (C) For any successful theory, the account must not be circular. The properties which are used in the vote must not themselves involve the notion of reference in such a way that it is ultimately impossible to eliminate. (Kripke 1980, 71)

And he comments:

> (C) is not a thesis but a condition on the satisfaction of the other theses. Theses (1)–(6) cannot be satisfied in a way which leads to a circle, in a way which does not lead to any independent determination of reference. (Kripke 1980, 71)

So one might suspect that an account where the description which is supposed to fix the reference of a term itself involves the term ‘reference’ or equivalent notions is circular.[54] For precisely this reason, Jackson tries to avoid any explicit mention of reference by proposing that what a layman associates with the word ‘beech’ is “having the property, whatever it is, that the experts associate with the word ‘beech’” (Jackson 1998b, 210; cf. also Jackson 2004, 271). He thus simply replaces ‘reference’ by ‘associated properties’ in the term’s A-intension, and since on his view reference is in

[54] Strictly speaking, since primary intensions need not be expressible by linguistic descriptions, a two-dimensionalist is not committed to holding that the term ‘reference’ (or a related expression) is used at all. But still, the fact that on the current proposal the reference relation is among the properties associated with the relevant terms might give rise to the same kind of worries.

fact secured by associated properties, this will boil down to the same thing. However, while it is possible that what Frank Jackson associates with deferential concepts is sensitive to associated properties as opposed to the reference relation itself, it would be quite contentious to assume that this is also true for Putnam or even for an average speaker who has not been trained in the philosophy of language. Jackson could insist that what is really relevant are a speaker's judgments given ideal rational reasoning and argue that since the description theory of reference is a priori true, in this sense every speaker will judge that reference is determined by speaker associations given ideal rational capacities. But firstly, at least as a response to an externalist, this only begs the question. Secondly, even if some version of descriptivism is true, it is still not obvious that it is a priori so.

But be that as it may, I fail to see how the mere fact that a term's primary intension involves the notion of reference has to lead to a circle. And in fact, this is not precisely what Kripke objects to regarding Strawson's account, either. To illustrate Kripke's actual point, consider a nice example which is due to Kroon: Imagine that before Smith's party, Jones tells someone that he wants his use of 'Bud' to stand for whoever is referred to by the first person at the party who uses the name 'Bud'. As it happens, Smith overhears a part of this conversation, but only grasps that Jones is talking about some Bud and when he later utters the name 'Bud', intending to defer to Jones' use, he is in fact the first to use the name 'Bud' at the party (cf. Kroon 1989, 376). In this example, there is indeed a circle and thus, the reference of 'Bud' is not grounded. Accordingly, when the reference of a term is determined by a description which appeals to the reference of someone else's use of the term, one can be led in a circle. But note that in the example just invoked, Smith and Jones were just unlucky. If for example Miller had been the first to use the name 'Bud' in a conversation about his favorite actor who is best known for his role in 'Harold and Maude' (Bud Cort), then there would have been no circle. So the current proposal to construe the primary intensions of names deferentially at least does not necessarily involve a circle. Kripke's worry,

however, is that in practice, it is hard to make sure that this does not happen (cf. Kripke 1980, 90). In his view, Strawson has to require that one knows from whom one has acquired the name (or whatever kind of term) in question, that one knows a chain leading from my use of the name back to its introduction, and that one knows that everybody in the chain uses the right kind of description.

Of course it is implausible that an average speaker knows all of these things. But in fact, why should a speaker be required to know any of this? Take a case where a speaker's primary intension of 'Gödel' is equivalent to 'the individual called 'Gödel' by those from whom I acquired the name'. Typically, the persons deferred to will also defer to those from whom they acquired the name and so on, until eventually a person defers to the person or persons who introduced the name in the 'original baptism'. In such a case, there is thus no problem apparent. But of course, there can also be cases in which there is no such regular chain of reference. In the following, I will distinguish three kinds of cases in which things do not go the normal way. As will become apparent, none of these cases creates a problem for the deferential understanding of speaker associations, either.

Take first a scenario where some person in the chain steps out of line and does not intend to use the name in accordance with those from whom she heard it. Let us say that Alvin hears the name 'Angela Merkel' and since he likes it, he decides to use it to refer to his motorbike.[55] Now when Alvin mentions the name 'Angela Merkel' in a conversation with Batu, who then defers to Alvin's use, the current version of descriptivism is committed to holding that when Batu uses the name 'Angela Merkel', her utterance does not refer to Germany's first female Chancellor, but to Alvin's motorbike.[56] However, if Alvin is really the only person from whom Batu heard the

[55] Kroon discusses several of such cases of what he calls 'indirect reference borrowing' (cf. Kroon 1987).

[56] Alternatively, the utterance could fail to refer, in case that the name is not used purely deferentially by Batu, who may for example take it to be a priori that it refers to a person.

name,[57] then this does seem to be the correct result. Against this, could the externalist just say that a speaker's intentions do not matter and that all that is required is the existence of a causal chain? I think it is hard to deny that a speaker has to intend to use the name in line with those from whom she heard it, not least since it must be possible for a speaker to endow a (generic) name which is already in use and which is familiar to her with a new meaning if she wishes to. It is thus not surprising that even externalists usually grant this condition (cf. e.g. Kripke 1980, 96).

Now to the second way in which the reference of a name can fail to be grounded in the way it usually is: Assume that everybody in the causal chain intends to use the name like the person from whom she heard it. But there is no proper 'original baptism'. The person who introduces the name may somehow fail to name anyone, or may name more than one person. In this case, the view of Strawson, Chalmers and Jackson predicts that the name is empty or ambiguous – but again, seemingly rightly so.

The third way in which reference can fail to be established was already discussed: Kroon's example showed that the speaker associations can lead into a circle. When two persons defer to each other's use of the name, then the name is empty. But again it is hard to see how the causal theory of reference could treat the case differently.

Similar considerations apply to deferential uses of other kinds of terms. Here, one should expect that the causal chain will mostly not lead to a 'baptism', but rather come to an end when a person deferred to, uses the term non-deferentially. Like in the case of names, the speaker neither has to know that reference will eventually be established this way nor to whom precisely she defers. She only has to intend to use the term in accordance

[57] In a case where subsequently, Batu is involved in a number of conversations with people who use the name to refer to the politician, things are intuitively less clear. But this need not be a problem for our view either, because it just means that we are not always or only inclined to use the name in accordance with the person from whom we first heard it. One would thus have to adjust the given description slightly to account for this. We could for instance say that in such a case, my use of the name is ambiguous or that it goes with its more established usage in my linguistic community.

with those from whom she learned it or alternatively (and in many cases more plausibly) to the relevant group of experts, whoever these may be. I conclude that Kripke's reservations against the current account which are based on an alleged threat of circularity are unfounded.

I think that a good case has been made that even where names are concerned or other uses of terms to which the arguments from Ignorance and Error can be applied, speakers do (implicitly) know something which can determine the term's reference: They know that a name 'N', if it refers, refers to the individual called 'N' by those from whom she acquired the name. Furthermore, our judgments about hypothetical scenarios seem perfectly compatible with the idea to understand the primary intensions of names deferentially. The view that with respect to each possible case, we judge our use of the name to concur with those from whom we (actually) acquired it is also well in line with Kripke's general insights concerning the importance of causal chains. (Unsurprisingly, one might add, since he himself derived his conclusions from considerations about hypothetical cases.) In cases of incomplete understanding, it also seems plausible that people do intend to defer to others and are thus disposed to look for the usage of other members in their linguistic community when they consider a hypothetical case as actual. Again, our evaluations of thought experiments which were invoked to establish externalism seem to confirm this – see for instance Burge's 'arthritis' case. However, I think our judgments about hypothetical cases raise some concerns in the context of deference which I will consider a bit later on. I will thus defer a fuller discussion of the question of whether deference to others is part of the a priori associations of names to 3.2.

3.1.2.3 Deferential concepts and apriority

If one incorporates the account of semantic deference just outlined into a two-dimensional framework, one can see that deferential concepts have some noteworthy features. When a term is used deferentially, this entails

that it is not semantically neutral, i.e. its primary and its secondary intension are not equivalent. Take Burge's 'arthritis' example to illustrate this: Both the doctor's and the patient's concept pick out an inflammation of joints. But while it may be a priori for the doctor that arthritis is an ailment of the joints, it is clearly not so for the patient. Moreover, the latter's concept rigidly refers to what the actual experts in her linguistic community pick out by 'arthritis'. If so, then although her and her doctor's concept differ concerning their primary intensions, their secondary intensions are equivalent. At the same time, the patient's concept shares its primary intension with that of her intrinsic duplicate, but not its secondary intension: The duplicate's concept rigidly picks out a disease which can affect a person's muscles.

This shows that whenever a concept is deferential, it is impossible to gain insight into the (metaphysical) essence of the category which it denotes by an analysis of the concept. The phenomenon of semantic deference can thus be highly relevant to issues concerning the epistemic value of conceptual analysis. Moreover, it can even be relevant to more general issues about the possibility of a priori knowledge. To see that latter point, suppose that there were deference involved in each of our utterances. In this case, the evaluation of any sentence with respect to a hypothetical scenario would be dependent on the existence of a linguistic community in this scenario. Thus, if we were to consider a completely empty scenario, a sentence like 'Language exists' would not have to be considered as false, but as lacking a truth-value. In fact, this sentence would only be false with respect to scenarios in which either 'language' or 'exists' have a different meaning than they actually have. And even worse, if each of our utterances was deferential, even a sentence like '2 + 2 = 4' would be false with respect to some worlds: If for example the number three was called '2' in a linguistic community (and vice versa), the sentence would be false with respect to this scenario. Thus, since we are concerned with worlds considered as

actual, '2 + 2 = 4' would not be a priori.[58] And since these considerations can be generalized straightforwardly, it follows that if every utterance was deferential, there could be no a priori true sentences, which would evidently be disastrous for the whole project of conceptual analysis.
The considerations just made suggest that in order to assess the prospects of conceptual analysis, it would be important to get a grip on how frequent deferential usage is. In the following, I will identify two methods for answering that question. As will be seen, a straightforward application of these methods reveals problems for the two-dimensionalist account of deference. I will tackle these problems in 3.2.

3.1.2.4 Two methods for detecting deferential concepts and two problems for two-dimensionalism

Burge's writings suggest that there is a strong link between deference and incomplete understanding. Indeed, this seems to be a reasonable assumption: Whenever a subject only has a partial grasp of a term, she has to defer to other people's usage. It is thus natural to try and approach the question how frequent deferential usage is by asking how frequent incomplete understanding is. One problem with this idea, however, is that it is quite unclear what complete understanding amounts to in the case of proper names and natural kind terms. It might mean to know the essential properties of the individual or kind in question. But this is obviously a very strong requirement. For firstly, it entails that in many cases, there is no-one within a linguistic community with a complete understanding of the term. And secondly, knowing the essential properties of Gödel or water seems unrelated to *linguistic* competence anyway. In the following, I will develop a notion of incomplete understanding which is applicable to all kinds of terms, based on the two-dimensional framework. I will then show that two-dimensionalism entails an intimate relation between deference and

[58] Of course, there could still be metaphysically necessary sentences. If '2' actually denotes the number two and '4' the number four etc., '2 + 2 = 4' would still express a necessary truth.

incomplete understanding, the nature of which, however, is problematic for epistemic reasons.

The following thesis about concept possession due to Chalmers and Jackson is already familiar from chapter 2:

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension.[59]

On a reading suggested by this thesis, insufficient understanding would mean that even given complete empirical information about the world (via a canonical description) and ideal rational capacities, a person is still not able to determine the concept's extension. Of course in this case, the person does not possess the concept in question according to (CJ). What we are looking for here, though, are cases in which a person does possess a concept despite of her incomplete understanding. Let me thus introduce the notion of *deference-dependency*:

(DD) A concept C is deference-dependent for a subject S in a linguistic community L iff i) S possesses C, and ii) when given complete empirical information about the world, S has to take into account information concerning the linguistic behavior of other members of L in order to determine C's extension.

In other words, whenever a subject's concept is deference-dependent, the subject would not possess the concept if she did not defer, since she would

[59] Incidentally, in the same paper they say that they want to restrict their discussion to non-deferential concepts. However, if one were to give up (CJ) for deferential concepts, this would evidently seriously weaken the scrutability thesis. Anyway, in their discussion of deference elsewhere, Chalmers and Jackson clearly assume that (CJ) holds for deferential concepts as well (cf. e.g. Chalmers 2002b; Jackson 1998b, 2003). Notice also that if one conceded that (CJ) does not apply to deferential concepts, which on the current reading includes names in their standard usage, this would undermine the thesis of metaphysical plenitude. I will say more about the relation between scrutability theses and metaphysical plenitude in chapter 6.

not be able to determine its reference with respect to every scenario. Due to (CJ), Chalmers and Jackson are committed to holding that every deference-dependent concept is deferential. But this commitment immediately raises a problem. Since deference is based on a subject's intention to use the relevant expression in accord with other speakers in her linguistic community, she has to know which of her concepts are deference-dependent. It is safe to say, however, that it is hopeless to demand from a subject to reliably judge whether the conditions specified in (DD) are satisfied. To see this, notice that (CJ) involves some idealizations which carry over to (DD): The subject is supposed to possess complete empirical information about the world, expressed in some limited vocabulary. And she is required to have 'unimpaired rational processes'. But needless to say, in real life it is not within the capacities of any human being to grasp such an amount of information, let alone process it. There is also no reason to expect that we are able to evaluate whether, if we were endowed with information of his kind and if we were furthermore able to handle this information, we would be in a position to determine the reference of a concept without looking at the usage of others. Notice that, for example, even the greatest living botanist cannot be sure to have a deference-independent understanding of the term 'elm', i.e. one which allows her to satisfy (CJ) without reliance on others' usage, because she cannot exclude the existence of an unknown tree species which is so similar to elms that she cannot tell them apart. Moreover, even if it were remotely plausible that we are able to judge whether a concept is deference-dependent or not, our disposition to defer does certainly not rely on considerations of the kind 'If someone provided me with complete physical, phenomenal and indexical information about the world, would I be able to determine the extension of my concept, given that I were ideally rational?'. One could of course insist that only a speaker's dispositions to judge under ideal conditions are really relevant. But this would mean that the two-dimensionalist account of deference is just unrelated to our linguistic and epistemic practice and thus to the data which it is (inter alia) supposed to explain.

There is thus a constitutive connection between incomplete understanding and deference in the two-dimensionalist framework, but it is of no practical value if one aims to determine which of our concepts are deferential. In the following, it will transpire that if one applies another natural method to reach that aim, the method of probing our intuitions about hypothetical scenarios, the result is even more worrisome.

Chalmers himself proposes the following heuristic: To evaluate whether some sentence S represents an epistemic possibility, one can ask: 'If it turns out that (some world) W is actual, will it be the case that S?' (cf. e.g. Chalmers 2004, 178). If one finds a scenario for which the answer is yes, then S is epistemically possible. Let me try to use this heuristic as a test to find out whether a given concept is deferential, by checking whether it is sensitive to information about the term's usage in one's linguistic community. Take Burge's 'arthritis' case first. Suppose the patient in the scenario were to evaluate the following sentence:

(1) If it turns out that 'arthritis' is used to denote an inflammation of joints by the experts in my linguistic community, then I don't have arthritis in my thigh.

Intuitively, this conditional would be acceptable for her, which confirms the idea that the patient's concept is deferential. But now consider another example, chosen arbitrarily:

(2) If it turns out that numbers are called 'colors' in my linguistic community, then 7 is a color.

This conditional seems to be just as acceptable, and if it is, the same considerations will plausibly go through for each of our concepts. The conclusion would be that all of our concepts are deferential, with disastrous consequences for the prospects of conceptual analysis – as outlined above.

A straightforward analysis of the notion of semantic deference within a two-dimensionalist framework thus leads into severe problems. In my view, both of the problems just pointed out have a common source. To address them, one has to be careful to distinguish between meaning in a public language and the content of mental states. In the following, I will

elaborate on the general relation between these notions and how they are dealt with by two-dimensionalism – or how I think they should be dealt with. The account I propose involves understanding (CJ) as a theory about the possession of mental concepts. It will, inter alia, provide the means to solve the problems just identified.

3.2 Linguistic meaning, mental content, and two-dimensionalism

Let me start with the second of the problems just identified. With the distinction between linguistic meaning and mental content in mind, one could try to disambiguate the sentences (1) and (2) invoked above. Take sentence (2) first:

(2*) If it turns out that numbers are called 'colors' in my linguistic community, then '7 is a color' is true (in the language spoken in my linguistic community).[60]

(2)** If it turns out that numbers are called 'colors' in my linguistic community, then my thought that 7 is not a color is false.[61]

If phrased this way, (2*) is just a metalinguistic statement. Sentences of this kind should be considered true by any speaker, regardless of whether her relevant concepts are deferential or not. This reading may at least partly account for the intuitive acceptability of (2). But it seems that the more relevant reading is (2**). And plausibly, this sentence will be considered as false by the average speaker. The term 'color' would then be an example of a non-deferential concept. Accordingly, if one understands Chalmers and

[60] It is not a trivial task to give an adequate formulation of this reading, because one may wonder which language is used in (2*). It obviously cannot be the language spoken in the patient's linguistic community. It is thus necessary to make an (additional) assumption to the effect that the language used in (2*) is English and that the language spoken in the (hypothetical) linguistic community mentioned is English*, which is identical to English with the only exception being the one described in (2*).

[61] The language used in (2**) is the language of the patient's thoughts, which we can, for the sake of argument, assume to be English.

Jackson's account of deference as a theory about mental concepts and the content of mental states, then there is a way to avoid the undesired consequences outlined above by restricting the domain of deferential concepts. If one applies this understanding to Burge's case, it again seems to yield the correct result:

(1*) If it turns out that 'arthritis' is used to denote an inflammation of joints by the experts in my linguistic community, then my belief that I have arthritis in my thigh is false.

Plausibly, this conditional would seem acceptable to the patient in Burge's thought experiment. The test introduced in the preceding section thus correctly indicates that the patient's 'arthritis' concept is deferential. It thus appears that Chalmers and Jackson's account of deference should best be taken as applying to mental content; and accordingly (CJ) should be read as a thesis about the possession of mental concepts.

It will be useful to consider what Chalmers and Jackson themselves take their two-dimensionalism to be a theory of. Jackson explicitly states that his account is about the meaning of words in a public language. The associated properties are connected with these words by (known) conventions. Thus construed, his brand of descriptivism cannot possibly apply to the content of mental states. For even if there is a language of thought, we do not know the words of Mentalese; and they certainly do not get their meaning by convention (cf. Jackson 2004, 275). Chalmers likewise holds that his two-dimensionalism is a theory of the meaning of linguistic expressions. However, he thinks that the two-dimensional account can also serve as a theory of mental content: Since primary intensions are shared between intrinsic duplicates, they can be taken to constitute a kind of narrow content (cf. Chalmers 2002c, 2003). Notably, he himself introduces his account of deferential concepts inter alia to counter Burge's argument, which is explicitly about belief states. Jackson in fact has a quite similar conception of mental content, except for the fact that he rejects the idea that mental states also have wide contents (cf. Jackson 2003).

In any case, it is possible to understand two-dimensionalism as a theory of mental content. Moreover, this understanding seems to be the appropriate one since we saw above that it is problematic to apply the linguistic reading to the two-dimensionalist account of deference.

This conclusion may be premature, however. On a charitable interpretation of Chalmers' and Jackson's views, (2**) is not relevant to their accounts. This is because they both concede that the primary intensions associated with an expression can vary between different tokens of a type; it is therefore at least possible that an associated primary intension does not reflect an expression's meaning in a public language. Accordingly, it is natural to think that more basically, primary intensions reflect a different aspect of meaning, which one might call utterance meaning. It is not entirely clear to me what theoretical work such a notion is supposed to do – one may well hold that a theory of linguistic meaning should capture precisely what is shared by different tokens of a type.[62] Let me nevertheless apply the reading just developed to the problem at hand to see whether it avoids the threat of ubiquitous deference. Here is the relevant version of (2):

(2*)** If it turns out that numbers are called 'colors' in my linguistic community, then my utterance '7 is a color' is true.

Intuitions about such sentences may of course diverge, but at least to me, this sentence seems acceptable. If it is, then the 'utterance meaning' interpretation of two-dimensionalism does not solve the problem, either.

Moreover, there are less sophisticated considerations which support the view that non-deferential utterances are at least quite rare. In Chalmers' own words, semantic deference means that a speaker "intend[s] to use the word for the same phenomenon for which others in the community use it" (Chalmers 2002a, 620). But is it not natural to suppose that, unless we are involved in a game or are particularly eccentric, we always want to accord with the practice of our linguistic community? After all, we usually do not

[62] For a defense of the idea that a semantic value which varies between tokens of a type can still deserve the label 'linguistic meaning' cf. Chalmers 2002b, 173ff.

invent the terms we use, we learn them.[63] I thus conclude that no account which treats two-dimensionalism primarily as an account of linguistic meaning can avoid the undesired consequence that deference is extremely widespread.

My proposal is therefore this: The primary intension associated with an expression token should first of all be considered as the content of the thought expressed by the utterer. This is in line with the fact that, as we have just seen, Chalmers' and Jackson's account of deference only works if understood as a theory about the contents of mental states. The proposal is also perfectly compatible with the fact that primary intensions can vary between different tokens of a type – it is even trivial that different speakers can use the same expression type to express different thoughts.[64] Moreover, there are independent reasons for holding this view: The primary intension reflects the properties which the speaker associates with the relevant expression. It is hard to see how it could fail to represent the thought which the speaker tried to express by uttering the expression.

One residual worry is that the proposal at hand cannot really give an adequate account of the phenomenon of deference with its distinctly social dimension, which is connected with a shared linguistic practice. In particular, it appears that the connection between deference and incomplete understanding of a word in a public language gets lost on my interpretation. However, I do not deny that deference has such a social dimension, and I think there clearly is a connection between deference and incomplete understanding: In general, when a subject believes to have an incomplete grasp of a term, she will be disposed to defer to others' usage, precisely in order to bring her use of the term into accordance with her linguistic community. Nevertheless, on the current view there is no *necessary*

[63] De Brabanter et al. probably have something like this in mind when they say that there is a deferential component which is involved in every communicative act, which they call 'default deference' (cf. De Brabanter et al. 2005).

[64] In Chalmers 2006, 96f., Chalmers also mentions that primary intensions can be taken to represent the content of the thought expressed by an utterance.

connection between deference and incomplete understanding. In the following, it will transpire that this is in fact a virtue.

Let me now turn to the other of the two problems identified in the preceding section. The problem was that Chalmers and Jackson's thesis about concept possession (CJ) seems to imply that whenever a subject has incomplete mastery of a term, or more precisely whenever a term is deference-dependent for her, she has to defer to others. As we saw, this requires her to know which of the terms she uses are deference-dependent, which is simply impossible. In practice, the epistemic inaccessibility of deference-dependency could become manifest in the following way: A speaker who intuitively possesses a concept of a public language could have no disposition to defer, in spite of the fact that her concept is deference-dependent, because she believes to have a sufficient grasp of the term's meaning.[65] Let me illustrate this by means of a somewhat extreme case: Let us suppose there is a person who, by all appearances, masters a certain term, without being in any way disposed to defer. In all of her applications of the relevant term throughout her life, she uses it in complete accordance with her linguistic community. Suppose further that somewhere in a remote place of the universe, there exist some entities which the subject would misclassify if she ever came across them. Therefore, if that person were given complete empirical information about the world and ideal cognitive capacities, she would assign a 'deviant' extension to the term: The assigned extension could be larger, maybe because the subject is unaware of some subtle criterion for the term's applicability – which would make this a case of 'ignorance' – or it could be slightly different, in a case of 'error'. In order to preserve the idea that (CJ) applies to possession of a

[65] What about the opposite direction: Can a subject be inclined to defer, although she has a sufficient grasp of the term? This is certainly possible, but it need not be a problem. If that person were disposed to revise her understanding in hypothetical circumstances in favor of some experts' judgments, those revisable associations would arguably not count as a priori associations.

concept of a public language, Chalmers and Jackson would have to deny that in such cases, the person masters the term in question, contrary to appearances. But of course, this verdict would have to be applied to many other cases as well: Because there is all the difference in the world between a person's disposition to defer and our disposition to take her to master a term on the one side and (dis)satisfaction of (DD) on the other, our judgments regarding this matter would have to be deemed to a large extent unreliable. Moreover, this unreliability would not only concern our prima facie judgments: Satisfaction of (DD) is nothing which is remotely epistemically accessible under realistic conditions, and thus the question of whether or not a person masters a certain term would frequently not be decidable, either, even under careful scrutiny. In any case, it seems to me that the whole idea of abandoning our pre-theoretical judgments on the basis of highly sophisticated theoretical commitments, such as (CJ), is misguided. Plausibly, whether or not a person masters a term is a question from a social context: To master a term is to successfully participate in a social practice. This question is quite obviously completely independent of the question of whether she satisfies (DD).

My proposed solution for this problem is, again, to move away from public language. One should deny that (CJ) holds for mastery of a term and take it to express a condition for the possession of a mental concept. If taken this way, (CJ) does not have to be considered as a thesis which goes beyond the general commitments of two-dimensionalism. It just says that whatever properties one associates with a word, and thus whatever extension one is disposed to assign to it if given sufficient information about the world (and with respect to various scenarios), constitutes the content of one's relevant concept. This might (again) raise the question what it takes for two people to share a concept. If one individuates concepts as fine-grained as was just suggested, then concepts will presumably rarely be shared. But as I already mentioned above, two-dimensionalism in itself is not committed to any particular thesis here. One *can* individuate concepts in a very fine-grained way. However, with respect to at least some kinds of concepts, it does make sense to adopt a more coarse-grained understanding by stating, for

instance, that for two people to share a concept, the associated primary intensions only need to have a certain degree of similarity.

Could one apply a similar move to argue that there is a close connection between (CJ) and mastery of a term after all – by holding that if a person masters a term, she assigns an extension from within a certain range to it? I do not think such a maneuver would work. For a start, it does not seem as if this is the right way to adjust the condition: A person's understanding of a term could only be slightly deviant, and yet, due to the nature of the actual world, she assigns a very different extension to it on the basis of her understanding. So one should presumably rather adopt a slightly different criterion by saying that if a person masters a term, she is disposed to assign an extension from within a certain range to it with respect to the set of scenarios (or maybe some subset of it) – i.e. in effect if she associates a primary intension from within a certain range to it. But firstly, one would then have to postulate that expressions are also associated with a community meaning, in order to determine the domain in which the speaker's primary intension has to be located. This would already mean to grant that there is at least an important kind of linguistic meaning which is not captured by the properties associated by individual speakers. And secondly, one still would not get the right result in many cases: In Burge's thought experiment, for instance, the patient's primary intension is very different from that of her doctor; Putnam's primary intension with respect to 'elm' apparently differs strikingly from whatever one can sensibly assume to be the term's community meaning. But the starting point of the considerations on deference and the division of linguistic labor was precisely that intuitively, Putnam and the patient do master the terms in question (or alternatively, that they possess the concepts).

Interpreting two-dimensionalism primarily not as an account of the meaning of linguistic expressions, but rather as one of the contents of the thoughts expressed by these expressions thus offers a number of advantages. Given this reading, it is unproblematic to assign varying semantic values to different tokens of an expression type. The current view also entails a more plausible understanding of Chalmers and Jackson's

account of deference, which permits us to restrict the domain of deferential utterances in a sensible way. And finally, it suggests reading (CJ) as a thesis about mental concepts, which again permits us to avoid severe problems.

There is a worry that in my interpretation of two-dimensionalism, it is hard to see how two-dimensional semantics could help to underpin conceptual analysis as a philosophical method. There seems to be little point in conceptual analysis if it can only reveal a person's concepts, which may be quite idiosyncratic. For precisely this reason, Jackson holds that the aim of conceptual analysis is to analyze the meanings of our folk terms (cf. e.g. Jackson 1998a, 30ff.). But in this case, it seems that the adherent of conceptual analysis needs a theory of linguistic meaning, not one of mental content. Against this objection, let me first note that I do not want to deny that primary and secondary intensions can be used to model linguistic meaning. In cases where, for example, the primary intension associated with an expression is widely shared, or at least approximately shared, among speakers of a linguistic community, it seems reasonable to say that this primary intension is associated with the expression by convention – and thus that it represents (an aspect of) the expression type's linguistic meaning. So I do not have to abandon the idea that public concepts, or folk terms, are the target of conceptual analysis. I just think that one should clearly distinguish two aspects of such an inquiry: What the conceptual analyst scrutinizes in the first place is the primary intension which she associates with an expression – i.e., her own concept. This is the a priori part of the inquiry. Now if it turns out that the properties she associates with the expression in question are commonly associated with it, then she has in fact analyzed the folk term.[66] But she cannot know a priori that she has done so. On this understanding, conceptual analysis has an essential a posteriori component. One should note, however, that this is not the result of my proposal to understand two-dimensionalism as an account of the thoughts expressed by linguistic expressions. If one takes conceptual

[66] Cf. Goldman 2007 for a very similar proposal.

analysis to reveal the meanings of our folk terms, then it cannot be completely a priori on any reading: Although, for instance, one may be a priori justified in believing that bachelors are unmarried men – in virtue of one's linguistic competence – one can never be a priori justified in believing that 'bachelors are married males' is a true English sentence. Notice that this does not necessarily imply that one has to conduct intricate empirical investigations to find this out. At least in many cases, we can be quite confident that our understanding of the expression is not idiosyncratic, due to our experiences as language users.

In this chapter, I discussed the epistemic arguments brought up by Kripke and others, whose conclusion contradicts the idea that names and natural kind terms have primary intensions. I tried to undermine these arguments by giving at least a rough outline of how the primary intensions of such expressions are to be conceived, in a way which is in line with our judgments about hypothetical cases. Following Lewis' idea, I suggested that the primary intension of 'water' should be taken to comprise a theory about a particular substance. It is natural to think that other natural kind terms can be treated in a similar way.
The primary intensions associated with names have to be construed quite differently, however. With respect to them, I defended an idea which goes back to Strawson and which was then taken up by Chalmers and Jackson who applied it to the two-dimensional framework: They hold that the primary intensions of names and generally of incompletely understood terms are deferential, i.e. the properties associated with such expressions include deference to the usage of the expression by other speakers in the linguistic community. However, it transpired that such deferential concepts raise a couple of problems. My solution to these problems was to understand two-dimensionalism as a theory of the contents of the thoughts expressed with linguistic utterances, which offers a number of advantages even beyond issues related to deference. Conceptual analysis can then be understood as a two-step process. The first of these steps involves an analysis of the primary intension associated by a subject, while in the

second step one has to check whether this primary intension is shared by other speakers in the subject's linguistic community.

4 Primary intensions, defining the subject, and communication

In this chapter, I will discuss two arguments for the existence of primary intensions which are central to Jackson's defense of descriptivism. Let me give a very rough sketch of what these arguments aim to establish:

Both of the arguments are supposed to show that primary intensions, i.e. associated properties, play an indispensable role in our linguistic and epistemic practice. The first one is about the role of primary intensions in defining the subject; the second one concerns their role in communication. As will be seen, each of the arguments aims to establish two theses, a semantic and an epistemic one. In the case of the first argument, the semantic thesis is that primary intensions define the subject in the sense that changing a term's primary intension amounts to changing the subject. According to the epistemic thesis, we need primary intensions in philosophical as well as in empirical inquiry to determine our subject matter. The second argument aims to show, firstly, that primary intensions are transmitted from speaker to hearer in communicative acts – this is the semantic thesis about communication. Secondly, it is supposed to establish the thesis that successful communication would not be possible if the linguistic expressions involved were not associated with primary intensions – this is the epistemic thesis about communication. It will transpire that Jackson's case for the relevance of primary intensions both in defining the subject and in communication needs to be qualified. I will therefore develop alternative accounts of the role of primary intensions in these areas.

The semantic theses seem more ambitious than the epistemic ones; at least on one reading, the epistemic theses are even entailed by the semantic theses. It is therefore plausible that the epistemic theses will be easier to defend. Eventually, I will argue that both of the semantic theses are strictly speaking false and I will develop an account which I think is more adequate

than Jackson's. At the same time, this account will leave the importance of primary intensions in our linguistic/epistemic practice largely intact.
One should note, however, that these issues are highly dependent on which kinds of terms are being considered. If one restricts the two theses to semantically neutral terms, then they seem hard to reject: With respect to such terms, the semantic theses are close to trivially true and the epistemic theses are at least very plausible. But nothing that has been said in the preceding chapters bears on the question of how many of our terms are of that kind. In fact, it has not even been shown that there are any semantically neutral terms. These questions will be addressed in chapter 5. In any case, Jackson's claims about communication and defining the subject have to be taken to apply to all kinds of terms if they are to support the two-dimensionalist thesis that every linguistic expression has a primary intension. Thus, since two-dimensionalists accept that there are epistemically opaque terms, i.e. terms whose secondary application conditions are not a priori accessible, these will have to be dealt with as well.

4.1 Defining the subject

In *From Metaphysics to Ethics* (Jackson 1998a), Jackson argues that primary intensions play a crucial role in philosophical as well as in any kind of empirical inquiry because they define the subject. And he makes a couple of remarks which indicate that he wants this to be understood literally, for example in his discussion of William Lycan's view. Lycan believes that most of the properties we associate with belief can turn out to be wrong. Here is Jackson's response:

> I of course hold against Lycan that if we give up too many of the properties common sense associates with belief as represented by the folk theory of belief, we do indeed change the subject, and are no longer talking about belief. (Jackson 1998a, 38)

This quote suggests that Jackson takes the associated properties to define the subject in a strict semantic sense, such that at least a substantial change

in the primary intension amounts to changing the subject. Beyond that, Jackson advocates an explicitly epistemic thesis. He thinks that we need the associated properties in order to determine or identify the subject matter in our epistemic practice. Jackson illustrates this idea by means of the following analogy: When the bounty hunter searches for a fugitive, he would be lost if he did not have something which tells him whom he is looking for, or at what point he has found the person he is looking for. For this reason, he has to rely on a handbill (cf. Jackson 1998a, 30). Here, the way the handbill represents the fugitive is supposed to stand for the properties associated with a term.

As I mentioned above, the semantic thesis that speaker associations define the subject is plausibly stronger than the epistemic thesis that we need these associations in order to determine the subject. One could therefore endorse the epistemic thesis without endorsing the semantic one, and this is in fact the view which I will eventually argue for. Nevertheless, the two theses seem to go together very well. In particular, the semantic thesis nicely underpins the epistemic thesis: If the semantic thesis is correct, then it is very natural to assume that primary intensions give us epistemic access to our subject matter. Generally speaking, it is an attractive idea not to treat the semantics independently of the epistemology – and of course, this is one of the key motivations behind two-dimensionalism. However, as I will argue, Jackson's view fails as a general model of what defines the subject. The account I propose, while being less straightforward than Jackson's, sustains a close connection between the epistemology and the semantics of determining the subject and therefore, I hope, preserves much of the appeal of Jackson's model.

The structure of my discussion on the role of primary intensions in determining the subject matter will be as follows: I will start by saying a few words about the importance of the question of what defines the subject in philosophy. Jackson's theses will prove to be attractive not only for the purposes of a defense of two-dimensionalism, but also because they promise to provide epistemically accessible criteria for changes of subject.

I will then turn to a discussion of the epistemic thesis. I should note, however, that I will develop my own version of the thesis by arguing that if one accepts a specific reading of the claim that we have the (conditional) ability to determine our subject matter (which is the by now familiar thesis (CJ)), then it is hard to avoid the conclusion that linguistic expressions are associated with primary intensions. My main goal in this section, i.e. 4.1, is to establish that version of the epistemic thesis. My argument will proceed in two steps: I will first defend the conditional claim that if one assumes that we have the ability in question, then one should also accept that this ability is applicable to other worlds considered as actual as well, by rebutting Laura Schroeter's alternative proposal on what defines the subject. The second step of the argument will draw on considerations due to Chalmers and Jackson. Its aim will be to show that, roughly speaking, the determinability of the subject matter with respect to all scenarios entails its a priori determinability with respect to all scenarios. As I will explain, the latter thesis is equivalent to the thesis that linguistic expressions have primary intensions.
In the final part of the section, I will be concerned with Jackson's semantic thesis that primary intensions define the subject matter. I will argue that the thesis ultimately fails and develop an alternative account of the relevance of primary intensions in ensuring (or determining) constancy in subject matter.

I mentioned above that in his response to Lycan, Jackson claims that if one ignores the properties associated with the term 'belief' or changes too many of them, one may thereby change the subject and no longer be concerned with belief. In philosophical debates the charge that someone has changed the subject is not uncommon. Many have argued that Quine, in his so-called 'naturalized epistemology', is no longer doing epistemology (cf. e.g. Kim 1988). Or take another example from epistemology: Especially in the early discussions on externalism about justification, internalists often claimed that their opponents have changed the subject, because they (the externalists) were no longer concerned with justification (cf. e.g. BonJour

1985; Chisholm 1989; Lehrer 1990). Adherents of a libertarian account of free will have occasionally accused compatibilists of changing the subject (cf. e.g. Bishop Bramhall's objections to John Locke, cf. Chapell 1999). In the philosophy of science, there was a huge debate about theory change: Philosophers such as Thomas Kuhn and Paul Feyerabend claimed that when one scientific theory (or 'paradigm') is replaced by another, the two theories are often incommensurable (cf. Kuhn 1962; Feyerabend 1962). Now, incommensurability seems to be the result of a particularly bad kind of change in subject matter: Not only do the theories differ fundamentally in their contents, there is also no way to translate statements made in their respective vocabularies and thus to compare them on a rational basis.

In most cases, changes of subject are of course completely unproblematic: In general, it is neither particularly hard to detect and nor is it in any way blameworthy when someone changes the subject. But as the examples just given suggest, there are exceptions. Those cases which are potentially problematic and which can evoke disagreement regarding the question of whether or not the subject has been changed often follow the following schema: The parties involved start with a question which they phrase by, for example, 'What is X?'. After a while, the question they are after is still phrased by 'What is X?', but nevertheless, it is no longer the same question. Presumably, such a case will have to involve some kind of change in the meaning of at least one of the key terms involved.

Against this background, the motivation for Jackson's view becomes apparent: Since primary intensions are a central semantic value, changing the primary intension of an expression entails meaning change, and thus a change of subject. What makes this view even more attractive is the fact that, since primary intensions are accessible a priori, it promises to offer epistemic access to changes of subject. Accordingly, Jackson's account seems to provide an epistemic criterion on the basis of which disputes of this sort can, at least in principle, be settled.

But of course, since this presupposes that there are primary intensions, it cannot be an argument for their existence. Moreover, even if one grants that expressions are associated with primary intensions, one is not thereby

committed to holding that they are what defines the subject. In order to evaluate Jackson's semantic thesis, it would thus be useful to identify independent criteria for changes of subject, or for meaning change. The most obvious criterion should be a (substantial) change in extension.[67] I think a (substantial) change in the secondary intension is also plausibly a sufficient condition for a change of subject. Take a typical case of philosophical inquiry where one is dealing with a question of the form 'What is X?': When the secondary intension of that sentence changes, this means, on both an externalist and a two-dimensionalist account, that the representational content of the key question has changed. It is hard to see how this can occur without a change in subject matter or even without meaning change.[68]

Jackson's remarks about the (epistemic) determination of the subject, including his story about the bounty hunter, indicate that he takes primary intensions to be essential for any kind of inquiry. The underlying idea is that the associated properties are what enables a subject to make judgments in response to evidence. For if there were no such associations, then we would be like a bounty hunter without a handbill – we would not even know when we have found what we are looking for. The associated properties thus enable us to determine our subject matter.

There is a natural route from this claim to the semantic thesis that primary intensions define the subject. For it is plausible that if the associated properties by means of which we determine the subject matter change radically, the subject matter will change as well.

Jackson's account seems intuitively appealing. Nevertheless, the considerations he invokes to support his theses do not yield a completely compelling argument. Firstly, even though it is clear that we need some basis for determining our subject matter, it is less obvious why nothing short of primary intensions can account for this ability. Why, for example,

[67] More precisely, this should be understood as a change in the atemporal extension, or alternatively in the extension with respect to a specific time.

[68] There are arguably contexts in which sameness of extension is sufficient for sameness of the subject, in particular in cases where the subject is an individual.

should empirical associations not be sufficient as well? Secondly, even if primary intensions are indeed required to account for the ability to determine the subject matter, then it has still not been shown that changing the associated properties amounts to changing the subject. But let me defer that latter issue until later and start with a discussion of (my version of) the epistemic thesis.

4.1.1 A case for the epistemic thesis

I think the key premise behind the epistemic thesis, namely that we are able to determine the subject matter if we are provided with relevant evidence, can naturally be expressed by the already familiar thesis (CJ):

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension.

On the face of it, (CJ) seems to be an attractive thesis. One can argue about the kind and the scope of the required information or believe that there are exceptions to the thesis, but in general, one should grant that we are able to make such judgments in the face of evidence, if one wants to avoid skepticism. Relatedly, when we are involved in an empirical investigation, we should know (at least implicitly) given which evidence some thesis or other would turn out to be true, or under which circumstances the inquiry has been successful. I will give a much more explicit defense of (CJ) itself in chapter 6. Here, I will assume that we have this ability, which is thus to be understood as a version of the assumption implicit in Jackson's account that we have the ability to determine our subject matter. The claim I want to defend is that primary intensions are needed to account for the ability in question.

In itself, (CJ) is too weak to establish the existence of primary intensions. Opponents of conceptual analysis can accept that we have such an ability, while denying that it is due to mere conceptual competence or some other a priori faculty. And in fact, as we will see (CJ) is endorsed by a number of

decided critics of two-dimensionalism. To see how it can nevertheless be used as a premise in an argument for the existence of primary intensions, consider the following two stronger theses:

(CJ+) If a subject possesses a concept and has unimpaired rational processes, then sufficient information *about any given scenario* puts a subject in a position to identify the concept's extension *with respect to that scenario*.

(CJ++) If a subject possesses a concept and has unimpaired rational processes, then sufficient information about any given scenario puts a subject in a position to identify the concept's extension with respect to that scenario *a priori*.

Given that one accepts (CJ), there is at least a prima facie case to the conclusion that (CJ++) holds as well:

Firstly, it seems natural to assume that it makes no difference whether we are confronted with actual information about the world or with merely hypothetical information about a scenario which is just considered as actual. On the face of it, there is no reason to think that the actual world is somehow special in this respect.

Secondly, once (CJ+) is established, it seems plausible that these conditional judgments from the information about the scenario to the extension of the concept in question are a priori: Given that the information in the antecedents of the conditionals is complete, there simply is no additional empirical information which could justify the judgments. Now recall that a primary intension is defined as an a priori accessible semantic value which assigns an extension to an expression with respect to every world considered as actual, where these worlds are presented via canonical descriptions. Grasping an expression's primary intension is thus being able to determine its extension with respect to each world considered as actual a priori, if provided with the world's canonical description. And consequently, embracing (CJ++) amounts to accepting the claim that to possess a concept is to grasp its associated primary intension. Notice that (CJ) to (CJ++) are versions of Chalmers' scrutability thesis – more

specifically of the scrutability of reference – which was introduced in chapter 2. There, I already pointed out that there are close connections between scrutability and the two-dimensionalist framework. As we just saw, (CJ++) is exactly the version required to establish two-dimensionalism.

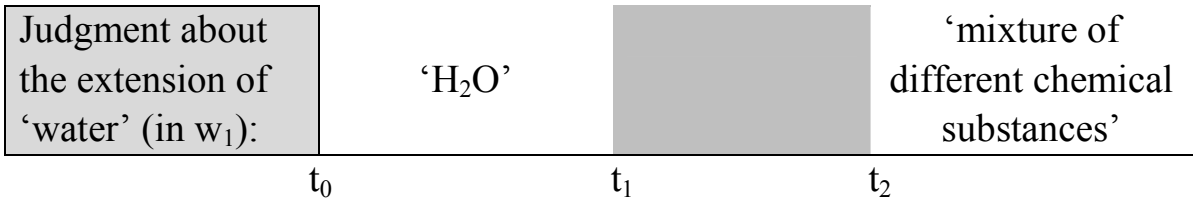
Accordingly, an opponent of two-dimensionalism and of the epistemic thesis who accepts (CJ) has to deny that it entails (CJ++), despite its prima facie plausibility. In fact, a number of critics have tried to undermine the transition in question. It seems that most of them have focused on the step from (CJ+) to (CJ++) (cf. Block & Stalnaker 1999; Polger 2008). I will discuss their objections later on. I believe that the step from (CJ) to (CJ+) deserves more attention than it has received so far. In the following, I will therefore discuss Laura Schroeter's arguments in some detail, who (to my knowledge) has to date been the only one to try to rebut the idea that (CJ) entails (CJ+). Schroeter's argumentation rests upon an alternative account of conceptual competence, which at the same time offers a view of what defines the subject which is quite different from Jackson's.

4.1.1.1 From (CJ) to (CJ+) – Schroeter's improv model

Schroeter calls Chalmers and Jackson's model of conceptual competence the 'template model' (cf. Schroeter 2006). In the case of the bounty hunter, the handbill serves as a template. This template is determinate, in the sense that changing the template means to change the subject. But according to Schroeter, this model does not adequately reflect what she calls our 'representational practice' because it is not sufficiently flexible. She believes that we are often required to revise our original conception, and thus our template, in the course of our investigations. Take for instance Jackson's bounty hunter: During his search, he could find out that the person he is after is not red-haired as it is depicted on the handbill, but rather blond, that it is not a man but a woman, etc. It could even turn out that the sought person did not rob a bank at all or that there was not even a bank robbery (cf. Schroeter 2006, 571). Schroeter thus thinks that it can be

rational to revise any of the properties associated with a concept. And crucially, she believes that this revision need not be accompanied by conceptual change, or a change of subject. This raises the question of how continuity of the subject is ensured on her account, which she calls the 'improv model' – the model of rational improvisation. She gives the following two conditions for continuity of subject matter: Firstly, each change in our conception must be based on rational reasons – this will mostly be due to newly acquired empirical evidence. Secondly, there must be something about the former conception which explains the success of our representational practice connected with it and which is thus preserved in the new conception (cf. Schroeter 2006, 573).

Here is a thought experiment which Schroeter herself invokes to illustrate her model: Suppose that we initially assume that water is a natural kind. At some point, we find out that the liquid in our ponds and creeks is really just a mixture of many different substances. The only pattern which connects our different uses of the term 'water', or maybe rather the most natural pattern which does so, is formed by certain culinary properties. After a while, we thus change our conception of water to something like 'basic drink'. Figure 2 illustrates our judgments about the extension of 'water' at different times in Schroeter's thought experiment. (I added one small detail to Schroeter's scenario, namely that at t_0, we do not only believe that water is a natural kind, but we already have a theory about the molecular structure of water: We think that water is H_2O.)



(Figure 2)

w_1 (the world inhabited by the subjects in the hypothetical scenario): the dominant clear, drinkable, … liquid in the oceans and lakes is an inhomogeneous mixture of different chemical substances

t_1 marks the discovery that the liquid in our rivers and lakes is actually a mixture of many different chemical substances. Subsequently, our conception of water changes. At t_2, it has changed to 'basic drink'. By now, we may have even forgotten that we once thought that water is a natural kind – according to Schroeter, it is not a rational requirement to keep track of our former beliefs (cf. Schroeter 2006, 579). Based on our new conception, we now judge that water is a mixture of various chemical substances. This seems to confirm that we are able to determine the extension of 'water' if given sufficient empirical information, in accordance with (CJ). But recall that Schroeter rejects (CJ+). In particular, she argues that we do not have the required ability with respect to other worlds which we consider as actual, if they differ too much from the 'real' actual world in relevant details. The second part of Schroeter's thought experiment is supposed to show why:

Starting from the hypothetical scenario just described, let us assume that the people in the thought experiment consider w_2 as actual, where the watery stuff is H_2O. Let them determine the extension of 'water' with respect to this scenario at t_0. The judgment seems clear: Based on their conception of water according to which it is a natural kind, they will say that water is H_2O.

As I said before, at t_1 these people discover that the liquid in their actual environment is not a natural kind, and thus their conception changes

between t_1 and t_2. If they now have another look at w_2 and if, as Schroeter assumes, they base their judgment on their newly-formed conception ('basic drink'), their judgment about the extension of 'water' will differ from that at t_0: Some isotopes of H_2O are not really drinkable (for example heavy water), non-liquid H_2O is not drinkable at all, pure H_2O does not taste very well and does not contain electrolytes which are important for our nutrition, etc. And thus, they will now conclude that it is only specific isotopes of H_2O with specific kinds of admixtures (in a specific aggregate phase) which fall under the extension of 'water'.

Consequently, and importantly, empirical discoveries (about the actual world) can change our judgments about other worlds considered as actual. And thus, (CJ+) is false according to Schroeter – we are not able to evaluate the extension of a term with respect to other worlds considered as actual, or only when the scenario in question does not differ from our world in relevant details.

As a side note: The extension of 'water' in w_2 when considered as actual does not change between t_0 and t_2. Rather, one of these judgments is false – but which one? It is not really clear to me what one should say here (cf. figure 3).[69]

[69] In personal conversation, Schroeter proposed that we should either say that the first judgment, at t_0, is false or that the extension of 'water' with respect to this scenario is indeterminate.

Judgment about the extension of 'water' (in w_1):	'H_2O' (F)		'mixture of different chemical substances'
Judgment about the extension of 'water' in w_2 considered as actual:	'H_2O' (?)		'specific isotopes of H_2O with certain admixtures' (?)
	t_0	t_1	t_2

(Figure 3)

w_1 (the world inhabited by the subjects in the hypothetical scenario): the dominant clear, drinkable, … liquid in the oceans and lakes is an inhomogeneous mixture of different chemical substances
w_2: the dominant clear, drinkable, … liquid in the oceans and lakes is H_2O

Unlike Chalmers and Jackson, Schroeter does not believe that concepts have a priori implications, even less primary intensions. But still, she wants to retain the idea that they guide our judgments in the face of empirical evidence, i.e. she wants to retain (CJ). So, on her view there are speaker associations which play an important role in our epistemic practice (or our 'representational practices'), but they are themselves empirically defeasible.
Chalmers and Jackson would say that what Schroeter calls our conception, namely the associated properties, constitutes the content of the concept in question. Therefore, any change in the conception entails conceptual change and at least in Jackson's view also a change of subject. However, Schroeter denies that her thought experiment involves such a change:

> All I'm asking you to do in this example is to take your new empirically informed beliefs about water and feed them back into the hypothetical reasoning process. […] Provided that your deliberation is fully rational, and provided that you really do know the relevant facts about your real-world environment, this methodology should guarantee that there is no change of meaning involved […]. (Schroeter 2006, 584)

So, according to Schroeter, what happens in the hypothetical case just described is that we simply refine our representational practice in the light of new empirical evidence, moreover in a seemingly completely rational way. One could of course insist that the case involves conceptual change, but Schroeter thinks there is no independent reason to do so. In her view, if a change in the conception is based on rational reasons, this even ensures conceptual constancy.[70] I think it is far from obvious that this latter point is correct. Can it not be rational to change the meaning of a term or to change the subject? Take for instance the bounty hunter: There could be plenty of good reasons for him, in the course of his investigations, to decide to search for another person instead: Maybe the man he was after is innocent, or he will bring the bounty hunter no reward, or maybe the handbill is a fraud and there is no such man. Schroeter insists that we have to distinguish such cases where the change in the conception is only pragmatically justified from ones where it is epistemically justified – which would mean that the bounty hunter merely changes his beliefs about the person he is after. But on what basis can this distinction be made within Schroeter's model? In fact, I think it is anything but clear that in her own story about the bounty hunter's search, his target remains the same. In the following, I will back up the suspicion that the rational belief revisions which she has in mind do not ensure constancy of subject matter. In particular, I will show that many cases which are in line with Schroeter's model should be considered as involving a change of subject.

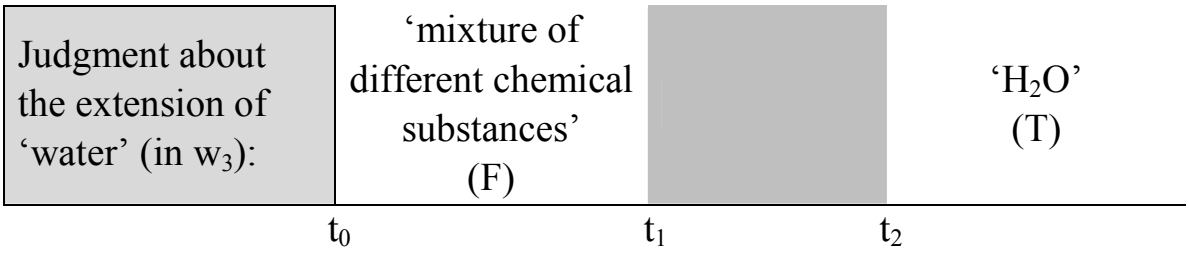
I just mentioned that Schroeter thinks there is no independent reason for saying that changes in the conception which conform to her improv model should be equated with conceptual change. Since I argued above that a change in extension would surely qualify as an independent reason for assuming conceptual change or a change of subject, my strategy will be to

[70] Above I mentioned that Schroeter gives an additional condition for conceptual constancy, namely that there must be some aspect of the former conception which explains the success of our representational practice connected with the concept. This requirement should probably be subsumed under the rationality condition.

show that Schroeter's model of conceptual continuity cannot rule out cases in which the extension of the concept in question changes.

It is plausible that in the scenario which Schroeter invokes, there is indeed no change in the extension. There is at least good reason to assume that 'water' referred to a mixture of chemical substances from the start – recall that the original judgment about the nature of water was false, anyway (cf. figure 3). So there is no problem apparent for Schroeter's model with respect to her own scenario. In the following, however, I will develop a thought experiment which does raise a problem for her account.

My thought experiment will, so to say, turn Schroeter's scenario upside down. To elaborate: Let us assume that initially (at t_0) people think, for whatever reason, that the liquid in their rivers and lakes is no natural kind. The conception they associate with 'water' is 'basic drink'. But then, at t_1, they find out that they were wrong – the stuff in their rivers and lakes which they call 'water' is a natural kind, namely H_2O. In the light of this empirical finding, they change their conception accordingly. Finally, at t_2, they believe that the term 'water' applies only to the natural kind samples of which can be found in their rivers and lakes. Figure 4 presents a subject's judgments about the extension of 'water' within that thought experiment.



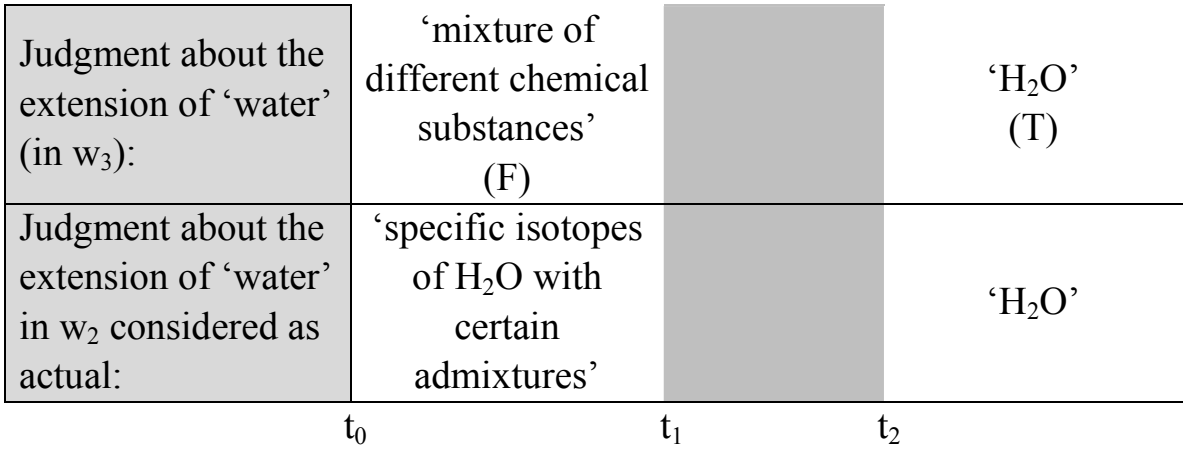
(Figure 4)

w_3 (the world inhabited by the subjects in the hypothetical scenario): the dominant clear, drinkable, … liquid in the oceans and lakes is H_2O

As can be seen, at t_0 people judge that the (alleged) culinary kind called 'water' is a mixture of different chemical substances. But at t_2, after they have learned that the stuff in their rivers and lakes is a natural kind after all and after their conception has changed accordingly, they will judge that all and only H_2O is in the extension of 'water'.
Now let us further suppose that these people evaluate the extension of 'water' with respect to a hypothetical scenario, which they consider as actual, at different times. In this hypothetical scenario, w_2, the dominant liquid in the rivers and lakes is H_2O. So what will their judgment about the extension of 'water' with respect to w_2 be at t_0, according to Schroeter's model? In her own scenario, each of a subject's judgments about a world considered as actual is based on her conception at that time. Accordingly, one should assume that the subjects in the present scenario judge that only specific isotopes of H_2O with certain admixtures fall under the extension of 'water', based on their conception of water as 'basic drink'.[71] By the same token, when they reconsider this scenario at t_2 when the conception they associate with water includes that it is a natural kind, they will judge that all (aggregations of) H_2O molecules in w_2 are in the extension of 'water'. Figure 5 illustrates a subject's judgments about the extension of 'water'

[71] In fact, this judgment is completely analogous to the one about w_2 at t_2 in Schroeter's scenario. I will return to this point below.

with respect to her own world, i.e. w_3, and with respect to w_2 considered as actual at different times.



(Figure 5)

w_3 (the world inhabited by the subjects in the hypothetical scenario): the dominant clear, drinkable, … liquid in the oceans and lakes is H_2O
w_2: the dominant clear, drinkable, … liquid in the oceans and lakes is H_2O

Thus, just like in Schroeter's thought experiment, the judgment about a world which is considered as actual changes over time. And in both cases, this is not because the respective subject initially misses some relevant detail about that world. Rather, it is because of changes in her beliefs about the (her) actual world. But in my thought experiment, this immediately raises a problem. This is because Schroeter explicitly accepts (CJ), the thesis that a subject can determine the extension of her concepts if provided with sufficient empirical information about the (actual) world. For w_2 is, at least in all relevant respects, identical with the (actual) world of the subject. So, due to (CJ), both of her judgments about the extension of 'water' with respect to w_2 – the one at t_0 and the one at t_2 – would have to be considered as true: Both are in fact judgments about the actual world of the subject. But this would imply that the extension of 'water' changed between t_0 and t_2. Thus, although the example I gave does seem to conform to Schroeter's

model since the subject simply changes the conception associated with ‘water’ in the light of empirical findings, it clearly involves conceptual change and a change of subject.

In the following, I will discuss a number of possible responses on behalf of the improv model to the counterexample I just gave. It will emerge that all of them ultimately fail. I will then make a more general diagnosis of the shortcomings of Schroeter’s account.

One way to resist my conclusion is to deny that in the case I just sketched, the subject’s judgment about the extension of ‘water’ with respect to w_2 at t_0 is ideally rational. The judgment certainly does mirror the subject’s conception of water at that time. However, one could argue that if she were to reflect more thoroughly about what she should say if w_2 was actual, she would conclude that it would turn out that water is simply the natural kind H_2O.

The problem with this response becomes apparent when one reconsiders Schroeter’s own scenario. Her argument against thesis (CJ+) is based on the idea that the subject’s judgment about the extension of ‘water’ in w_2 changes between t_0 and t_2. In order to show this, she herself assumes that the subject’s judgments simply reflect the conception associated with ‘water’ at that time. Moreover, the epistemic situation of the subject in Schroeter’s scenario at t_2 is equivalent to that of the subject in my scenario at t_0: Both believe that the liquid in the rivers and lakes of their environment is a culinary kind, and chemically a mixture of many different substances. So if it is rational for Schroeter’s subject to judge at t_2 that ‘water’ refers only to specific isotopes of H_2O with certain admixtures, then the same should be true for the subject in my scenario at t_0. I actually believe that there are good reasons to doubt that the judgment in question is perfectly rational. (I will return to this issue below.) But as was just shown, this line of reasoning is not available to Schroeter, since it would undermine her own argument.

One might wonder whether my scenario is disanalogous to Schroeter's in some other relevant way. The improv model requires that any change in the conception is rationally mandated by the newly acquired empirical evidence. But is this really the case in the example I gave? I.e., on discovering that the liquid in the rivers and lakes is a natural kind, should the subject really change her conception and then conclude that water is H_2O?

I think it is at least plausible that she should. A natural kind appears to be a better referential candidate than a culinary kind. For someone who is not convinced by this reason alone, one could modify the scenario slightly, for instance by adding that between t_0 and t_2, the subject acquires additional evidence to the effect that there is no such thing as a basic drink. But maybe even more importantly, if Schroeter held that the change in the extension is not rational, this would entail some undesired consequences. To see this, suppose that in the thought experiment I gave, upon discovering that the substance in the rivers and lakes is H_2O, the subject should not change her conception. Now assume that the scenario is such that once, at t_{-2}, the subject had already believed that water is a natural kind, but was then, at t_{-1}, confronted with strong (but ultimately misleading) evidence which suggested that this is false. She then changes her conception and at t_0, she believes that water is a culinary kind, namely the basic drink. (Notice that these events between t_{-2} and t_0 correspond exactly to those between t_0 and t_2 in Schroeter's scenario.) At t_1, the subject gets all the relevant information about her environment. In particular, she learns that the liquid in the rivers and lakes around her is a natural kind after all. By our current assumption, she will not change her conception and insist that 'water' refers to the culinary kind. However, if she had never been confronted with evidence which suggested that water is not a natural kind, her judgment at t_2 would have been based on the conception of water which she had from the very beginning. Accordingly, her judgment at t_2 would have been very different, even though in both cases, she eventually has complete and correct information about the world. Once again, in conjunction with thesis (CJ), this implies that the extension of 'water' must

differ as well between these counterfactual variations which are only distinguished by the evidence presented in an intermediate step.[72] Since this consequence is patently absurd, I conclude that this response to my argument fails as well.

Schroeter's improv model includes another condition for conceptual continuity: There must be some fact or feature which explains the previous success of our representational practice connected with the concept and which therefore must not be lost when one changes one's conception.[73] I must confess that it is not altogether clear to me how exactly this requirement is to be understood. It could mean that there must be some explanation for the success of our former applications of the term or concept in question, and thus the fact which explains this success (despite the defectiveness of the previous conception) has to be mirrored in the new conception. Or it could mean that there must be something about the former conception itself which was good and which therefore has to be retained in the new conception. But in each case, I do not think there is a way to spell out this requirement which could help to rebut my argument: If the condition is just that there must be some fact which explains the success of our representational practice and which is incorporated in the new conception, then I think it is clearly met in my example: The required fact is simply that there is a natural kind to which the subject referred in at least most of her applications of the term 'water'. If, however, the explanans has to be a part of our former, and thus also of our new conception, then it is still hard to see how it could be met in Schroeter's example without being met in mine. First of all, since the cases are symmetrical, the earlier and the later conceptions in her scenario obviously have as little or as much in common as those in mine. Furthermore, it seems that whatever aspect of

[72] For someone who wonders whether the counterfactual case is really relevant here, one can alternatively assume that there is another subject involved who starts with the same conception as the first one, is also provided with complete empirical information at t_1, but who is not given the misleading evidence at t_{-1} and thus never changes her conception of water.

[73] Schroeter briefly talks about this requirement in Schroeter 2006, 573.

the conception which explains the success of the representational practice in her example can also be taken as the explanans in mine – for instance, that there is some unifying kind which underlies the applications of the concept. Therefore, the two cases again appear to be completely analogous.

Another possible option for Schroeter is to make the constancy of an expression's extension an explicit additional condition for conceptual constancy. She could then say that the scenario I invoked does not conform to her model simply because it involves a change in extension. More specifically, she could concede that in my scenario, the change in the conception is accompanied with conceptual change due to the change in extension, while insisting that in her own scenario, there is no change in extension and thus conceptual constancy. This would allow her to maintain that a change in the conception is compatible with constancy of meaning, and consequently with constancy in the subject matter.
One immediate consequence of this proposal is that the criteria offered by Schroeter's model would then no longer be epistemically accessible to a subject. That is to say, when a subject is confronted with evidence which undermines her conception of a given subject matter, she is not in a position to know whether a change in the conception would involve a change in the extension or whether it would not. Another, related consequence is that the improv model would then entail externalism about rationality. To see this, recall that one of Schroeter's central theses is that if a change in the conception is rational, conceptual constancy is guaranteed. Consequently, the current proposal implies that the change in extension between t_1 and t_2 in my scenario is irrational, in spite of the subject not being in a position to realize this. I do not think that this response to the argument against the improv model is particularly promising. Besides from those just outlined, it does have a number of consequences which I think are undesirable. But I will not discuss these any further, because it seems to me that even accepting all of them would not save the improv model. For note that in the scenario I sketched, the problem arose not because there was a straightforward change in the extension of 'water', but rather because

of a divergence of two judgments which were *de facto* about the world inhabited by the subject. So suppose that in Schroeter's own scenario, the subject would consider w_1 – which is, unbeknownst to her, her own world – as actual at t_0. It is not completely clear what she would judge the extension of 'water' to be. But if she, as Schroeter herself assumes throughout, bases her judgment exclusively on her conception of water at that time, which includes that water is a natural kind, then she will surely not conclude that water is just the inhomogeneous mixture of chemical substances which serve as a culinary kind, as she does at t_2 after her conception has changed. It thus seems that if Schroeter holds that the subject's adjustment of her conception in my scenario is not rational because it involves a change in extension, then she will have to conclude the same with respect to her own scenario.

The last possible defense of the improv model I want to discuss starts from the claim that considering a world as actual in which something is the case is different from actually learning that something is the case, and that therefore, thesis (CJ) should be read as saying only that a subject is able to determine the extension of her concepts if she is given information (which she takes to be) about her own world, and not about a scenario which is considered as actual.[74] Inter alia, this would mean rejecting the Bayesian principle of conditionalization, which implies that if you believe that if p then q, and then come to know that p, you should believe that q. One could then say that my counterexample is irrelevant because the problems it reveals only concern hypothetical judgments, in response to hypothetical evidence.
The following remark from Schroeter seems to indicate that she does not endorse such a view. As she sketches her hypothetical scenario, she assumes that the subject is confronted with new (hypothetical) information E. Then she elaborates:

[74] Thanks to David Chalmers for making me aware that this is a position in logical space.

> You do not know whether or not E is true of your real environment. Nevertheless, you can reason about E hypothetically, treating E as if it were true. […] I shall assume it is uncontroversial that when E is true of your real environment, your verdict that water = H_2O will be true as well. This is thesis (A) [which corresponds to (CJ), J.K.], which I endorsed in §II. (Schroeter 2006, 576)

Schroeter thus seems to assume that it does not matter whether the subject takes the evidence to be real or not. However, later we read: "[W]hen you engage in real empirical enquiry, you are no longer making hypothetical judgements – you are actually revising your beliefs." (Schroeter 2006, 579) This comment suggests that she thinks there is a difference between the two kinds of judgments. And in fact, this might be more in line with her general account. Her argument against thesis (CJ+) implicitly depends on the assumption that on the one hand, when a subject considers a scenario as actual, her judgment will be based exclusively on her conception, but on the other hand, when she is confronted with actual evidence, she will change her conception and then ultimately base her judgment about the extension of 'water' on the new one.

In any case, I think the whole idea of making a distinction between hypothetical and actual judgments in the way just outlined is flawed. Conditionalization is a central principle of Bayesian epistemology. Even if one thinks there are exceptions to its validity, it is hard to believe that it fails in such mundane cases of empirical inquiry.[75] Moreover, rejecting this principle merely in order to save the improv model, without offering any independent reason for doing so, would clearly be ad hoc. In my view, the more plausible conclusion to draw is that the judgments about hypothetical cases as assumed by Schroeter are not really ideally rational, and thus, what Schroeter calls the conception associated with a term or concept does not correspond to its primary intension.

[75] Typical counterexamples to conditionalization involve paradoxes connected with self-locating beliefs (cf. e.g. Artzenius 2003, Titelbaum 2008).

Let me formulate some insights to be gained from the discussion of possible responses to my argument against the improv model: It transpired that Schroeter's account has trouble dealing with cases where subjects are (initially) presented with misleading evidence. More generally, the improv model seems to imply that the order of presentation of the evidence affects the ultimate judgment which is based on complete evidence.[76] This is plausibly to be explained by either non-ideal reasoning from the side of the subject or conceptual change.

A more general lesson to draw is that the conception associated with a term or concept as construed by Schroeter is not equivalent to the associated properties, because judgments which are based on the former are not always ideally rational. For the same reason, these conceptions cannot ground (CJ). Therefore, Schroeter's attempt to refute Jackson's thesis that primary intensions determine the subject by undermining the step from (CJ) to (CJ+) fails as well, together with her improv model.

These considerations do not prove that the step from (CJ) to (CJ+) is valid. There might be ways to question the idea that we have the required ability with respect to remote scenarios which are very different from Schroeter's objections. Nevertheless, the failure of Schroeter's account demonstrates that it is not easy to argue that there is a relevant difference between judgments about the actual world and those about worlds considered as actual, if one is given complete information about these worlds. Moreover, since, as I argued above, the step from (CJ) to (CJ+) is quite natural and since to date, Schroeter's improv model represents the only attempt to

[76] Here is one more example to demonstrate this: Suppose that subjects A and B start with the same conception of X, according to which X has the properties a, b, and c. In response to new evidence, A's conception changes to bcd; then to cde; and eventually to def. Meanwhile, B does not acquire any new evidence. Finally, A and B are given all the required empirical information, leading B to conclude that her conception of X being abc is correct. But evidently, this conclusion is no longer open to A, thus her judgment will differ from B's.

reject it, I conclude that this step in my argument for my version of Jackson's epistemic thesis is at least in good standing.[77]

4.1.1.2 From (CJ+) to (CJ++)

As I mentioned above, Block, Stalnaker (cf. Block & Stalnaker 1999) and Polger (cf. Polger 2008) take issue with the second step required for my argument, the one from (CJ+) to (CJ++). They agree with Chalmers and Jackson that we can determine the extensions of our concepts if we are given empirical information in a limited vocabulary (in particular, microphysical information) and also seem willing to grant the same for hypothetical information about non-actual scenarios,[78] but they deny that the knowledge thus gained is a priori. Block, Stalnaker and Polger believe that the judgments in question are usually based on empirical methods or background beliefs, such as inferences to the best explanation or simplicity assumptions. Note, however, that the mere fact that such judgments can be justified empirically or even that they are often empirically justified does not speak against the claim that they can also be justified a priori. So, in order to show that the transition from (CJ+) to (CJ++) fails, one has to argue that the judgments referred to in (CJ+), or at least some of them, rely essentially on empirical justification. In fact, in their joint paper from 2001, *Conceptual Analysis and Reductive Explanation*, Chalmers and Jackson themselves give an argument which demonstrates that empirical information cannot play an essential role in justifying these judgments (cf. Chalmers & Jackson 2001, 347–349). Let me give a slightly modified – and in fact simplified – version of the argument:[79] Take the derivation of a fact

[77] I will discuss more general considerations for and against the transition from (CJ) to (CJ+) below.

[78] In this aspect, Polger is more explicit than Block and Stalnaker, whose view on this particular issue is not completely transparent from their writings.

[79] I think the simplification does not take anything from the argument's force. At the same time, my specific version of the argument will turn out to be useful below.

F from a canonical description D_i of some scenario: $D_i \rightarrow F$.[80] Block, Stalnaker, and Polger would claim that very often, some empirical background assumption B (for example the belief that the world is simple) will be required to justify such a conditional. But what is the status of B itself? Recall that within the two-dimensionalist framework, an a priori true sentence is one which is true with respect to every world considered as actual, as stated by thesis (2D2) (cf. chapter 2). Therefore, since B is not an a priori truth, it will be true with respect to some scenarios and false with respect to others. Furthermore, by assumption it should be derivable from descriptions of those scenarios where it is true and its negation should be derivable from descriptions of those scenarios where it is false. Now assume that B is false in the scenario represented by D_i. It is clear that in this case, B cannot be necessary for justifying $D_i \rightarrow F$. Therefore, B can only essentially play a justifying role in the judgments about such scenarios where it is true. But now consider such a scenario. According to what was noted above, B itself is derivable from a canonical description of that scenario, which we may call D_j. So even if B is indeed required to justify $D_j \rightarrow F$, it still does not have to be taken as an empirical background belief, because there is always the alternative route at hand to derive B from D_j and then to use the thus derived assumption (or the conditional $D_j \rightarrow B$) to justify F. Thus, the only remaining option for someone who rejects the step from (CJ+) to (CJ++) is to try and argue that the justification of $D_j \rightarrow B$ is itself dependent on some other empirical background assumption C. But this would not help either because one could then run the same type of argument to show that C cannot play an essential role in the justification of this conditional, since it is itself derivable from any scenario where it holds.[81]

---

[80] Strictly speaking, the argument targets the (a priori) scrutability of truth, rather than the scrutability of reference which underlies (CJ) to (CJ++). The difference should not matter here, though.

[81] I noted in chapter 2 that Chalmers requires a priori justification to be conclusive. Given such an understanding of apriority, the argument just presented is not sound. This is because, since empirical justification need not be conclusive, $D_j \rightarrow F$ may be

We saw above that Schroeter's case against the step from (CJ) to (CJ+) fails. But beyond that, besides from some very general considerations, I did not give a positive argument to show that this step is valid. Since, as was just shown, the other relevant step in my argument for the epistemic thesis, the one from (CJ+) to (CJ++), is quite straightforward, it is the first step which should be resisted by an opponent. I therefore think it is worthwhile to have another look at this issue.

It is very plausible that the key reason for someone to reject the step from (CJ+) to (CJ++) is that one does not believe that linguistic expressions are a priori associated with properties, or maybe just that it is not a priori associations which guide our judgments. For if these judgments are based on empirical associations, then it is natural to assume that they can only be made with respect to nearby possible worlds, and in particular not with respect to worlds where the relevant correlation fails to hold. Unsurprisingly, such a view is also held by Schroeter. One could thus try to take an alternative route to establishing (CJ++), namely via the following thesis:

(CJ_{ap}) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension a priori.

(CJ_{ap}) adds to (CJ) only the claim that the conditional judgments about the actual world are a priori. It is therefore not committed to the claim that we have the relevant abilities with respect to every possible world. However, if one succeeded in establishing (CJ_{ap}), then the step to (CJ+) (and (CJ++)) would be greatly facilitated. And to do this, one can apply the same kind of

---

empirically justifiable, but not conclusively. On a more modest understanding of apriority, according to which the justification conditions for a priori knowledge are the same as those for empirical knowledge – a view which I defended in chapter 2 – the argument does go through, though.

argument which I just used to support the step from (CJ+) to (CJ++):[82] Let me call the canonical description of our actual world $D_@$. According to (CJ), the extension of each of our concepts (or alternatively the truth-value of each thought) is either explicitly specified in $D_@$ or derivable from it. Now take the derivation of some fact F from $D_@$ which is allegedly not a priori. In this case, the judgment has to be essentially based on an empirical fact or background belief, call this B. At least if F is really known, then it cannot be based on a falsehood; thus, B has to be true. But in this case, B itself is either explicitly specified in $D_@$ or derivable from it. One therefore does not need the empirical background assumption B to justify F or $D_@ \rightarrow$ F: One can always derive B from $D_@$ and then use it in turn to derive F. And if one thinks that B itself is only a posteriori derivable from $D_@$, then any putative empirical background belief or fact on which the step from $D_@$ to B is supposed to rely can also be derived from $D_@$; etc. These considerations strongly suggest that if the information contained in the canonical description enables a subject to determine the extension of all concepts and thoughts, then the corresponding judgments at least can be justified a priori.[83]

The argument just given only aims to support (CJ_{ap}) and therefore does not make a direct, positive case for (CJ+). However, as I mentioned above, once (CJ_{ap}) is established, a key motivation for rejecting the step from (CJ) to (CJ+) disappears.

I conclude that there are very good reasons to endorse my version of the epistemic thesis that primary intensions are required to determine the

[82] Chalmers and Jackson's original argument in fact presupposes (CJ+). For this reason, my modified version which only relies on (CJ) is more suitable at this point.

[83] I should note that this version of the argument which starts from (CJ) relies on an additional premise about the epistemic position of actual subjects. If, for instance, a subject was omniscient, then she could trivially derive all truths from an arbitrary set of sentences. But of course, this would not show that these derivations are a priori. Since actual subjects only know a fraction of these facts, this is not a problem for the argument.

subject and are therefore indispensable for empirical inquiry. But recall that the way I phrased it, the epistemic thesis should be read as a conditional claim: If we have the ability to determine the extensions of our concepts and thoughts in the face of empirical evidence, as specified by (CJ), then we need primary intensions to account for this ability. So an opponent of two-dimensionalism can (and maybe should) reject the antecedent of the conditional. The tenability of (CJ), which is therefore vital for the case for two-dimensionalism, will be discussed in detail in chapter 6.

4.1.2 The failure of the semantic thesis

Now let me turn to Jackson's semantic thesis that primary intensions define the subject, such that at least a significant change in the primary intension of the (or a) key term relevant to the inquiry amounts to a change of subject. In the following, I will argue that this thesis is false, at least with respect to many epistemically opaque terms such as natural kind terms. The resulting view will nevertheless be compatible with the general two-dimensionalist framework. I will conclude with a few remarks on how according to my account of what defines the subject, changes of subject can be recognized or avoided.

Let me start with a toy example: Assume that I introduce the term 'flim' which is supposed to refer rigidly to the place one meter to the left of me. At some point, I move two meters to the left and then stipulate that now, 'flim' is to refer rigidly to the place one meter to the right of me. The properties associated with the term have changed considerably, but it would be odd to claim that I have changed the subject. We often use indexical expressions in a similar way to retain constancy of subject matter in spite of (or even by means of) a shift in the associated properties. The only difference between the cases is that in the case of indexicals, we use different expressions to ensure constancy of subject matter – say, by using 'yesterday' instead of 'today' –, while in the example I gave the same term is used. Still, in both cases there is constancy of subject matter despite the change in the associated properties.

My next argument is based on the idea that there are close connections between constancy in subject matter and disagreement. Let me illustrate this by reference to a case of a putative change of subject. Suppose the question is 'Is X F?'. But in the course of the investigations, the meaning of 'X' changes, so that we have in fact two questions, 'Is X_1 F?' and 'Is X_2 F?' Now, if we say that these are really two different questions in the relevant sense, i.e., if we take this as a change of subject, then we should also be prepared to say that if one person says 'X is F', meaning that X_1 is F and another person says 'X is not F', meaning that X_2 is not F, there is no genuine disagreement between them.[84] Note that such a case can occur when one of the parties involved has changed the subject, but it need not. Maybe they were (unknowingly) addressing two different questions from the outset. Now my claim is that in a conversation involving the term 'X', it can happen that two parties genuinely disagree over some feature of X despite the fact that they associate 'X' with different properties. Such cases will often involve names, natural kind terms, or generally deferential uses of a term, i.e. epistemically opaque terms. As an example, take Burge's 'arthritis' thought experiment which was already discussed in chapter 3: Although the doctor obviously associates different properties with the term 'arthritis' than his patient, he genuinely disagrees with the patient's opinion that he has arthritis in his thigh. Otherwise, it would not make sense for the patient to revise his belief.

Given the connections between disagreement and constancy of subject matter just outlined, I hold that if there can be genuine disagreement between two parties in spite of a significant divergence in the associated properties, then one should also be ready to say that there can be constancy of subject matter in spite of a significant change in the associated properties. Again, the conclusion is that a change in a term's primary intension does seem compatible with constancy of subject matter: It turns out that Jackson's semantic thesis that primary intensions define the subject is false with respect to specific kinds of terms. An important observation

[84] There are some subtleties to take account of here, but in general, this seems correct.

here is that what the cases in which Jackson's model fails share is that despite the change in the primary intension, the extension as well as the secondary intension remain the same.
The failure of Jackson's account (re-)raises the following two questions: Firstly, what is the practical relevance of this issue; i.e., are there actual cases in which the primary intension of a term changes without a change in the secondary intension and thus of the subject matter? And secondly, if there are such cases, how does this bear on the epistemic problem connected with changes of subject, i.e., the problem of identifying them?

On the first question: If one understands natural kind terms as theoretical terms in the way described in chapter 3, a view which is by all appearances endorsed by Jackson, then one should expect changes in primary intensions without changes in secondary intension to be quite common. When we learn new things about the kind in question, or maybe when our relation to it changes, the theory about the kind changes as well. Even if one believes (as I do, cf. chapter 3) that not all of our beliefs about some kind are part of the term's primary intension, it is still plausible that there are often revisions concerning properties which are among the associated properties. If 'water' is a theoretical term, then plausibly, such properties as 'being an element' were once part of the theoretical role connected with water, and 'being H_2O' should be now. And thus, when it was discovered that water is not an element, but is rather composed of hydrogen and oxygen atoms, this did affect the primary intension of 'water'; but plausibly, it did not affect its secondary intension and thus did not imply a change of subject. I thus think there will often be such changes in the primary intensions of natural kind terms, in virtue of which the associated theory becomes more accurate, vagueness is reduced, and so on.[85] Thus, there will often be changes in the primary intension without changes of subject.

[85] I even suspect that there are cases, in particular among scientific terms, where at some point such a term will stop being a theoretical role term and instead simply be defined via some of the kind's essential properties. One example might be acid: Originally, 'acid' was plausibly a theoretical role term, while it may today be a priori

Let me end with some considerations about the second question I asked above, namely how this result bears on the epistemic problem of recognizing changes of subject.[86] One major advantage of Jackson's view which was already mentioned is that it seems to solve this problem straightforwardly: Since primary intensions are accessible a priori, we should also be able to check whether the primary intension associated with a term, and thus the subject, has changed. But if, as I propose, a change in a key term's secondary intension is the most basic criterion for a change of subject, then the epistemic problem seems to arise anew – after all, secondary intensions need not be accessible to a speaker at all.
Nevertheless, in the examples I gave there is no serious problem apparent. There was no change of subject when, for instance, our theory of water was refined, and it seems like there was never a real danger that this could happen. But this is not to say that in practice, Schroeter's idea is correct after all, that as long as we make 'rational' revisions in our conception (in her sense), there is no risk of changing the subject. I insist that her improv model fails to offer a useful criterion for constancy of subject matter. And while I reject Jackson's semantic thesis that primary intensions define the subject, I still think they play an essential role in ensuring such constancy. For note that there is no reason to assume that a two-dimensionalist is committed to the thesis that primary intensions define the subject in Jackson's sense. What she is committed to, though, is the weaker claim that primary intensions determine a term's reference with respect to a context, and its secondary intension with respect to a (centered) world considered as actual. I therefore hold that in the dispute between Jackson and Lycan mentioned above, both are wrong: Against Jackson, I claim that a change in the associated properties, however drastic, need not amount to a change of subject. But unlike Lycan, I think we cannot just ignore the originally

---

(at least regarding the scientific concept of acid) that acids are proton-donors. I think one can argue for something similar in the case of terms like 'star', and probably many other terms.

[86] Note that this problem is not the same as the one discussed in the previous section of determining one's subject.

associated properties; not least because an unconstrained change in the associated properties usually will lead to a change of subject. My view is simply that if we want to change the properties associated with an expression, we have to make sure that the primary intensions before and after this change pick out the same category in every possible world. This may sound like a very ambitious task, so one should note that it only requires to make sure that they actually pick out the same kind and then to 'rigidify'. Of course, our reasons for assuming that two primary intensions refer to the same kind will always be empirical and fallible. But they can nevertheless be extremely compelling. Besides from that, we cannot expect to have an a priori guarantee for continuity of subject matter anyway since, for instance, we can never exclude the possibility of memory failures and the like.

My view thus leaves room for the possibility of changing the theory about a kind without thereby changing what the theory is about. At the same time, it assigns a crucial epistemic role to the associated properties in ensuring constancy of the subject matter. The resulting picture is, I think, genuinely two-dimensional: Changes in the secondary intension are a better criterion for changes of subject than changes in the primary intension, as witnessed by the fact that when terms are involved which are not semantically neutral, there can be changes in the associated primary intensions without changes of subject matter. Nevertheless, primary intensions play an important role in ensuring the constancy of the subject matter in our epistemic practice.

4.2 Two-dimensional communication

Jackson argues in a number of places that primary intensions play a crucial role in communication (cf. Jackson 1998b, 2001, 2004, 2007, 2010). His line of reasoning goes roughly as follows: Language is a system of representation, which means that it can be used to indicate the distribution of properties in the world. This is made possible by the fact that linguistic expressions are associated with these properties by convention. Now, one of the central functions of language, perhaps its central function, is that of

being a medium in the exchange of information, i.e. in communication. And this requires not only that there are associations between words and properties, but that these associations are known to the speakers.

From these considerations, Jackson draws the conclusion that primary intensions are what is communicated when subjects pass information to each other. Secondary intensions are not eligible for this job because, as I just mentioned, the associations between words and properties have to be known to the subjects involved. To give an example invoked by Jackson, it does not seem reasonable to say that when someone in 1750 uttered 'There is water around', he was thereby transmitting the information that there is H_2O around (cf. Jackson 2004, 262f.). For since neither of them knew that water is H_2O, we would then have to say that neither did the speaker know which information he was transmitting, nor did the hearer know which information he was receiving. And this idea leaves it quite mysterious why we attach such great value to information.

Let me extract two theses about the role of primary intensions in communication from the above remarks, a semantic thesis and an epistemic thesis: According to the semantic thesis, primary intensions are what is communicated from speaker to hearer in a conversation. The epistemic thesis says that in order for language to be able to play its role in our practice of passing information to each other, there have to be primary intensions – i.e. known associations between linguistic expressions and properties.[87]

Semantic thesis (ST): Primary intensions are what is linguistically communicated.

Epistemic thesis (ET): If there were no primary intensions, linguistic communication would not be possible.[88]

[87] Jackson explicitly endorses both of these theses in Jackson 2004: For the semantic thesis, cf. Jackson 2004, 261; for the epistemic thesis, cf. Jackson 2004, 265f.

[88] The thesis might seem inappropriately named. The following way of phrasing it should make its epistemic dimension more obvious: If there were no primary intension, language would not be suitable for mediating the exchange of information.

These versions of Jackson's theses may be a bit too strict – at least Jackson does not commit himself to precisely these formulations. In the following, I will also discuss whether qualified versions of the theses can be held.
I take it that Jackson's argument for the essential role of primary intensions in communication is primarily an argument for primary intensions. But recall that unlike me, Jackson believes that primary intensions primarily reflect the meaning of expressions in a public language (cf. chapter 3). One therefore has to make a distinction: One can either try to argue for the claim that linguistic expressions are associated with primary intensions by linguistic convention, such that primary intensions reflect the public meaning of these expressions – this is the thesis which Jackson tries to support. Or one can hold the weaker view that linguistic expressions are just associated with primary intensions by subjects, i.e. they can be associated with thoughts whose contents are primary intensions. As will become apparent in the discussion, the semantic thesis about communication is intimately connected with the former of these claims. It seems plausible, for instance, that for (ST) to be true, there have to be primary intensions in that stronger sense. (ET), however, can be read as saying that successful communication requires the existence of primary intensions in the weaker sense, i.e. subjects individually associating expressions with primary intensions.

The structure of this section is as follows: I start by discussing (ST). I will argue that the thesis can be held with respect to indexical expressions, albeit with some qualifications. I will then show that it is doubtful whether it applies to communication involving names. Finally, I will argue that the thesis fails with respect to many natural kind terms. This result threatens to undermine the support for (ET) as well, and thus Jackson's whole account of the importance of primary intensions in communication. However, I will argue that primary intensions nevertheless play a vital role in communication, by proposing two possible ways to defend a version of (ET) in the face of the failure of (ST): The first of these proposals involves arguing for a restricted version of (ST). The second proposal aims to

establish (ET) in a way which is independent of (ST) – it invokes an argument to the effect that primary intensions can play an important role in communication even in cases where they are not transmitted from speaker to hearer.

4.2.1 The semantic thesis

Jackson's general picture of the role of primary intensions seems quite appealing. However, his account is seriously threatened by the fact that there can be intersubjective variation in the primary intensions associated with an expression.[89] To show why such variation raises a challenge to Jackson's view, let me give an outline of what I take to be the basic structure of communication: An act of successful communication involves at least two subjects, a speaker and a hearer. The speaker makes an utterance addressed to the hearer which expresses a thought. Then, upon hearing (and understanding) the speaker's utterance, the hearer comes to entertain the thought expressed by the speaker. It is very plausible that this transfer of thought contents from speaker to hearer is a necessary condition for successful communication. In turn, this entails that the thought expressed by the speaker and the thought acquired by the hearer need to have the same content. Let me call this requirement (CI), for Content Identity:

(CI) Successful communication requires that speaker and hearer entertain a thought with the same content.

If one combines (CI) with Jackson's semantic thesis that primary intensions are what is transmitted in linguistic communication, this suggests the following thesis:

(CIP) Successful communication requires that speaker and hearer entertain a thought with the same primary intension.

[89] This worry is also raised by Kroon (cf. Kroon 2004, 2009).

Now suppose that in a conversation, the speaker expresses her thought by uttering the sentence S. If the hearer associates S with a different primary intension than the speaker, then she will come to entertain a thought with a different primary intension. But by all appearances, communication works fine in many such cases. (Some of these cases will be considered below.) I think this observation raises doubts about both of Jackson's theses. For, once it is established that sameness of primary intensions is not a necessary condition for successful communication, there seems to be little reason left to assume that primary intensions play any role in the transfer of information, which puts into doubt the epistemic thesis as well.

One note about methodology: It is plausible that many kinds of expressions are semantically neutral, i.e. their primary intension coincides with their secondary intension. Accordingly, if communication via such expressions involves the transfer of primary intensions, then it also involves the transfer of secondary intensions. Such expressions are thus not suitable for testing the semantic thesis. I will therefore focus on three kinds of expressions which are at least widely regarded as two-dimensional: indexicals, proper names and natural kind terms.

It will be useful to distinguish two kinds of intersubjective variation in primary intensions: Firstly, primary intensions yield centered contents, i.e. contents which are relativized to an individual at a time (cf. chapter 2). (I will call this individual 'Center' in the following.) This implies that when expressions are involved which are sensitive to centering, the primary intensions can vary from subject to subject as well. Differences in centering can most obviously become relevant in the case of indexical expressions like 'I', 'here', or 'now'. However, if my discussion in the preceding chapter was adequate, then natural kind terms and names also have a hidden indexical (or 'self-locating') component: Following Lewis, Chalmers, and Jackson, I proposed to analyze the primary intension of 'water' (very roughly) as 'the watery stuff of *our* acquaintance'. The primary intension of a name such as 'Gödel', I argued, can be expressed by something like 'the individual called 'Gödel' by those from whom *I* acquired the name'. So the question how such centered contents can be

communicated will be relevant for all the kinds of terms which I am going to consider.
Secondly, there can be variation in the primary intensions associated with an expression which is independent of differences in centering, which one might call variation in descriptive content. As we will see, such variation is possible, and troublesome, in the case of names and in particular of natural kind terms.

4.2.1.1 Communication involving indexical expressions

It is plausible that in the case of indexicals, variation in the associated descriptive content (which may correspond to Kaplan's 'character') will not occur among competent speakers. But the question how the associated centered contents are to be communicated still needs to be discussed. For, as Robert Stalnaker pointed out already in 1981 in his *Indexical belief*, it is unclear how such communication could work (cf. Stalnaker 1981). Suppose for instance that in a conversation with Heimson, David Hume says 'I am Hume'. It is obviously undesired that Heimson thereupon acquires the belief that he, i.e. Heimson, is Hume. However, this is precisely what is predicted, given that (as I argued above) communication involves the transfer of thought contents. For, Hume's and Heimson's beliefs have the same primary intension. And this problem affects Jackson's semantic thesis that communication involves the transmission of primary intensions as well.
There have been some attempts to show how the transfer of centered contents is nevertheless possible. Stephan Torre and Dilip Ninan introduce worlds with multiple centers, one for each participant in a conversation (cf. Torre 2010; Ninan 2010). Importantly, at least in Ninan's account, each center stands for a specific subject in a conversation. So assume there is a conversation involving two subjects, Alvin and Batu. The contents of what they communicate are given by sets of worlds with two marked centers. These centers are in turn ordered in what Ninan calls a 'conversational sequence'. Let me take the conversational sequence for the present

conversation to be <Alvin, Batu>. This means that for each utterance made in the conversation, whether it is produced by Alvin or by Batu, the first center will always pick out Alvin.

To illustrate: Assume that in the course of the conversation, Alvin says 'I am funny'. The content of this utterance can be given as follows: {<w,x,y> x is funny at w} (or more simply as 'Center$_1$ is funny'). And when Batu expresses what she has learned from Alvin's remark by saying 'You are funny', the content of her utterance is also {<w,x,y> x is funny at w} ('Center$_1$ is funny'). If, however, she were to say 'I am funny', the content would be this: {<w,x,y> y is funny at w} (Center$_2$ is funny). Accordingly, Alvin's 'I am funny' and Batu's 'I am funny' have different contents, while Alvin's 'I am funny' has the same content as Batu's 'You are funny'. This seems to resolve the problem concerning the communicability of such contents. In particular, the proposal satisfies (CI), since what is said by Alvin is precisely what is learned by Batu. If one were to construe primary intensions analogously, one could thereby maintain (CIP).

In fact, this kind of solution does not require that one postulates worlds with multiple centers (although there may be other reasons to do so). It suffices to have one center if that center constantly picks out the same person in a conversation. This seems to be Jackson's proposal in his recent book (cf. Jackson 2010, 140f.).

However, the problem I see with all of these proposals is this: They may (or may not) provide plausible accounts of linguistic meaning or utterance content, but not of the contents of the associated thoughts. Take Torre and Ninan's account involving multiple centers: On their view, when I utter 'I am funny' in the presence of another person, the associated content will be doubly-centered, when there are five more persons present, the worlds will have six centers, and presumably, when I am just talking to myself, the content will just be singly-centered. However, it would be quite odd to say that the content of my belief is dependent on how many people are present.[90] Something similar applies to Jackson's account involving worlds

[90] To avoid that consequence, one would have to say that each thought content has a huge, possibly infinite, number of centers, one for each potential conversation partner,

with one center: Assume that in a conversation between Alvin and Batu, Alvin represents the center; simultaneously, Coco is the center in a conversation between Coco and Dario. Now when these four meet for a new conversation, it is impossible for both Alvin and Coco to retain their status as center. So let us assume that in the second conversation, Coco represents the center. Then the content of Alvin's utterance of 'I am funny' will vary, depending on whether it was made in the first or in the second of these conversations. I am not sure I like the idea that the content of such utterances depends on factors such as the number of speakers present. But I think it is just out of the question to say that this also applies to the content of the thought which is expressed by the utterance.

It is thus possible to introduce a version of centered content such that this content is shared between different speakers in a conversation. The trick is to have the center or centers pick out the same individual throughout the conversation. Nevertheless, this proposal cannot save (CIP), because the content in question is surely not that of the thoughts expressed by the speakers. In any case, such a proposal would be against the spirit of the account of Lewis who popularized the idea that thought contents are centered. For, that account which is in many ways a precursor of Chalmers and Jackson's two-dimensionalism is motivated by the idea that a subject's beliefs represent what the world is like from that subject's perspective. Consequently, Alvin and Batu do (in one sense) not have the same belief in the example I just gave; while Hume and Heimson do have the same belief, when each of them believes that he is Hume. That consequence is not per se a bad one since there clearly is a sense in which Hume and Heimson share the belief in question. But as we just saw, that view conflicts with an equally plausible view about communication.

I conclude that (CIP) fails: In many cases of successful communication involving indexical expressions, the speaker and the hearer do not come to entertain a thought with the same primary intension. Nevertheless, I do not think that it is reasonable to conclude that communication involving

---

such that each center is determinately assigned to an individual. I think such a view would be no less absurd.

indexical expressions merely consists in the transmission of uncentered contents. Consider for example the following conversation:

Alvin: 'I am funny.'
Batu: 'I agree, Alvin is funny – unlike you.'

I think it is obvious that although the singular thought that Alvin is funny has been transmitted from Alvin to Batu, communication failed here. It is important to note that this cannot be attributed to the fact that Batu fails to realize that Alvin is the speaker. To see this, compare the following conversation:

Alvin: 'Alvin is funny.'
Batu: 'Yes, he is – unlike you.'

In the second case, it is likely (though not certain) that, in Stalnaker's terminology, the context set is relevantly defective (cf. Stalnaker 1978), i.e. at least one presupposition relevant for the conversation is not shared by speaker and hearer. For it is plausible that Alvin presupposes that he, i.e. the speaker, is Alvin when making his utterance. But still, it seems perfectly natural to say that Batu has understood what Alvin told him – and thus that they communicated successfully – if she understands that Alvin is funny, even though she fails to realize that Alvin is also the person who is currently talking to her.

The examples suggest that when indexical expressions are involved, successful communication essentially involves centered contents. Plausibly, in the case where Alvin says 'I am funny', it is necessary that Batu comes to entertain not just the singular thought that Alvin is funny; rather, she needs to understand that the individual who is currently speaker-related to her is funny – which is of course to have another centered thought.

I conclude that in order to model communication involving indexicals, one should draw on primary intensions. There are (at least) two ways to do this: The first option invokes primary and secondary intensions and says that successful communication involving indexical expressions requires a) shared secondary intensions between speaker and hearer and b) an

appropriate relation between the associated primary intensions.[91] The second option is to try to do without secondary intensions. Since, as was mentioned in chapter 3, Jackson holds that thoughts only have primary intensions, such an account would be better in line with his commitments. One would then have to drop (CI) altogether and say that the contents of the thoughts which are entertained by a speaker and a hearer only have to be appropriately related. But notice that it is not sufficient to say that successful communication involves suitably related primary intensions. For, when indexicals are involved the primary intensions can never be appropriately related simpliciter, but only relative to a context – i.e., one would have to add that the communicating parties must also be related accordingly.

Let me sum up what we have found so far. The general problem was this: There are good reasons to assume that mental content is centered. But in precisely those cases where centered contents are needed – for example those involving indexicals – this assumption clashes with an otherwise very plausible account of communication, in particular with (CI). Nevertheless, it transpired that centered contents play an essential role in communication involving indexicals. These facts can be accommodated in two different ways. On both of the resulting views, successful communication requires that speaker and hearer entertain suitably related primary intensions. Their difference lies in their attitudes towards uncentered contents and thereby also towards (CI): On the first of the accounts I outlined, successful communication also involves the transmission of secondary intensions. It can thereby save at least a version of (CI). The second account tries to do without secondary intensions and thus also without (CI). However, as I noticed, one would then have to add that besides from the primary intensions, the communicating parties also have to be suitably related.
Since primary intensions play a crucial role in both of the accounts just sketched, the epistemic thesis is clearly not in danger. It is less clear how the above considerations bear on the semantic thesis. If one takes the thesis

[91] This seems to be the view held by Chalmers in Chalmers 2011b, 619–621.

to entail (CIP), then it obviously fails. But this may be an overly uncharitable interpretation of Jackson's view. If the semantic thesis that primary intensions are what is communicated is understood in a slightly looser sense, it could be considered compatible at least with the latter of the proposals just mentioned.

4.2.1.2 Communication involving proper names

Many hold that successful communication involving names only requires that the name's reference is shared, such that speaker and hearer will come to entertain the same singular thought. On such a view, the relation between the associated primary intensions is obviously irrelevant. But let us see whether Jackson's account can nevertheless be applied to communication involving names.

Following the account defended in the previous chapter, I will understand the primary intension of a name N deferentially, as in 'the individual called 'N' by those from whom I acquired the name'. Accordingly, the semantic thesis requires that communication involving names is construed as the transfer of centered contents. I think to motivate such a view, one would first of all have to give an account of how subjects derive centered information from an utterance involving a name. In the following, I will try to spell out what such an account would involve.

Consider again Alvin and Batu who are having a conversation about Coco. Alvin says: 'Coco is into art'. If everything goes right, then Batu will come to entertain a thought with the same primary intension as Alvin's, which we can phrase as 'The individual called 'Coco' by those from whom Center acquired the name is into art'. Notably, such a case is even in line with (CIP). In order to get a grip on how the communication of centered contents involving names should be conceived, it will thus be useful to compare it with cases involving the indexical 'here'. When, for instance, Alvin says something like 'It is dangerous here' to Batu, then there are plausibly situations such that successful communication requires that Batu comes to entertain a thought with the same centered content. Here is one

way in which one might construe the epistemology of this case, i.e. the way in which Batu derives centered information from the centered information conveyed by Alvin.

Alvin: 'It is dangerous here.'
Batu knows (or comes to know) that
(1) the speaker is at a dangerous place [via linguistic competence, possibly capacities for the evaluation of the reliability of testimony]
(2) Center is such-and-such related* to the speaker [via perception]
(3) Center-place is dangerous. [via inference]

* There is some place such that the speaker and Center are both located there or near enough (at the time of the conversation).

Compare this to the case involving a name, if we construe it as a case where centered contents are transmitted:

Alvin: 'Coco is into art.'
Batu knows (or comes to know) that
(i) the individual called 'Coco' by those from whom the speaker acquired the name is into art [?]
(ii) Center is thus-and-thus related** to the speaker [?]
(iii) the individual called 'Coco' by those from whom Center acquired the name is into art. [via inference]

One may suspect that the latter case is epistemically more demanding than the former one. Let us have a closer look at how it could work. Start with step one: Upon hearing Alvin's utterance, how could Batu come to know (i)? I think a reasonable suggestion is to say, in Jackson's terms, that it is part of our (implicit) folk theory that names are used deferentially. And since according to Jackson our folk theories determine the meanings of our expressions, one can then say that Batu comes to know (i) on the basis of her linguistic competence (and, possibly, of her ability to evaluate the reliability of testimony).

(ii) is slightly trickier. The required relation between Batu and the speaker is roughly this:
** There is an x such that the speaker acquired the name ‘Coco’ from x and there is a y such that Center acquired the name ‘Coco’ from y, and x refers to exactly one individual with ‘Coco’ and y refers to exactly one individual with ‘Coco’ and the individual thus referred to by x is identical with the individual thus referred to by y.
In order to know that this relation obtains, Batu again has to rely on her knowledge about how names are used. But this alone is not sufficient, because the name ‘Coco’ is not unique; and this applies to names quite generally. Therefore, linguistic competence will not enable Batu to know that the person or persons from whom she acquired the name refer to the same individual as those from whom the speaker acquired the name. Accordingly, in order to know (ii) Batu has to rely on some kind of additional background knowledge. Notice, however, that this is not a distinctive problem of the theory at hand. On any account, the subject will need some background knowledge if she wants to make sure that she co-refers with the speaker in using a name.
Finally, if one supposes that Batu somehow knows or comes to know (ii), then the step from (i) and (ii) to (iii) is trivial.

This, I take it, is roughly how speakers could derive centered information from the speaker-centered information provided by an utterance. As we just saw, such an account seems epistemically speaking quite intricate; it is not entirely clear whether it provides a plausible interpretation of how speakers gain knowledge from utterances containing names. But I will not elaborate on this issue. Rather, I will focus on another potential problem which becomes apparent once one realizes that the account at hand can only be applied if both speaker and hearer use a proper name roughly in the way I suggested above (i.e., deferentially).
For consider Dario, who has known Coco for a long time and is a close friend of hers. He has no tendency to use ‘Coco’ deferentially. Rather, he is disposed to apply the name to that person with whom he shared certain

experiences, who has specific noteworthy properties, etc. Let me thus assume that the primary intension associated by Dario with the name ‘Coco’ is something like ‘the person of Center’s acquaintance who actually plays the Coco-role’, where the Coco-role stands for some of Coco’s noteworthy properties, her relation to Dario, etc. When Dario says to Batu ‘Coco is funny’, thus expressing his belief (with the primary intension) that the person of his acquaintance who actually plays the Coco-role is funny, then Batu will come to entertain a thought with a quite different primary intension, namely that the individual called ‘Coco’ by those from whom she acquired the name is funny. Notice that this kind of variation in the primary intensions cannot be traced back to differences in centering. Nevertheless, as long as the name co-refers among Dario and Batu, communication seems to work just fine in this case, which suggests that it is irrelevant which primary intension Dario and Batu attach to ‘Coco’. If so, then communication involving names stands in contradiction to the semantic thesis that communication involves the transmission of primary intensions from speaker to hearer.

Accordingly, a defender of the thesis is committed to saying that it does matter which primary intension a subject associates with a name and thus, that something goes wrong in the case at hand. I think the most promising way for her to do this is to hold that if it is indeed part of our folk theory that names are used deferentially, then someone like Dario who associates a significantly different primary intension with a name thus violates a linguistic convention.

It is far from clear, however, that there really is such a linguistic convention concerning the use of names. Consider again Dario’s non-deferential use of the name ‘Coco’. For him, Coco is that person with whom he shared certain experiences and so on, regardless of how other people call her or whether this is the name she was given at birth. To me, this case does not seem particularly unusual. It is at least not obvious that people tend to defer to others’ use of a name in cases where they are closely familiar with the referent. To say that in such cases, the usage of the name is in some way incompetent would at least be counterintuitive.

Let me give a brief overview of the discussion up to this point. The main problem for the semantic thesis that communication involves the transmission of primary intensions is the possibility of diverging primary intensions between speaker and hearer. At least where the divergence is not due to centering, such cases seem incompatible with the semantic thesis. It transpired that there are cases of successful communication involving names where primary intensions are not shared among those involved in the conversation. I suggested that the most promising way to save the semantic thesis from this objection is to hold that in such cases, at least one of the participants in the conversation violates a linguistic convention. However, as we saw it is not clear whether such a convention exists. Notice also that even if it could be held that there is a linguistic convention to use names deferentially, it would still be questionable whether communication involves the transmission of the relevant centered contents, since it is unclear whether such an account provides an adequate characterization of the associated epistemology. I conclude that although it has not been conclusively shown that the semantic thesis fails with respect to communication involving proper names, it is implausible that such communication involves the transmission of primary intensions.
Next, let me turn to a discussion of communication involving natural kind terms. Here it is even clearer that there are cases with respect to which the semantic thesis fails.

4.2.1.3 Communication involving natural kind terms

Let me start my discussion with the paradigmatic example of ‘water’. According to the account I developed in chapter 3, the primary intension of the term ‘water’ is roughly this: ‘the actual watery stuff of Center’s[92]

[92] There, I mainly phrased it as ‘the actual watery stuff of *our* acquaintance’. The difference between these formulations does not matter for my discussion here. As I briefly mentioned in chapter 3, I consider the way of phrasing the primary intension of ‘water’ on which I draw here to be more precise (cf. footnote 44).

acquaintance'. Once again, this kind of content is sensitive to centering. But I will leave aside the question of how these centered contents are communicated. Rather, I will focus on the question of whether the associated descriptive contents are shared among speakers. Consider the example, brought in by Chalmers and Jackson, of the city-dweller and the beach-dweller (cf. Chalmers & Jackson 2001, 328): The city-dweller uses the term 'water' for the liquid that comes out of faucets (knowing nothing of oceans), and a beach-dweller uses 'water' for the liquid in the oceans (knowing nothing of faucets). Just like with diverging primary intensions in the case of names, it is at least not immediately apparent why a city-dweller should not be able to pass information about water to a beach-dweller. And again, like in the case of names it is arguable that neither violate a linguistic convention. Many have argued that the only requirement for a subject to master the term 'water' is that she uses it in a way such that she refers (rigidly) to H_2O (cf. e.g. Soames 2002, ch. 10, 2004; Lycan 2009). But in fact, I think the example of 'water' which is usually discussed in this context is a favorable one for the proponent of the semantic thesis. For recall that on Jackson's view, the meaning of an expression is given by the properties commonly associated with that expression – by what he calls the folk theory. Now it is very plausible, given the importance of water in our everyday life, that there is a folk theory of water. The case of the city-dweller and the beach-dweller is indeed a strange one, since one would assume that it is common knowledge that some water comes out of faucets and some water flows in the oceans.

However, it is very unlikely that the same is true for all natural kind terms. Consider for example terms like 'sea bass' or 'guinea fowl' or Putnam's examples 'elm' and 'molybdenum'. One will be hard pressed to identify folk theories for such natural kinds which have to be known to competent speakers. Primarily, the problem is that there is just too little common knowledge about these natural kinds. Subjects may use the terms in question deferentially, or associate them with some theoretical role if they are more familiar with the kind. But I think this theoretical role will still differ significantly between different speakers if for instance they know the

natural kind in question from very different contexts. For example, an engineer in a nuclear power plant (where an alloy of molybdenum is often used in the construction of pressure vessels) will presumably associate a very different theoretical role with ‘molybdenum’ than a biologist (molybdenum is inter alia an essential component of the active site of a number of enzymes).

I conclude that for many natural kind terms, it is highly implausible to assume that they are associated with a specific primary intension by convention. This entails the failure of the semantic thesis: In such cases, communication cannot involve the transfer of primary intensions from speaker to hearer.
This result also threatens to undermine the case for the epistemic thesis: For if primary intensions do not have to be shared between speakers and thus do not have to be transmitted in a conversation, then there seems to be little reason to hold that communication would not be possible without primary intensions. In the following, I will therefore look at ways to defend the epistemic thesis in the face of the failure of the semantic thesis.

4.2.2 The epistemic thesis

I see the following two ways to argue for the epistemic thesis which are not dependent on the truth of the semantic thesis. The first option is to say that communication usually, though not always, involves the transmission of primary intensions from speaker to hearer. This means to grant that there are exceptions to the semantic thesis, while insisting that it does hold in the majority of cases. The downside of this proposal is that in itself, the resulting thesis is too weak to vindicate the epistemic thesis. For if it is conceded that communication does work without the transmission of primary intensions, then the claim that they are required for language to play its role in the passing of information seems unfounded. For this reason, the present idea could be fruitfully combined with the following thesis: In those cases where primary intensions are not transmitted,

communication is somehow deficient, even though it does not altogether fail. And therefore, without the transmission of primary intensions language could not play the role it does play. Such a view entails that primary intensions are usually shared among speakers of a linguistic community. Any argument for it would therefore also help to defend conceptual analysis, the utility of which – as I already pointed out in chapter 3 – is crucially dependent on the question as to what extent primary intensions are shared.

The second way to adjust Jackson's argument for primary intensions is to decouple the epistemic thesis from the semantic thesis. One could say that even though communication does not always involve the transfer of primary intensions, they nevertheless play an important role in promoting communicative success, even in those cases where they differ significantly between speaker and hearer. Of course, if one were to follow this line, one would have to specify what exactly that role is, i.e. one would have to provide supporting evidence for the epistemic thesis which is independent of the semantic thesis. In the following, I will discuss the tenability of the options just sketched in turn.

4.2.2.1 The importance of shared primary intensions

According to the first of the proposals at issue, it would be bad if communication did not usually involve the transmission of primary intensions. To evaluate this claim, it will be useful to have an independent understanding of the requirements of successful communication. One minimal condition which is, I would guess, universally accepted is extensional equivalence. It should also be uncontroversial that this does not just involve the extensional equivalence of the thoughts transmitted from speaker to hearer (or entertained by speaker and hearer), but also that of the thoughts' constituents or the concepts associated with the speaker's utterance. Extensional equivalence is important for the preservation of truth. When a speaker expresses a true belief with an utterance and the hearer, upon hearing this utterance, comes to entertain and then to accept an

extensionally equivalent thought, then, trivially, she will have acquired a true belief.[93] The extensional equivalence of the associated concepts is important precisely to ensure the preservation of truth. There is thus a direct route from the importance of extensional equivalence of the thoughts transmitted to that of the subjects' vocabularies: Suppose that Alvin's and Batu's vocabularies are not perfectly extensionally equivalent, i.e. it is not the case that every concept associated with an (atomic) expression in Alvin's vocabulary is extensionally equivalent with the concept corresponding to the expression in Batu's vocabulary. Then it can happen that when Alvin expresses a true thought, the thought which Batu comes to entertain upon hearing the utterance will be false. Extensional equivalence of the vocabularies is thus required to ensure the preservation of truth. Adding to this the assumption that the rules of composition of Alvin and Batu's idiolect are equivalent already suffices for the preservation of truth in most communicative contexts – but not in all. In order to ensure the preservation of truth with respect to modal or other intensional contexts, it is also necessary to include the intensional equivalence of the vocabularies.[94]

Note that an opponent of two-dimensionalism, and of Jackson's semantic thesis in particular, can accept all of what has been said so far. She can hold that successful communication involves the transmission of, in two-dimensional terms, (appropriately structured) secondary intensions and that therefore, it is only sameness of the secondary intensions of the speaker's and the hearer's thoughts which is required. Accordingly, we still need an argument for the use of coinciding primary intensions in communication.

I just mentioned that intensional equivalence of the vocabularies of the participants of a conversation (to be understood as equivalence of the associated secondary intensions) is important for preserving truth in intensional contexts. The resulting intensional equivalence of the transmitted thoughts itself will also be useful: To start with, it will help to

[93] This point is also made in Heck 1995, 86.

[94] I set aside residual problems connected with context-sensitivity here, some of which were discussed above.

avoid relativism with respect to such intensional contexts. Furthermore, it will foster agreement about modal matters between speaker and hearer, i.e. they will tend to draw similar modal conclusions from what is communicated.[95] So why not argue along similar lines for the importance of shared primary intensions between speaker and hearer? For when the primary intensions of the relevant thoughts concur, this will tend to bring about agreement about epistemic possibilities. To illustrate: Whenever the primary intensions of two thoughts differ, they differ in their truth-conditions concerning at least one epistemic possibility. On the assumption that the subjects involved are rational, this implies that they will disagree over the evaluation of at least one hypothetical scenario. This can be harmful, since it implies that the subjects are disposed to draw different conclusions from potential evidence.

One has to be careful not to presuppose too much here. One can of course try to identify ideal conditions for communication. For instance, it would be desirable if the conversation partners were ideally rational and shared all the relevant background beliefs. This would ensure that they draw exactly the same conclusions from what is communicated. But this is neither a requirement of successful communication nor is it often satisfied in practice. So maybe something similar is true for the transmission of primary intensions in communication: It would be good if it were usually the case, but in our communicative practice, it is nevertheless a rare phenomenon. Therefore, if one tries to argue that our communicative practices rely on shared primary intensions, one does not only have to show that the absence of shared primary intensions involves significant disadvantages, it also has to be plausible that we are in fact, at least often, better off.

I think a promising way to demonstrate this is to refer to the success of mutual inquiries. Let me elaborate: In the previous section, I defended the idea that we need primary intensions because they guide us in empirical (and also non-empirical) inquiry. More specifically, they tell us what it is that we are looking for and under which conditions we have found it. Now

[95] This point is also made in Lewis 1999, 356.

consider an investigation which is conducted by more than one person. It is obviously important that those involved do not only have a common goal, i.e. a common subject matter, but that they also agree at least roughly on how to reach it; in particular, they should agree on what has to be the case for the investigation to yield a specific result. Shared primary intensions promote agreement about such matters; i.e., subjects who associate the same properties with a given expression will tend to pass the same judgments in response to specific evidence. For this reason, it is important that the investigators associate the key term(s) with at least similar primary intensions. According to this line of thought, shared primary intensions greatly facilitate the success of mutual inquiries, inter alia (but not only) by facilitating communication between those involved. Moreover, such mutual inquiries are both very common and play an important role in our epistemic practice – in science and elsewhere. This is evidence for the thesis that the sharing of primary intensions among communication partners is not only desirable, but that it is likely to be the rule rather than the exception. If successful, this argument could lend support to a qualified version of Jackson's semantic thesis, while maintaining the epistemic thesis that primary intensions play an indispensable role in our epistemic practice. The resulting view is that while communication does not necessarily involve the transmission of primary intensions from speaker to hearer, it still does so typically. And fortunately so, for if intersubjective variation were frequent and the transmission of primary intensions therefore rare, mutual inquiries would often be doomed to failure from the start.

An opponent of two-dimensionalism could object to the argument just given along the following lines: She could concede that if linguistic expressions were associated with something like primary intensions and these were often shared among speakers, then one could appeal to them in order to explain the feasibility of mutual inquiries. But there is an alternative explanation. The role played by primary intensions in the two-dimensionalist account can also be played by empirical background beliefs which are shared by those involved; there is thus no need to appeal to primary intensions. I admit that this is a good objection. Generally

speaking, it is hard to prove that only primary intensions can do the job in question. But one might still try to argue that the two-dimensionalist explanation of the possibility of mutual inquiries is superior to its rivals. On that note, let me propose two possible rejoinders to the objection just sketched. The second of these proposals, which I think is the more promising one, will draw on considerations from section 4.1.

It would be nice if the relevant background assumptions did not just happen to be shared, but if we could rather rely on this fact. A two-dimensionalist can argue that this condition is in fact satisfied, since primary intensions are associated with linguistic expressions by convention. But the force of this point is disputable. An opponent could well respond that in practice, it suffices if we rely on the fact that we share the relevant beliefs because of our common cultural background or, like in science, our common educational background. She could even argue that many empirical background beliefs are in fact associated with linguistic expressions by convention, for example by referring to Putnam's stereotypes. A more promising way to counter the objection that shared empirical beliefs are sufficient leans on the argument from (CJ) to (CJ++) given in 4.1. For note that not just any shared background beliefs will do to secure the success of mutual inquiries. As I argued above, the inquirers have to share the relevant dispositions to evaluate the outcome of the inquiry in the face of evidence. But the upshot of my discussion in 4.1 was that given a plausible thesis about this ability, the (implicit) beliefs which enable us to make such judgments in the face of evidence cannot be merely empirical; we need primary intensions to account for them. If one accepts this argument, then it is highly doubtful that shared empirical background beliefs can be a sufficient basis for successful mutual inquiries. For, if the primary intensions associated by those involved in the inquiry differ significantly, then even if they all have access to the same evidence and share all relevant background beliefs, their judgments about the outcome of the inquiry will nevertheless diverge.

Let me briefly take stock: The failure of an unrestricted version of Jackson's semantic thesis led to the question of whether one could hold that

the semantic thesis nevertheless applies in the majority of cases, which would mean that although communication does not always involve the transmission of primary intensions, it still does so mostly. This would imply that primary intensions are usually shared among communication partners. In order to support this latter thesis, I referred to the importance of mutual inquiries in our epistemic practice, and in turn to the importance of shared primary intensions for the feasibility of these inquiries. Notice that these considerations were mainly relevant to expressions whose secondary intensions differ from their primary intensions. For semantically neutral terms, the thesis that their primary intensions are generally shared among competent speakers (and the connected thesis that primary intensions are transmitted in communication) is very plausible. So the answer to the question of how often primary intensions are shared among speakers of a linguistic community also depends on the relative frequency of semantically neutral expressions in our everyday discourse. This issue will be addressed in the next chapter.

4.2.2.2 How primary intensions help to promote co-reference even when they are not shared

The view just discussed offers a rather indirect defense of the epistemic thesis that we need primary intensions for language to play the role it does play in our epistemic practice which proceeds via the (semantic) thesis that primary intensions are typically transmitted from speaker to hearer in communication. The downside of this strategy is that one thereby seems forced to limit the scope of the epistemic thesis to the same extent to which one has to limit that of the semantic thesis. But as I pointed out above, one might additionally defend the epistemic thesis in a way which is independent of the semantic thesis, by holding that primary intensions facilitate communication even in those cases where they are not shared between a speaker and a hearer – and to which the semantic thesis thus does not apply. In the following, I will demonstrate how primary intensions

serve to coordinate co-reference among speakers even when they are not shared, using the example of proper names and natural kind terms.

4.2.2.2.1 Proper names

Let me start with proper names: As I argued above, it is likely that communication involving names does not include the transmission of primary intensions. There are plausibly many cases where the primary intensions associated with a name by a speaker and a hearer differ significantly. Even in those cases where the primary intensions are shared because both of them use the name deferentially, it is questionable whether the associated centered contents are communicated. As I will try to show in the following, the primary intensions associated with names nevertheless play a crucial role in enabling communication.

Above I pointed out that sameness of extensions is a necessary condition for successful communication and that sameness of associated secondary intensions is at least very useful. Now assume that the only requirement for competent use of a proper name is to associate it with a specific extension or secondary intension. Then speakers should strive to make sure that their uses of the name have the same referent or secondary intension. The tricky part of this is to secure co-reference (since one can then simply intend to use the name rigidly). I think the simplest way in which this can be achieved is if each speaker is disposed upon acquiring a name to use it in accordance with the person(s) from whom she acquired it. In so doing, the speakers will ensure that when the name spreads in the linguistic community, its reference will remain constant. It is thus no accident that the primary intensions of names are deferential, as was proposed in chapter 3.

In my discussion of possible intersubjective variation in the primary intensions associated with names, I conceded that not every competent speaker has to use a name deferentially. I gave the example of Dario, who uses the name ‘Coco’ rather like a theoretical role term. But note that since Dario knows Coco so well, has interacted with her and talked with others about her many times, it is highly unlikely that he thereby fails to co-refer

with other speakers when he uses the name. If, however, people were to use names non-deferentially in cases where they are not well familiar with the referents, the risk of being out of line with others would be much higher. If one assumes that speakers aim to secure sameness of reference for the purposes of communication, one would thus expect non-deferential uses of names to be very rare in such cases. And I think the idea that names are generally used deferentially if we are not well familiar with the referents is confirmed by Kripke's discussion of actual and hypothetical cases involving names like 'Einstein', 'Gödel' or 'Feynman'. Its intuitive appeal indicates that we are disposed to defer and also expect others to defer to common usage in such cases.

Deferential uses of other kinds of terms seem to follow a similar pattern: Typically, when a speaker believes she has an insufficient understanding of a term, she will be disposed to use that term deferentially. Again, it is plausible that this mainly serves to secure shared extensions (and shared secondary intensions) among speakers and thus to enable communication.

There is of course an alternative explanation of how the sharing of reference / secondary intension is achieved in a linguistic community. The story told about the respective mechanisms told by externalists will, to some extent, be quite similar to the one I just gave. But they will deny that there is any need to invoke primary intensions. In their view, the way reference is fixed, and thus the way the desired coordination among the speakers is achieved, is dependent on factors external to the subject. They will insist that, for example, Kripke does not refer to Gödel when he says 'Gödel' because he associates the name with some metalinguistic property, but rather because his use of the name is linked with a causal chain leading to Gödel.

I do not aim to refute such an externalist position here. But still, my discussion in chapter 3 showed that the mere existence of a causal chain is not sufficient; without the subject's intention to defer to others, co-reference will not be established. The externalist thus cannot maintain that the coordination is achieved purely by causal or other external factors. A two-dimensionalist can therefore at least offer a simpler explanation of how

co-reference is coordinated. In any case, it has become clear that she can uphold the thesis that primary intensions play a crucial role in facilitating communication even in the case of names and generally deferential uses of linguistic expressions, to which the semantic thesis does not apply.

4.2.2.2.2 Natural kind terms

Take a case where two subjects associate a natural kind term with completely different primary intensions. This seems clearly possible. What is less clear is how often such cases actually occur and whether we should say that the two subjects share the relevant concept. But be that as it may, as I pointed out above communication is still possible between these subjects, as long as they refer (rigidly) to the same kind. In view of this fact, it is hard to see what primary intensions are supposed to contribute to the success of such communication. My aim in the following will be to show that, like in the case of names, primary intensions serve communication involving natural kind terms by helping to coordinate co-reference.

One might wonder how co-reference is secured in cases where the primary intensions differ dramatically. It may of course be due to a mere coincidence. But in general the possibility of communication should not depend on coincidences. So is there any way for potential communication partners to make sure that they refer to the same kind, if their theories of that kind are completely different? One could be tempted to argue that since this seems hard, one should expect that primary intensions are usually approximately shared among speakers, ideally due to a linguistic convention. The current problem therefore once more underlines the usefulness of shared primary intensions for communication. Nevertheless, I pointed out above that with respect to many natural kinds, there is plausibly no folk theory which is shared among speakers. So here is an alternative proposal as to how co-reference can also be achieved which bears some resemblances to what I said above about names: It might be that when a person acquires a natural kind term, she will initially use the term

deferentially. But when she becomes better acquainted with the kind, thereby gaining some confidence that it is this kind which others refer to when they use the term etc., she develops a 'theory' of the kind and may at some point no longer be disposed to defer to others' use. Now when another member of the relevant linguistic community does the same, the theory she will develop may be very different, depending on the natural kind and the circumstances. In such a case, it would be no mere coincidence that the subjects in question co-refer, even though the primary intensions they associate have, in the extreme case, nothing in common. Here, the role played by primary intensions in facilitating communication is a bit more subtle than in the previous cases, but still significant. By deferring to other members of her linguistic community, the subject is able to co-refer with them until she has gained sufficient knowledge about the kind, which includes knowledge about the linguistic behavior of others with respect to that kind. At this point, these latter (empirical) beliefs secure co-reference to a sufficient degree, such that the subject need no longer defer to others.

The upshot of the foregoing discussion is that a two-dimensionalist can maintain that primary intensions play a crucial role in enabling communication even in cases where they are not shared between speaker and hearer and to which the semantic thesis therefore does not apply. This is because at least in many such cases, primary intensions serve to secure co-reference among the communication partners, which is a necessary condition for successful communication.

In this chapter, I considered two ways to argue for the need of primary intensions. In section 4.1 I discussed the idea that, as Jackson puts it, primary intensions define the subject of any kind of inquiry. This claim can be divided into a semantic thesis and an epistemic one. According to the semantic thesis, any significant change in the primary intension associated with one of the key terms of our inquiry amounts to a change of subject. The epistemic thesis says that we need primary intensions to determine the subject, since they guide our judgments in the face of evidence.

I argued that the semantic thesis fails: At least in some cases, changes of the associated primary intensions are compatible with constancy concerning the subject matter. The epistemic thesis, on the other hand, can be defended. On the assumption of an attractive thesis about our ability to make judgments in the face of evidence, expressed by (CJ), there are indeed good reasons to believe that we need primary intensions to account for this ability.

In section 4.2, I discussed the claim that primary intensions are necessary for successful communication. I again distinguished a semantic thesis – that communication involves the transmission of primary intensions from speaker to hearer – and an epistemic thesis – that without primary intensions, language could not play the role it does play in the passing of information between subjects. Since the possibility of intersubjective variation in primary intensions showed that the semantic thesis fails to apply to all communicative situations, I proposed to modify it. On the resulting account, primary intensions are typically communicated from speaker to hearer. This view is supported by the importance of mutual inquiries in our epistemic practice, which would be greatly hindered if primary intensions were not typically shared among speakers, and thus transferred from speaker to hearer.

Finally, I argued that one can defend the epistemic thesis independently of the semantic thesis. In particular, I showed that primary intensions can play a crucial role in enabling communication even in cases where they are not transmitted from speaker to hearer.

I admit that none of these considerations proves that linguistic expressions are associated with primary intensions. An opponent may try to invoke alternative explanations of the phenomena I discussed, such as our ability to make judgments in the face of empirical evidence or the role of language in passing information, which do not involve the presence of primary intensions. However, I do not think this will be easy to do. In any case, unless alternative, or even better, explanations have been given, these phenomena lend support to an account which posits primary intensions.

5 Epistemic transparency and epistemic opacity

In the preceding chapters, I defended the central two-dimensionalist claim that every linguistic expression which is a candidate for having an extension has a primary intension: In chapter 3 I showed that this thesis can be upheld even with respect to proper names and natural kind terms, which are often believed to have no a priori implications. In chapter 4, I discussed more general arguments in support of primary intensions and argued that they play important roles in the epistemic determination of one's subject matter and in communication. Now it is surely a crucial step in a defense of conceptual analysis to establish the thesis that words and sentences are indeed associated with primary intensions. But still, even if that much has been achieved, the usefulness of conceptual analysis as a method in philosophy (and elsewhere) seems highly dependent on issues concerning epistemic transparency and epistemic opacity, which will be discussed in this chapter.

According to the analyses I gave of names, natural kind terms and any terms which are deferentially used, these terms are not semantically neutral, i.e. their primary intensions differ from their secondary intensions. This entails that they are epistemically opaque: the (metaphysically) necessary and sufficient conditions for membership in the respective categories are not accessible a priori. Since the notions of epistemic transparency and semantic neutrality will be at center stage in this chapter, it will be useful to explicitly define them:

(SN) An expression is **semantically neutral** iff its primary intension coincides with its secondary intension.

(ET) An expression is **epistemically transparent** iff its secondary intension is accessible a priori.

Epistemic opacity should then be considered as the negation of epistemic transparency, i.e. an expression is epistemically opaque if and only if its

secondary intension is *not* accessible a priori. The relation between (SN) and (ET) will be spelt out in detail in 5.2.

Epistemically opaque terms are at least potentially problematic for the prospects of a conceptual analysis. Since their secondary intensions and thus their metaphysical (or secondary) application conditions are not accessible via the possession of the relevant concepts, it seems questionable if anything of philosophical interest can be learned from them. Another, related problem is that there is often intersubjective variation in the primary intensions of epistemically opaque terms, as we saw in chapters 3 and 4. As I pointed out at the end of chapter 3, such intersubjective variation is a genuine problem for conceptual analysis, understood as a public enterprise. The question of what can be learned from (an analysis of) epistemically opaque terms will be addressed in section 5.3. But first, in 5.1, I will discuss and ultimately reject arguments from Ruth Millikan, Hilary Putnam, and Hilary Kornblith, which are supposed to show that epistemic opacity is ubiquitous. Then, in 5.2, I will deal with the question of whether and how one can determine which kinds of terms are epistemically opaque and which are epistemically transparent.

5.1 Arguments for ubiquitous opacity

Most philosophers agree that the arguments from Kripke, Putnam and others have shown that proper names, natural kind terms and indexical expressions are epistemically opaque. However, Putnam's claim that the same kinds of considerations apply to "the great majority of all nouns, and to other parts of speech as well" (Putnam 1975, 242) has met with considerably less approval. I nevertheless do not think that one can safely consider epistemic opacity as a phenomenon which is clearly limited in scope. For even though Putnam's own attempts to back up his claim may not be entirely convincing,[96] other authors, such as Tyler Burge, Millikan

[96] I will say more about this issue in 5.1.2.

and Kornblith, have brought forth different arguments to show that many other kinds of expressions exhibit opacity as well. I should note that these philosophers typically deny that the relevant expressions have Fregean senses at all, and would therefore most likely reject the idea that they have primary intensions. But since the existence of primary intensions has been defended in chapters 3 and 4, I will leave these issues aside here and focus on questions concerning epistemic opacity.

5.1.1 Millikan

Millikan is probably the most radical of the authors just mentioned. She herself notes that her view goes beyond that of Putnam, Burge, and others (cf. Millikan 2005, 67). Millikan thinks that 'empirical concepts', including even concepts such as 'square', are generally not associated with Fregean senses (cf. Millikan 2010, 50f.; cf. also Millikan 2005, 66). In her view, linguistic expressions are associated with conceptions. But these are very different from senses. A conception corresponds roughly to a collection of (empirical) ways of recognizing the concept's referent. Conceptions neither determine the reference, nor are they generally shared among speakers (cf. Millikan 2005, 67–71). For Millikan, possession of a concept is essentially constituted by a subject's ability to distinguish the referent. This ability need not be infallible; in fact, it can even be based on a mechanism which is only reliable in the subject's environment. Consequently, for two persons to share a concept, and to be able to communicate successfully, they only need to be able to distinguish the same referent. As Millikan puts it, all that is required is agreement in judgment, not in the method(s) of recognizing the referent (cf. Millikan 2010, 56f.).

As a general model of what it takes to possess and to share a concept, Millikan's account seems highly counterintuitive. It may have some plausibility for some kinds of concepts, such as natural kind concepts; but not for concepts such as 'square'. Take for instance a subject who lives in an environment where most (or maybe all) square things are red. Due to her excellent color vision, she can recognize these reliably. By applying the

term ‘square’ to all and only red things, she may even agree in judgments with others and thus, on Millikan’s account, communicate with them. But this person may be unable to recognize shape properties; she may have no idea what shapes are or, somewhat less extremely, she may not know this particular kind of shape. I think that in such a case, we would have no inclination to say that this person possesses the concept ‘square’. (It is a lot more plausible that she possesses the concept ‘red’.)

But be that as it may, a two-dimensionalist is not committed to any particular thesis about the individuation of concepts, i.e. about what it takes for two persons to share a concept (as I pointed out in chapter 3). What she is committed to, though, is that the properties a subject associates with a concept determine the reference with respect to any world considered as actual. As I said above, I will not discuss this thesis again in this chapter. Here, I will try to establish two theses which are also important for a vindication of the method of conceptual analysis: Namely, that at least in the case of some kinds of terms, these associated properties a) are shared between speakers; b) determine the term’s secondary intension independently of features of the actual world.

Both of these claims are rejected by Millikan. However, as she herself concedes sometimes defining features are “passed on explicitly from generation to generation” (Millikan 2005, 71), for instance in the case of terms for geometrical figures. Now if these features determine the terms’ secondary intensions, as is indeed very plausible, then why should one not say that such terms are epistemically transparent terms whose primary intensions are furthermore shared among the speakers of a linguistic community? I think that as it stands, Millikan’s view is not very plausible. At least some of the concepts to which her account is supposed to apply are by all appearances epistemically transparent. If one wants to hold that epistemic opacity does not just concern natural kind terms, names and

indexicals, then one should give some positive arguments to back up this claim.[97]

5.1.2 Putnam

An obvious way to support the idea that epistemic opacity does not only concern names, natural kind terms and indexicals is to try and apply the same kinds of arguments which have been used to show that these terms are opaque to other kinds of expressions. Putnam thinks that 'Twin Earth style' considerations are applicable to many other kinds of terms, such as artifact terms. Recall the following key insights of his about natural kind terms: Firstly, these terms pick out the nature of the corresponding kind rigidly. Secondly, we do not have a priori insight into the essential properties of such kinds and thus into the (secondary) application conditions of the terms. Furthermore, Putnam argues that typically nothing we associate with natural kind terms is a priori.[98] Now, Putnam's remarks about artifact terms exactly parallel those about natural kind terms, as in his famous discussion of the term 'pencil': According to Putnam, it could turn out that pencils are not artifacts, but living organisms:

> We cut them open and examine them under the electron microscope, and we see the almost invisible tracery of nerves and other organs. We spy upon them, and we see them spawn, and we see the offspring grow into full-grown pencils. (Putnam 1975, 242)

---

[97] For the sake of fairness, I should add that Millikan does make a positive case for her view, which is mainly based on the Burgean idea that all or nearly all kinds of concepts can be possessed despite the subject's incomplete understanding. Since Kornblith's argument, which will be addressed below, is based on this idea as well, I will postpone a discussion of this issue.

[98] In this respect, Putnam's view seems to be stricter than that of Kripke who holds that at least in some cases, possession of a natural kind concept requires a priori knowledge of a reference-fixing description (cf. the discussion of 'light' and 'heat' in Kripke 1980, 131).

The hypothetical scenario suggests that not even the fact that pencils are artifacts can be known a priori. From this, Putnam infers that the term 'pencil' has no a priori associations. Nevertheless, he does not think that it is metaphysically possible that pencils are organisms. In his view, if the things we call 'pencils' are in fact, as is very likely, artifacts, then they are necessarily artifacts. This is because "[w]hen we use the term 'pencil', we intend to refer to whatever has the same *nature* as the normal examples of local pencils in the actual world." (Putnam 1975, 243) The upshot of Putnam's considerations is that there is no difference in the semantics of natural kind terms and artifact terms. Both kinds of terms refer rigidly to a hidden nature which is not a priori accessible.

Stephen Schwartz was one of the first to criticize Putnam's account of artifact terms (cf. Schwartz 1977, 1978, 1980). He conceded that Putnam's 'pencil' example[99] shows that it is epistemically possible that pencils are not artifacts and thus that it is not a priori that they are. But this by no means shows that 'pencil' has no priori associations or that it is not synonymous with any associated description. It only shows that 'being an artifact' is not part of the correct description (cf. Schwartz 1978, 569). Something similar holds for many other artifact terms, such as 'salt shaker', 'lamp' and 'screwdriver'. It is not a priori that these are artifacts. But there are other features which are more plausibly a priori associated, for example those specifying the function and maybe the form of these items.

Admittedly, it is often not easy to spell out such a priori associated characteristics precisely. But this is just an instance of the general problem that it is often difficult to give definitions of the expressions we use.[100] The more relevant question in the current context of epistemic transparency and opacity is whether we can determine the extension of such terms in counterfactual scenarios a priori. This is the case if their secondary intensions do not vary with contingent empirical facts. In my view, the second part of Schwartz's critique of Putnam in which he invokes a

[99] According to Putnam himself, he borrowed the example from Rogers Albritton (cf. Putnam 1975, 242).

[100] I will discuss this issue in some detail in chapter 7.

hypothetical case bears precisely on this issue: Suppose that every pencil in our environment, or maybe even every pencil in the universe, is an artifact. According to Putnam, this would imply that pencils have an 'artifactual nature' and are thereby necessarily artifacts. But this seems implausible, since it should nevertheless be possible that people start to grow pencils (cf. Schwartz 1978, 570). Elsewhere, Schwartz discusses the following case: Suppose, by analogy with Putnam's story about cats, that all actual salt shakers turned out to be spying devices from Mars. If we then started building small cylindrical objects with small holes at the top and used them to dispense salt, these would nevertheless be salt shakers (cf. Schwartz 1983, 479, fn. 6).
Schwartz concludes that, against Putnam, natural kind terms function very differently than artifact terms, which, following Locke, he takes to be 'nominal kind terms'. While the former ones thus refer to a hidden nature which can only be revealed empirically, the latter ones have application conditions which are accessible to a subject via the possession of the relevant concepts.

In my view, Schwartz' critique of Putnam is essentially correct. My case for the claim that typical artifact terms like 'salt shaker' or 'chair' are epistemically transparent will in part be based on considerations similar to the ones made by Schwartz. But before I set forth my own line of argument, let me sketch Kornblith's argument for ubiquitous opacity, which is based on considerations quite different from the ones of Putnam just outlined.

5.1.3 Kornblith

Kornblith is unimpressed by Schwartz' arguments. He insists that the semantics of artifact terms is no different from that of natural kind terms (cf. Kornblith 1980, 2007b). To show this, he appeals to a line of argument which is already present in Putnam's and Kripke's writings and which was developed further by Burge: Kornblith argues that successful reference in

spite of incomplete understanding, which figures in the argument from Ignorance and Error, does not only occur in the case of names and natural kind terms, but also of artifact terms.[101] In his view, this shows that the reference (and the content) of artifact terms is not determined by speaker associations, either. Consequently, the possession of such concepts does not give a subject access to their (secondary) application conditions.

Kornblith believes that he can nevertheless account for the difference in our intuitions about hypothetical scenarios involving natural kind terms such as 'water' and gold' on one side and 'pencil' and 'salt shaker' on the other. In his view, the difference in our verdicts about cases does not indicate a difference in the semantics of these kinds of terms, but rather in our empirical acquaintance with their referents. He thus admits that subjects usually have better insight into the essence of salt shakers and pencils than of many natural kinds, which is reflected in our judgments about hypothetical cases. But as the possibility of incomplete understanding shows, this is not necessarily so. It is therefore not the mere possession of a concept which provides the subject epistemic access to the nature of the denoted category, but rather empirical knowledge. In other words: Artifact terms are epistemically opaque. Accordingly, Kornblith holds that arguments from Ignorance and Error show that there is no (epistemically) relevant semantic difference between natural kind terms and artifact terms.

There is little reason to think that the same line of reasoning cannot be applied to many other kinds of terms as well. As was already argued by Putnam and especially by Burge, incomplete understanding is a ubiquitous phenomenon which concerns many different kinds of terms. In Kornblith's case, it is natural to suspect that he does not take the relevance of his result to be restricted to artifact terms, either. To support this suspicion, let me quote a remark from a paper of his on philosophical methodology: "I see the investigation of knowledge, *and philosophical investigation generally*, on the model of investigations of natural kinds" (Kornblith 1998, 134, my

[101] I already gave two examples of artifact terms where this will plausibly be the case for a great number of speakers in chapter 3, namely 'MRI' and 'transistor'. Kornblith also lists a couple of such terms, for example 'rheostat' and 'buckboard'.

emphasis). Now, although Kornblith does in fact hold that knowledge is a natural kind (cf. e.g. Kornblith 1998), it would be odd to assume that all philosophically relevant terms refer to natural kinds (or artifacts). A more charitable explanation of the remark just cited is that Kornblith believes that his argument concerning the semantics of artifact terms generalizes and can thus be used to establish the ubiquity of epistemic opacity. In its most ambitious form, it would go like this: All kinds of terms can be subject to incomplete understanding. Therefore, reference (and content) is never determined by speaker associations, which implies that there are no epistemically transparent terms. This conclusion could then in turn be used to argue that conceptual analysis is generally useless. In fact, this might be precisely what Kornblith, who is a well-known critic of conceptual analysis, has in mind.

One should note that when a concept is epistemically opaque, this need not imply that it is ineligible for any kind of conceptual analysis; I will say more about this issue in 5.3. Nevertheless, if all terms, or nearly all terms, were opaque, this would be worrisome for conceptual analysis. Furthermore, although it is not completely obvious that successful reference despite incomplete understanding is possible for all terms, I concede that it can concern a great number and variety of terms, including philosophically relevant ones. So I think if one aims to defend conceptual analysis against the objection from incomplete understanding, one has to reject the step from the premise that a subject can refer by using a term which she understands incompletely to the conclusion that that term is epistemically opaque. Prima facie, this seems difficult: In chapter 3, I argued that the argument from Ignorance and Error does not show that such concepts have no primary intensions and proposed to consider them as deferential. But the analysis I gave there showed that the primary intensions of deferential concepts differ from their secondary intensions; they are thus epistemically opaque.

One way in which a proponent of conceptual analysis might want to attack Kornblith's argument is by attempting to show that there has to be a difference in the semantics of natural kind terms and artifact terms, since

there must be speakers with complete understanding of an artifact term. It will be instructive to consider why such an argument fails. After that, I will set forth my own argument, in which I concede that artifact terms, and other kinds of terms as well, can be subject to incomplete understanding to the same extent as, say, natural kind terms.

The idea of Putnam and others is that in the case of natural kind terms, nature does the linguistic labor. When someone introduces such a term, she only needs to be acquainted with an exemplar of a kind and intend to refer to the particular natural kind exemplified. In the case of artifact terms, however, nature will hardly do the linguistic labor. This suggests that when a person has an insufficient grasp of the category denoted by a term, there have to be other members in her linguistic community – the 'experts' – with complete understanding to whom she can defer and who thus do the linguistic labor for her. But in fact, this need not be so, as an example invoked by Kornblith (cf. Kornblith 2007b, 146) which I modified slightly shows: Suppose that archaeologists find a number of artifacts from a long extinct culture which they take to belong to a common kind and introduce a term for that kind of artifact.[102] Such a case, in which there would be no-one in the linguistic community who knows the metaphysical conditions for membership in the relevant kind, seems clearly possible. It is therefore not necessary that there are subjects with complete understanding in the case of artifact terms, either.

Admittedly, the account I just gave of the introduction of the artifact term is a bit too simple. For instance, the ancient artifacts discovered may have had many different functions and thus exemplify more than one kind of artifacts. Therefore, to secure reference to a particular kind of artifact, it will not do to point at some exemplars. So prima facie, there has to be at least one person, namely the one who introduces the term, who knows more about the category denoted by the term after all.

In the case of natural kind terms, there is a very similar problem, which Devitt & Sterelny call the 'qua problem' (cf. Devitt & Sterelny 1999, 90–93): A particular specimen of a natural kind usually exemplifies many

[102] In Kornblith's example, the term introduced refers only to a single artifact.

different kinds; for example a eukaryote, a carnivore, a mammal, a tiger, … Accordingly, it has to be specified to which kind the term is to refer in the case of natural kind terms, too. And therefore, one cannot establish that there is a semantic difference between natural kind terms and artifact terms with an appeal to the problem just mentioned, anyway. Furthermore, this problem only concerns the introduction of a term. But the person who once introduced the term may have long forgotten what she intended the term to refer to or may no longer be alive. And maybe most importantly, even if there is someone whose knowledge or understanding of the term is sufficient to secure reference without deferring to others, the term can still be epistemically opaque for her.
A lesson to be drawn from these considerations is that any semantic difference between natural kind terms and artifact terms which one can hope to find will be a contingent one: No such difference can be derived from the difference in the metaphysical status of their referents alone.

In the following, I will argue that despite the fact that incomplete understanding can concern artifact terms to the same extent that it concerns natural kind terms, there is an epistemically relevant semantic difference between terms like ‘water’ and ‘gold’ on one hand and ‘salt shaker’ and ‘pencil’ on the other. The simple idea I want to defend is that the latter ones, along with many other terms which can be subject to the division of linguistic labor, are epistemically transparent for those with complete understanding of the relevant terms. It is clear that Kornblith’s argument from the possibility of Ignorance and Error does not undermine this claim. In fact, the phrase ‘incomplete understanding’ already suggests that there can also be complete understanding. I noted in chapter 3 that it is hard to say what complete understanding could mean in the case of natural kind terms. In the case of artifact terms, I think it is intuitively more plausible that they have a one-dimensional, i.e. semantically neutral, meaning which just need not be known to every speaker. In the following, I will yield some support to this idea.

Generally speaking, in order to show that a term is epistemically transparent, one has to establish two things: firstly, that it has a primary intension; secondly, that its reference with respect to counterfactual worlds does not vary with contingent features of the actual world. In my view, hypothetical cases of the kind invoked by Schwartz are perfectly suited to show this. More specifically, I will argue in the following that, against Kornblith, the best explanation for the difference in our judgments about hypothetical scenarios involving terms like ‘gold’ and ‘water’ on the one hand and ‘salt shaker’ and ‘pencil’ on the other is that the latter ones, unlike the former, are epistemically transparent.

Consider again Schwartz’ discussion of the term ‘salt shaker’: His considerations suggest that it is epistemically necessary that salt shakers have a specific function (and maybe form). It seems inconceivable that objects which were not designed, never used and never even intended to be used to dispense salt and which are moreover not usable to do so are nevertheless salt shakers. Furthermore, Schwartz’ thought experiment about the spying devices from Mars suggests that such functional properties also reflect the term’s application conditions with respect to counterfactual scenarios: No matter what the things we call ‘salt shakers’ turn out to be, an object with the appropriate function (and maybe form) would still be a salt shaker. Kornblith does not dispute the correctness of these intuitive judgments about hypothetical cases. But he denies that they are a priori; rather, he says, they are empirically shaped. Most of us are closely familiar with salt shakers and know very well what salt shakers are and what they are for. On the other hand, we are often much less familiar with and have much less well-established beliefs about natural kinds. In Kornblith’s view, this is what explains the difference in our judgments about scenarios involving artifacts on the one hand and natural kinds on the other. There is thus no need to appeal to a difference in the semantics of the corresponding terms.

It is anything but clear, however, that empirical (background) beliefs really influence our conceivability judgments, or more specifically our judgments about worlds which are considered as actual or worlds considered as

counterfactual relative to one considered as actual.[103] Consider the following example: Every raven I have seen in my life was black. Moreover, I firmly believe that there are no green ravens. Nevertheless, I can easily imagine green ravens. Maybe Kornblith would object that theoretical beliefs have to be taken into account as well: For instance, I believe that having a specific color is normally not among the properties essential for membership in a species, so my (empirically-based) background belief that there could have been green ravens may explain why I can imagine them. So here is a different example: Every cat I have seen in my life was an animal and I firmly believe that all the other cats in the world which I have not seen yet are animals as well. Even more so, since I trust in Putnam's and Kripke's insights concerning natural kinds, I believe that cats are necessarily animals. Nevertheless, I can easily conceive of a scenario where cats turn out to be robots.[104]
It is thus hard to see how an appeal to a difference in our empirical acquaintance with the referents of these terms could explain the difference in our judgments about hypothetical cases. A much more plausible explanation is that there is a difference in the semantics of these terms: While the secondary application conditions of 'cat' are dependent on contingent features of the actual world, those of 'salt shaker' are not. Since it is very plausible that the same kinds of considerations apply to many other artifact terms such as 'pencil', 'washing machine' or 'screwdriver', I conclude that these kinds of terms are epistemically transparent.

Let me briefly sum up what has been shown in this section: None of the accounts I considered succeeds in establishing the claim that the phenomenon of epistemic opacity extends far beyond names, natural kind terms and indexicals. The most promising argument in support of such a view, which is based on the premise that many other kinds can be subject to the division of linguistic labor as well, does not undermine the idea that these kinds of terms are epistemically transparent for those with complete

[103] Cf. also my discussion of Schroeter's account in the previous chapter.
[104] Cf. Putnam's famous example in Putnam 1962, 660.

understanding. Furthermore, at least in the case of artifact terms such as ‘salt shaker’ or ‘pencil’, our judgments about hypothetical scenarios involving them are most naturally explained by the fact that they are epistemically transparent.
But as I just mentioned, this is only true for subjects with complete, or non-deferential, understanding of these terms. And while I think it is safe to assume that incomplete understanding of terms like ‘salt shaker’ or ‘pencil’ is quite untypical, one could still wonder more generally if and to what extent incomplete understanding presents a problem for conceptual analysis. I will discuss this question briefly at the end of this chapter. In the following, I will have a closer look at how one can generally determine whether an expression is opaque or transparent.

5.2 Revealing opacity

Since, as was just argued, there are epistemically transparent as well as epistemically opaque terms, it would be useful to get a grip on which (types of) expressions fall into which of these categories. The discussion in the previous section may suggest that while indexicals, names and natural kind terms are opaque, artifact terms are transparent. But in fact, the example about the ancient artifacts which I borrowed from Kornblith shows that this need at least not generally be so. Kornblith therefore rightly remarks that one cannot read off the semantics of a term from the metaphysical status of the category it denotes (cf. Kornblith 2007b, 145). But I think this observation cuts both ways. Natural kind terms are not always epistemically opaque, either; a very simple example is ‘H_2O’: It is both a priori and necessary that H_2O consists of two hydrogen atoms and one oxygen atom.[105] In the following, I will therefore propose two alternative ways to determine whether an expression or a type of expression is opaque or transparent.

[105] I am inclined to grant, however, that proper names and indexical expressions are generally opaque: A transparent name would at least be unusual, while indexicals are opaque pretty much by definition.

5.2.1 Revealing opacity via the function of a term

One way to approach the task of revealing opacity is to consider the function of a term. The main functions of proper names, for instance, seem to be firstly to secure reference to a specific individual which is stable even with respect to modal contexts, and secondly, to coordinate the reference with other speakers.[106] Since we usually do not know the essential properties of the individual referred to, it thereby becomes apparent why the secondary application conditions of names are opaque.

Similar considerations apply to natural kind terms. Often, these terms are introduced in the following way: We come across a number of objects, or phenomena, and come to think that these objects are belonging to a common kind, or that the phenomena are caused by a common kind. When we introduce a term for that kind, we need not, and often do not, know what is essential to it. So it is no surprise that many natural kind terms are epistemically opaque.

In chapter 3, I gave an outline of how stable and rigid reference is achieved in the case of names and natural kind terms. The basic idea is that we draw on contingent features of the individual or kind and then rigidify. In some cases, these are not 'objective' features of the referent, but ones which specify its relation to us – this is where centered contents come into the picture. In other words, the associated properties do not, and are not meant to, reflect such a term's secondary application conditions. Rather, secondary intensions depend on the nature of the individual or kind which actually has these properties.

There are many other terms, however, which serve very different purposes. Take artifact terms: Since artifacts are typically produced to fulfill a specific function, it makes sense to introduce terms which simply pick out any object serving, or being intended to serve, that function. The 'nature' of actual exemplars of the 'artifactual kind' is therefore not relevant for the applicability of such terms. In this respect, the term for the ancient artifacts in the example I discussed above is an exception. However, that term is

[106] I already talked about the way this coordination is achieved in 4.2.

introduced in a way which is similar to how I described the introduction of many natural kind terms: People come across a number of objects which they suspect belong to a common kind and introduce a term for that kind, without knowing what exactly unites these objects, i.e. without knowing what makes them samples of a common kind. This is hardly a typical way of introducing an artifact term. Often, artifact terms are introduced even before any artifacts of that kind exist. In these cases, no story akin to the one I told about the introduction of natural kind terms can possibly apply. Moreover, as I indicated above, natural kind terms do not always work the same way, either. Sometimes, for instance, the existence of a natural kind is postulated (or hypothesized) for theoretical reasons; arguably, essential properties of this hypothetical kind will then be a priori associated with the corresponding term, like (maybe) in the case of 'Higgs boson'. In other cases, like with 'H_2O', the expression is supposed to pick out the essence of a familiar natural kind.

I just argued that typical artifact terms such as 'salt shaker', 'pencil' and 'screwdriver' simply pick out objects with a certain function. This may be a rather trivial point, but I think it reveals an influential mistake made by Kornblith and maybe even more clearly by Putnam: They seem to assume that since referential expressions serve to refer to an object or a class of objects, their (secondary) application conditions must be dependent on the nature of actual exemplars of the relevant category. But there is no reason to assume this, and in fact there are a number of reasons to believe that many kinds of terms work quite differently.

To sum up: In this section, I proposed that considering the function(s) of a term can help determine whether it is opaque or transparent. In turn, to determine its function(s), it can be helpful to think about the possible circumstances of a term's introduction. It transpired that if one applies such an approach to natural kind terms and artifact terms, one is led to expect that 'typical' natural kind terms are opaque, while 'typical' artifact terms are transparent, which confirms the verdict from the previous section.

5.2.2 Revealing opacity via considerations about hypothetical cases

The verdict just mentioned about the semantics of terms like ‘salt shaker’ as opposed to that of terms like ‘cat’ was based on judgments about hypothetical scenarios. The difference in the behavior of transparent and opaque terms with respect to such scenarios thus yields a second way to identify them. This approach can be modeled very well in a two-dimensionalist framework. Consider the two-dimensional matrix for ‘water’:

‘water’	$w_@$	w_2	w_3
$w_@$ (WS: H_2O)	H_2O	H_2O	H_2O
w_2 (WS: XYZ)	XYZ	XYZ	XYZ
w_3 (WS: ABC)	ABC	ABC	ABC

(Figure 6)

The diagonal of the matrix, which represents the primary intension, is accessible a priori – via judgments of the form ‘If the world is such and such, then water is …’. So, as should be familiar by now, grasp of a term’s primary intension enables a speaker to determine its extension with respect to any world considered as actual, if provided with complete empirical information about that scenario. Now recall, from the introduction to two-dimensionalism in chapter 2, that the only reason why we sometimes cannot determine the extension of a term with respect to a counterfactual situation is because we lack knowledge of some relevant contingent features of the actual world. If, for instance, we know that the clear,

drinkable, … liquid in our rivers and lakes is H_2O, then we are in a position to know that the substance on Twin Earth would not be water.

If one conjoins these two ideas, one arrives at the view that we are able to access a term's (hypothetical) extension with respect to a counterfactual world relative to any world considered as actual. Here is an example of such a judgment: 'If the clear, drinkable, … liquid in our rivers and lakes is / had turned out to be XYZ, then the substance on Twin Earth would have been water.' In principle, by running through every counterfactual world relative to a specific world considered as actual, we are thus able to access the (hypothetical) secondary intension of an expression relative to any world considered as actual. One thereby arrives at judgments like 'If the clear, drinkable, … liquid in our rivers and lakes is XYZ, then water is necessarily XYZ' or 'If the realizers of the cat role are robots, then cats are necessarily robots' (where the cat role corresponds to the theoretical role associated with 'cat'). Accordingly, it is not just the diagonal of the matrix which can be accessed a priori; the whole matrix is a priori.[107] The only thing about that matrix which we cannot know a priori is in which of the rows we are located, i.e. which of the centered worlds on the left which can be considered as actual is in fact the actual world. Two-dimensionalism thus underpins the kind of judgments necessary to determine whether a given term is epistemically transparent: One hypothetically varies characteristics of the actual world to see whether this affects a term's extension with respect to counterfactual scenarios and thus its secondary intension. If it does, then the term is epistemically opaque, if it does not, it is transparent. This is exactly in line with the way Schwartz established the thesis that there is a difference in the semantics of 'pencil' and 'salt shaker' on the one hand and 'water' and 'cat' on the other: By reference to (judgments about) hypothetical cases, he showed that while the extensions of the latter ones are dependent on the 'nature' of actual exemplars – and

[107] In Chalmers' terminology, the matrix is the 'two-dimensional intension'; formally speaking, the two-dimensional intension is a function from pairs of worlds considered as actual and worlds considered as counterfactual to extensions. In Chalmers 2006, 102f., Chalmers also claims that the two-dimensional intension is accessible a priori.

thus on contingent features of the actual world –, those of the former ones are not.

5.2.3 Can opacity be determined a priori?

The foregoing considerations suggest that it is generally a priori whether a term is transparent or opaque. But this view faces potential counterexamples. Take the case of 'jade', as it is often told in the philosophical literature:[108] People once thought that jade is a natural kind. But then it was discovered that what is called 'jade' are actually two different, though superficially similar minerals, jadeite and nephrite. Although the jade role, i.e. the cluster of superficial characteristics shared by jadeite and nephrite, and those associated with 'jade', is, at least in our environment, only realized by two substances, there could have been many more substances which shared these features. It seems plausible that these substances would be jade, too. If one has doubts about this, one can change the example slightly: Suppose that it had turned out that the jade role is actually realized by a large number of substances and mixtures of substances. In this case we would surely be inclined to call anything 'jade' which satisfies the jade role.

These considerations suggest that 'jade' is (or could have easily been / turned out to be) a purely functional term, picking out anything which realizes the jade role in every possible world. If, however, it had turned out that what we call 'jade' is actually a natural kind, the term would have rather behaved like a typical natural kind term and picked out the same substance with respect to every world.

Although the case of 'jade' is surely unusual, I nevertheless think it is anything but unique. Take for instance 'air': For a long period in human history, air was believed to be an element. One can reasonably assume that if this had turned out to be correct, 'air' would have turned out to refer

[108] This story is most likely not very accurate, so one should take the case to be partly hypothetical.

rigidly to that element. But in fact, air is just a mixture of gases in varying proportions. Plausibly, ‘air’ picks out anything which plays the air role with respect to every possible world – where ‘the air role’ stands for the theoretical role which is a priori associated with ‘air’. Let me illustrate this with a two-dimensional matrix:

‘air’	$w_@$	w_2	w_3
$w_@$ (AS: mixture of gases)	mixture of gases	A	AB
w_2 (AS: A)	A	A	A
w_3 (AS: AB)	AB	AB	AB

(Figure 7)

(Here, ‘AS’ abbreviates ‘airy stuff’, in analogy with Chalmers’ ‘watery stuff’, and stands for whatever plays the air role in the relevant world. In $w_@$, the air role is played by a mixture of gases, in w_2 by the element A, and in w_3 by the molecular compound AB.)

As I just said, the secondary intension of ‘air’ picks out whatever plays the air role in every world considered as counterfactual. The term is thus semantically neutral, i.e. its primary and secondary intension are equivalent. Terms which have this feature are often considered to be epistemically transparent. But as the matrix shows, there is an important difference between ‘air’ and other semantically neutral expressions: If, for

instance, we consider w_2 as actual, we see that relative to this scenario, the secondary intension picks out the element A in every world considered as counterfactual, including those where A fails to play the air role. So relative to w_2 considered as actual, the primary and the secondary intension of 'air' are not equivalent. This mirrors the fact that if the occupant of the air role had turned out to be an element, the term would have behaved like a standard natural kind term. It is thus epistemically contingent that 'air' is semantically neutral. And this seems to imply that it is not a priori that 'air' is epistemically transparent, in violation of the thesis formulated above.

One possible way to react to the examples I just gave is to say that they involve meaning change. When it turned out, for example, that the stuff called 'air' is not an element (or some other natural kind), it really turned out that the term 'air' is empty. But since it would have been impractical not to have a term for the stuff surrounding us, the meaning of 'air' changed, such that it now refers to the mixture of gases.

It is not easy to argue against such a view. But I think that generally, the default assumption should be that the meaning of an expression remains constant. Admittedly, there are pragmatic reasons to have a term for the stuff surrounding us, but these are likewise reasons to associate the term with a primary intension which does not make the existence of air dependent on it being a natural kind. Furthermore, the cases at hand are at least compatible with the idea that the primary and the secondary intension (and the whole matrix) remains constant, in case of which there would be no reason to say that there was a change in meaning. So even if one holds that the primary or the secondary intensions of 'jade' and 'air' changed when the relevant discoveries were made, there is little reason to think that there are no such terms for which it is not a priori whether they are semantically neutral.

On closer inspection, these examples do nevertheless not really show that epistemic transparency is sometimes only empirically evaluable. To demonstrate this, let me first note that they cannot speak against the idea that we are able to access the whole matrix a priori. If anything, they presuppose precisely this ability – in particular the ability to determine the

extension of an expression with respect to a world considered as counterfactual relative to a world considered as actual. So let me have a closer look at what the cases do show, using the example of ‘air’: Recall that the air role is the theoretical role associated with ‘air’. It is thus a priori that air plays the air role. Since the term’s secondary intension coincides with its primary intension, it is also necessary that air plays the air role. However, as we have seen it is not a priori that air necessarily plays the air role: If for instance w_2 had turned out to be actual, it would have been (metaphysically) possible that air does not play this role. So, although in some loose sense one may say that we know the nature of air a priori, we do not know a priori what the nature of air is. For all we know a priori about air, it could also be essentially A, or AB, or ...

Consequently, since we do not have a priori insight into the secondary application conditions of ‘air’, we should say that the term is not epistemically transparent, despite the fact that its primary and secondary intension are equivalent. Accordingly, semantic neutrality does not entail epistemic transparency. It is only those terms whose secondary intensions coincide with their primary intensions with epistemic necessity which are epistemically transparent.

I conclude that terms such as ‘jade’ and ‘air’ are in fact epistemically opaque. Since it is also a priori that they are opaque, they present no counterexample to the thesis that we are able to determine a priori whether a term is transparent or opaque. What we cannot say a priori of them, however, is whether they are semantically neutral. It was for example only the empirical discovery that air is not a natural kind which enabled us to recognize that the secondary intension of ‘air’ coincides with its primary intension. This entails that in some cases, we cannot say a priori whether an a priori associated property is also a metaphysically necessary property of the denoted category.

This result has implications for the question of our epistemic access to the general modal status of a statement. Kripke holds that although we cannot know the truth-values of many necessary statements a priori, we can still

know a priori that whatever truth-value they have, they have it necessarily (cf. Kripke 1980, 109). If one adds to this the claim that we can also say of contingent statements that they are contingent, i.e. either contingently true or contingently false, one arrives at the thesis that we can generally determine the modal status of a statement a priori.[109] Now consider the sentence 'Air plays the air role': From what I argued above, it follows that the sentence is a priori and expresses a necessary truth. But we cannot know a priori that it expresses a necessary truth, nor can we know a priori that if it is true, it is necessarily true. As I already pointed out above, for all we can say a priori, w_2 could be actual, in which case the statement would be contingent.

Further counterexamples to the thesis that the general modal status of a sentence is determinable a priori can easily be generated. Take for instance 'If air exists, then A exists'.[110] That statement is contingently false, but for all we know a priori, it could be necessarily true – for instance in case w_2 turned out to be actual. So the thesis that we can generally determine the modal status of a statement a priori is false. In fact, that thesis is closely related to the claim that we can say a priori whether an expression is semantically neutral, so it should not come as a surprise that they fall together.[111]

---

[109] The phrase 'general modal status' was introduced by Albert Casullo, who discusses, and seems inclined to accept, this thesis in Casullo 2003 and Casullo 2010. For an explicit defense of the thesis cf. Horvath 2009.

[110] Since the term 'water' works similarly – cf. below –, one can also use the following, somewhat nicer example: 'If water exists, then oxygen exists.' That sentence is necessary, but if water had turned out to not be a natural kind (or even a different natural kind), it would have been contingent.

[111] A sentence is semantically neutral if it is either epistemically and metaphysically contingent or epistemically and metaphysically necessary. Since it is plausible that we can determine a priori whether a sentence is epistemically contingent, any sentence of which one cannot say a priori whether it is semantically neutral is such that we cannot say a priori whether it is metaphysically contingent or necessary. But this is just to say that its general modal status is not determinable a priori.

There are most likely further terms of which we cannot say a priori whether they are semantically neutral. For example, there are plausibly at least some terms which are in fact natural kind terms, but which could have turned out to be not. Take 'water': It is natural to think that if Lavoisier had found out that the stuff he investigated is just an inhomogeneous mixture of many different substances, he would not have thereby discovered that there is no water. Rather, in this case the term 'water' would have been applicable to anything playing the water role. It is therefore epistemically contingent that the secondary intension of 'water' differs from its primary intension – the term could have turned out to be semantically neutral.
Nevertheless, it would be wrong to think that all of our terms work like this and that therefore, there are no epistemically transparent terms. There is a large number of semantically neutral terms whose neutrality is not dependent on contingent features of the actual world and which are thus epistemically transparent.[112] Schwartz' discussion of terms such as 'pencil' and 'salt shaker' illustrates this nicely. Furthermore, as I pointed out above, it is also plausible for reasons independent of the evaluation of hypothetical scenarios that there are many terms concerning which the 'nature' of actual exemplars is just irrelevant for their applicability. If so, then despite the counterexamples given, there are still many epistemically transparent terms, i.e. semantically neutral terms whose neutrality can be recognized a priori.

So far, I argued that there are both epistemically opaque and epistemically transparent expressions, which raised the question of whether and how we can decide into which of these categories a given term falls. I defended the view that even though it is sometimes an empirical question whether an expression is semantically neutral, one can still tell a priori whether it is transparent. To put it briefly, an expression is epistemically transparent if it

[112] Likewise, there are presumably many natural kind terms such that if it turned out that there is no such natural kind, then the term is empty.

is semantically neutral with epistemic necessity.[113] This leaves open the following two questions, both of which are highly important with respect to the prospects of conceptual analysis: Firstly, on which side of the divide can we expect to find those terms which are (most) relevant for philosophical inquiry? Secondly, what can be learned from opaque terms or, to put it differently, what can result from a conceptual analysis of such terms?

I will just say a few words about the former question, because I do not think it has a definite answer: Bealer claims that most philosophically interesting concepts can be known determinately (cf. Bealer 1998), which roughly corresponds to the view that they are transparent if understood completely.[114] However, I suspect that philosophically interesting expressions form a very inhomogeneous class, presumably even less homogeneous than natural kind terms and artifact terms. So although I suppose that many terms which have been subject to conceptual analysis, such as 'free will', 'knowledge', 'truth' or 'good'[115], are indeed transparent, Bealer's claim that this is true for most philosophically interesting expressions is quite bold and also hard to evaluate. In any case, one cannot reasonably hope that all terms which might be of philosophical interest are transparent. Therefore the question of what can be learned from opaque terms, which will be addressed in the following, is definitely significant for conceptual analysis.

---

[113] Let me add the following reservation, though: Intuitions about possible cases and thus also about opacity can sometimes be inconclusive. Some terms are probably borderline or ambiguous between being transparent and being opaque (as Lewis seems to suggests for 'water' (cf. Lewis 1994, 424)). I nevertheless believe that most terms can be conclusively classified as either opaque or transparent.

[114] A similar view is held by Alvin Goldman (cf. Goldman 2007).

[115] The idea that moral terms are transparent is contested, though (cf. e.g. Sayre-McCord 1997).

5.3 The value of opaque terms in conceptual analysis

When a term is epistemically opaque, one cannot determine its secondary, i.e. metaphysical application conditions a priori. Does this mean that any kind of conceptual analysis is futile if it is concerned with such terms? As I argued mainly in chapter 3, opaque terms nevertheless have primary intensions. These a priori associations could serve as the basis for a conceptual analysis, even if they are insufficient to give us access to the associated secondary intensions. On the other hand, I also suggested, in chapter 3, to take the primary intension of the opaque term 'water' as comprising a theory about the natural kind water and proposed that at least one of its key functions is to secure reference to that natural kind. One might therefore suspect that, at least insofar as other opaque terms are similar, the primary intensions of such terms themselves are not particularly important.

But these are quite abstract considerations, and their implications for conceptual analysis are not obvious. Take an epistemically transparent expression such as 'the most massive organism in the world': This expression will pick out the most massive organism, whatever it is, in every world considered as actual and in every world considered as counterfactual. Rigidifying that expression yields 'the actually most massive organism in the world'. That expression is opaque; in order to determine its reference with respect to worlds considered as counterfactual, one needs to have empirical information about the character of the actual world. Nevertheless, it is hard to see why (the primary intension of) this latter expression should contain less information than (that of) the former one. On the face of it, anything which can be extracted from the unrigidified expression can also be extracted from the rigidified one. Similar considerations apply to other opaque expressions: Assume for example that natural kind terms like 'water' are, as was argued in chapter 3, rigidified theoretical terms. Why should such a term be less interesting for conceptual analysis than a corresponding term which picks out, say, the respective occupant of the water role in any metaphysically possible world, instead of rigidly picking

out the actual occupant of the water role? In each case, the same theoretical role can be carved out by an analysis of the term's primary intension. So prima facie, if one takes the aim of conceptual analysis to be an analysis of an expression's primary intension, then it need not matter whether that term is transparent or opaque.

Things are not quite that simple, however. Firstly, in many cases we are interested in a category's 'metaphysical essence' which is expressed by the secondary intension of the corresponding expression. And secondly, if the main function of the primary intensions of opaque terms is indeed to secure reference to the denoted category, this may suggest that they are typically not shared among speakers. Since both of these considerations put the value of the conceptual analysis of opaque expressions into doubt, I will address them in the following.

5.3.1 Discovering essences

Let me start with the former of the considerations just mentioned. If, for instance, one aims to reveal the nature of water, an analysis of the term 'water' alone is obviously insufficient. So the fact that 'water' is opaque at least diminishes the value of an analysis of the term. One could try to generalize this point. Suppose that one deals with a question of the form 'What is X?', and someone comes up with the answer 'X is the actual occupant of the X-role'. Even if the 'X-role' is spelt out in detail, the question will still at most be half-answered: We still need to know what it is that actually occupies the X-role. This suggests that what we are really interested in in philosophical inquiry and elsewhere is secondary intensions. And if so, then the utility of conceptual analysis involving opaque expressions seems questionable.

I think it is not so clear that we are always interested in the secondary intension associated with an expression or that we should always be interested in secondary intensions. At least I see no reason why a term's secondary application conditions are generally to be considered a more worthwhile target of inquiry than its primary application conditions.

Nevertheless, it cannot be denied that we can and often do try to reveal secondary intensions. In such cases, conceptual analysis alone cannot give us what we want if the relevant expression is epistemically opaque. This does not mean that conceptual analysis has no role to play in such an enterprise, which can rather be construed as a two-step process: In the first step, the primary intension of the relevant expression is revealed. The second step is to determine empirically what the primary intension picks out.[116] It is true that the first of these steps does not have to be made explicitly. Nevertheless, as I argued mainly in chapter 4, primary intensions are required to determine the secondary intensions even of opaque terms.

5.3.2 Variation in primary intensions

Now let me turn to the second potential source of concern mentioned above regarding the value of opaque expressions in conceptual analysis. I just suggested that it can be a worthwhile project to analyze the primary intensions even of opaque expressions. This presupposes that there is *a* primary intension which is associated with the relevant expression, i.e. one which is at least approximately shared among speakers. But if the main function of opaque expressions is to secure reference to the relevant category, this provides reasons to doubt that the primary intensions of opaque expressions are usually shared. In fact, Anti-Fregeans often point out that the conception (or whatever they call it) a subject associates with for example a natural kind term is irrelevant. Even a Martian for whom 'water' denotes a poisonous gas sometimes used as a chemical weapon

[116] Compare for example Jackson's reconstruction of the identification of the temperature in gases with mean molecular energy:
"Pr. 1 Temperature in gases = that which plays the temperature ('T') role in gases. (Conceptual claim)
Pr. 2 That which plays the temperature role in gases = mean molecular kinetic energy. (Empirical discovery)
Conc. Temperature in gases = mean molecular kinetic energy. (Transitivity of '=')" (Jackson 1998a, 59)

could still possess the concept of water, as long as its ‘water’ utterances or thoughts referred to H_2O.
As we saw in 4.2, it is not obvious that it does not matter which primary intension a subject associates with an opaque term; there are at least some opaque terms where a specific primary intension has to be associated, most clearly in the case of indexical expressions.
Recall, moreover, that in 4.2, I also gave an argument based on the success of (communication in) mutual inquiries in support of the claim that the primary intensions of opaque expressions are usually shared among speakers. If that conclusion is correct, this is sufficient for an analysis of such expressions to be potentially useful. There is thus reason to think that most epistemically opaque expressions are suitable for a conceptual analysis.

Let me sum up the results of this chapter: I first showed that there are opaque as well as transparent expressions and identified ways to show on which side of this divide an expression falls. It transpired that semantic neutrality and epistemic transparency can come apart: While it is sometimes only empirically evaluable whether an expression is semantically neutral, epistemic transparency can be discovered a priori. Then I discussed the potential value of opaque terms in conceptual analysis. I conceded that in many cases where opaque terms are involved, conceptual analysis (alone) cannot give us everything we want. But I also noted that much depends on what one aims at. If one is interested in the ‘metaphysical essence’ of the category denoted by an opaque term, then conceptual analysis can only do part of the job. Nevertheless, an analysis of such a term – revealing its epistemic application conditions – can be of interest in its own rights. To this I should add that conceptual analysis need not be understood as a method which always results in an explicit (linguistic) analysis of a term. The Canberra Plan (which will be introduced in chapter 7), for example, which essentially involves conceptual analysis, can be applied to epistemically transparent and to opaque expressions equally well. And for conceptual analysis to play a role in reductive explanation,

the terms involved do not have to be epistemically transparent, either. I will discuss the potential aims and applications of conceptual analysis in detail in chapter 7.

Addendum: Does incomplete understanding present a problem for conceptual analysis?

At the beginning of this chapter, I conceded that firstly, deferential concepts are epistemically opaque and secondly, all kinds of expressions can be used deferentially, but argued that this does not entail that opacity is ubiquitous. The reason is that such terms can still be (and often are) transparent for those with complete understanding. What I left open, however, was whether the phenomenon of deferential/incomplete understanding itself presents a problem for conceptual analysis. At least at first glance, there are reasons for thinking so. For a start, the primary intension of a deferential concept does not seem to contain particularly useful information. Above I said that the analysis of opaque expressions can be worthwhile; but discovering something like 'X is the stuff called 'X' by the relevant experts' seems hardly of interest.[117] However, this observation only reveals a problem if people with an incomplete or deferential understanding of a term were to try and analyze it. One could thus give the simple advice only to analyze expressions which one understands completely or non-deferentially – which seems natural enough. There is a catch, however: As I already pointed out in chapter 3, while it is a priori whether a concept is deferential, incomplete understanding, and in particular, misunderstanding, may not be accessible to a speaker. There is therefore a risk that a person analyzes a term which she does not understand correctly. In such a case, the analysis may provide substantial information, but only about the subject's individual concept.

[117] I should note, however, that as was already mentioned in chapter 3, scrutability applies to deferential concepts as well.

In fact, this is just a special case of a problem which was discussed in a slightly different context at the end of chapter 3. There, I said that conceptual analysis has an a priori part and an empirical part. The a priori part consists in an analysis of a subject's individual concept, i.e. of the properties she associates with a given term. In the empirical part, one needs to check whether this concept is shared by other members of the linguistic community (and whether the associated properties can thus be said to reflect the term's linguistic meaning). That latter part is necessary to ensure that there is a common subject matter, since we can never know a priori whether a concept is shared by others. In this respect, incomplete understanding does not pose a special problem to conceptual analysis.

There is another problem which could be raised by deferential/incomplete understanding. There may be terms which are incompletely understood by the majority of those who use them. In fact, I think this is not a rare phenomenon. Just consider terms such as 'boson', 'transistor', 'microchip', 'vertebrate', etc. Now the problem is that I said above that for conceptual analysis to be (part of) a common enterprise, the concepts in question have to be shared among speakers of a linguistic community. But if there is only a small group of people – the 'experts' – with a complete understanding of the term, as is plausible in many cases, then this criterion seems to be violated. Or maybe rather, what is widely shared is the deferential understanding, not the complete understanding.

I think this problem is not all that worrisome. First of all, it can be doubted that such terms are especially interesting for a conceptual analysis anyway: Those terms which have been the target of conceptual analysis are typically everyday terms which are well understood by the average speaker. It would presumably make less sense, for example, to analyze technical terms. Goldman argues roughly along these lines for the claim that philosophically relevant terms are understood completely by the average speaker (cf. Goldman 2007, 18).

While I think there is something to this idea, I concede that it would be too bold to claim that we are never interested in an analysis of a term which is incompletely understood by many. So here is a second reason why I think

the present challenge to conceptual analysis is not too serious: When most of the speakers within a linguistic community defer to a group of experts, then they concede, at least implicitly, that the properties associated by these experts are the relevant ones. Therefore, in such cases it is the application conditions identified by an analysis of the experts' concept which are of interest. These might even be said to mirror the term's linguistic meaning: At least I see no principled reason why one should not say that only a relatively small group of speakers within a linguistic community has complete understanding of a term.

I conclude that the phenomenon of deferential or incomplete understanding does not provide any principled obstacles to conceptual analysis. The main reason is that it is a priori whether a concept is deferential; we can thus leave the analysis of concepts to those with non-deferential (and thus typically complete) understanding.

6 Scrutability, primary intensions, and conceptual analysis

The scrutability thesis, according to which we are able to determine the extension of our concepts and thoughts if we are given sufficient empirical information about the world, plays a key role in the two-dimensionalist framework. More specifically, there is an intimate connection between scrutability and primary intensions: Primary intensions determine the extension of an expression with respect to every world considered as actual. Since they are a priori accessible, this entails that a subject is able to identify an expression's extension[118] with respect to these scenarios, provided that she is ideally rational. The scrutability thesis spells out precisely this ability: Whenever a scenario is presented to a subject, that subject will be able to determine the extension of any expression with respect to the scenario, if she possesses the relevant concept.

At the same time, the scrutability thesis seems tailor-made for conceptual analysis. This is because in philosophical practice, conceptual analysis is often based on thought experiments. Typically, a hypothetical scenario is sketched of which one then has to judge whether a specific concept applies to it or not: Does John Searle sitting in his room, handing out Chinese symbols *understand* Chinese (cf. Searle 1980); does Smith *know* that the person who will get the job has ten coins in his pocket (cf. Gettier 1963); is Donald *responsible* for voting for the Democrats (cf. Frankfurt 1969); etc. Consequently, if one could establish the scrutability thesis, this could provide the theoretical underpinning for this practice.

Apart from these general connections with the topic of this book, there are a few issues related to scrutability left open from previous chapters. For example, in chapter 4 I gave an argument for the existence of primary

[118] Recall that in 3.3, I argued that primary intensions are fundamentally a property of (mental) concepts and thoughts. So when I say that a subject is able to determine the extension of an expression, this should be read as saying that the subject is able to determine the extension of the concept or thought she associates with the expression.

intensions which was based on a restricted version of the scrutability thesis, namely (CJ). But since I did not provide a thorough defense of (CJ) itself, that argument for primary intensions may be considered incomplete. There are thus a number of reasons to have a closer look at (some versions of) the scrutability thesis, its role in the two-dimensionalist framework and its practical use for conceptual analysis.

The structure of the chapter will be as follows: In the first part, i.e. in 6.1, I will first try to clarify the content of the scrutability thesis and spell out its commitments. Then I will explain how exactly the thesis is connected with the idea that linguistic expressions are associated with an a priori semantic value. Finally, I will discuss the plausibility of (CJ), which is arguably the most basic version of the scrutability thesis. In the second part of the chapter, i.e. in 6.2, I will be concerned with the utility of the scrutability thesis for the practice of conceptual analysis. Much of my discussion will turn on potential problems arising from idealizations, and, more specifically, on the question of how our actual judgments relate to those ideal judgments invoked in the scrutability thesis.

6.1 Scrutability and primary intensions

In chapter 4 I introduced the following version of the scrutability thesis:

(CJ++) If a subject possesses a concept and has unimpaired rational processes, then sufficient information about any given scenario puts a subject in a position to identify the concept's extension with respect to that scenario a priori.

As I pointed out there, (CJ++) is of vital importance for the two-dimensionalist framework, because it entails that concepts are associated with primary intensions. Let me therefore have a closer look at the commitments of the thesis.

First of all, it has to be specified what is meant by 'sufficient information'. Otherwise, it is easy to interpret the thesis in a way which trivializes it (cf. also chapter 2). For example, if the description of a scenario comprises

explicit information about the extension of the concept in question, then it is certainly possible to derive its extension from the description, but a subject's ability to do this is hardly evidence for the existence of primary intensions. One should therefore require that the information about the scenario is expressed in a limited vocabulary, in what Chalmers calls a canonical description. For present purposes, a vocabulary can be considered as limited simply if it is small enough to not make the scrutability thesis trivial. As I mentioned in chapter 2, Chalmers and Jackson propose that a canonical description of the actual world can be given by PQTI, which is a conjunction of **P**hysical facts, facts about subjects' **Q**ualitative states, a '**T**hat's all' clause which states that there is nothing else besides from what is given by P and Q, and **I**ndexical information which permits one to evaluate (sentences involving) broadly indexical or context-sensitive expressions.

There are surely many kinds of descriptions which could be used to specify the character of the world, yielding different scrutability theses. A two-dimensionalist is not committed to any specific base vocabulary, nor is a proponent of conceptual analysis. It is in any case clear that the expressions used in PQTI cannot serve as a canonical vocabulary for every scenario: Additional kinds of expressions will be required to describe worlds where properties are instantiated which are non-physical and non-phenomenal (cf. 2.1.3). Nevertheless, it is a natural idea to take PQTI as a canonical description of the actual world – I will say more about this below.

Next, it has to be clarified what it means when the information thus specified *puts a subject in a position* to determine a concept's extension: The highlighted phrase should be read as saying that the information suffices in principle to make the relevant judgments. The thesis thus involves some idealizations: A subject could make these judgments only if, firstly, she were able to grasp the information given by the canonical description and, secondly, if she were equipped with ideal rational capacities. Let me have a closer look at what would have to be the case for a subject to meet these requirements.

First of all, the subject would have to be able to grasp the information given by the canonical description. To do so, she has to possess the relevant concepts – i.e., not (only) those whose extension is to be determined, but (also) those used in the canonical description. In the case of PQTI, this first of all means that she has to possess the concepts used in P, which is phrased in the language of a completed physics. Obviously, an ordinary subject does not possess the concepts used by a completed physics; it is very likely that no actual subject does. The subject would also have to possess all kinds of phenomenal concepts, in order to grasp the information given by Q.[119] These will, inter alia, include the concepts expressing a bat's phenomenal states which are evoked when it orientates itself via echolocation, those of a shark sensing electric fields with its lateral line, and (maybe) even more alien ones of extraterrestrial life forms. And these are only the concepts needed to grasp information about the actual world. For other worlds, one will presumably need many other kinds of concepts expressing properties which are not instantiated in the actual world.

Furthermore, the mere possession of all of these concepts does not suffice to actually grasp the relevant information. Needless to say, the canonical description of a whole world will be quite extensive. It seems unlikely that a description of a world can exist within that world itself, or that a subject can grasp complete information about her world even in principle. And even if one idealizes away from these problems, it would still take superhuman capacities to process the information and to draw the relevant conclusions from it.

All in all, it is clear that the idealizations involved in the scrutability thesis are huge. The requirements for a subject to determine the extension of a concept on the basis of a canonical description of a world are so remote from the conditions present in our epistemic and linguistic practice that one might question the relevance of the scrutability thesis for the general project which I am concerned with. Such concerns can be divided into two categories: Firstly, there are concerns about whether there is any relation

[119] Of course this requirement ceases to apply if the phenomenal facts supervene (a priori) on the physical facts.

between the judgments of an ideal thinker who is confronted with a complete description of a possible world and those judgments of actual subjects on the basis of rather fragmentary (hypothetical) evidence. These issues will mainly be addressed in 6.2. Secondly, there are concerns about how the scrutability thesis is supposed to have any bearing on whether linguistic expressions are associated with an a priori semantic value. I will discuss those kinds of questions in the following.

6.1.1 From descriptivism to the scrutability thesis

Suppose it is true that our ability to make judgments in the face of evidence, or (maybe more generally) our ability to evaluate hypothetical scenarios, is based on a grasp of an a priori semantic value associated with our terms. It is at least not obvious how this relates to the question of whether we could determine a concept's extension if we had the information in PQTI and ideal rational capacities. For a start, our actual judgments are not aided by the kind of cognitive capacities necessary to derive a concept from PQTI. But this need not be a problem. It even seems desirable that a semantic theory has such a normative dimension – one should, to borrow a distinction from linguistics, surely allow that there is a gap between our performance and our competence. What is more worrisome is that our actual judgments are based on information which is very different from that involved in the scrutability thesis. As I mentioned above, we do not possess most of the concepts from PQTI, even less those used in canonical descriptions of other possible worlds. So even if we grant that (CJ++) holds, this only seems to show that our concepts have a priori associations which are of no practical use since we do not even possess the concepts which are connected with our concepts.
At the same time, the fact that we do not possess the concepts relevant to the canonical description may raise doubts about (CJ++) itself, for why should our concepts have a priori associations with concepts we do not possess? From this perspective, the supposedly intimate connection between the scrutability thesis and the idea that expressions are associated

with primary intensions is not apparent. Given the importance of both of these theses, it will be worthwhile to explicate in some detail why there is nevertheless such a connection.

Let me start by sketching a seemingly more natural picture illustrating the idea that our expressions are associated with an a priori semantic value, starting from Jackson's descriptivist account. According to Jackson, linguistic expressions are associated with properties. These properties determine an expression's extension with respect to every world considered as actual, and thus its primary intension. In chapter 2 I said that one should construe the talk of properties in a liberal sense here. Most basically, properties can be interpreted as extensions across possible worlds. Understood in this way, Jackson's brand of descriptivism fits nicely with the two-dimensionalist framework, but at the cost of making the descriptivist thesis vacuous if one is already committed to the claim that expressions are associated with primary intensions. Note that this does not mean that the account is essentially incomplete or circular. One could simply say that the grasp of a primary intension is nothing but a disposition to ascribe an extension to the expression with respect to any hypothetical scenario, without specifying what underlies this disposition. But for present purposes, in order to illuminate the relation between the scrutability thesis and the thesis that there is an a priori semantic value, it will be useful to develop a slightly more committing view of how the associated properties are to be construed. As I pointed out in chapter 2, such a view should not be too committing: There is no reason to assume that we associate expressions only with particularly natural properties, for instance. Nevertheless, one can take these properties to be real features of (often) mind-independent entities, or to be more precise as features which such entities have or could have.[120] Often, these will be macroscopic features to which we have

[120] Can we not ascribe properties which could not possibly be instantiated, say because they involve a hidden contradiction? Jackson would probably say that there are no such properties, and thus no corresponding contents. At least this is suggested by his discussion in Jackson 1998a, 125ff., where he argues that there are no logically

epistemic access. Take for instance the theoretical role associated with 'water', as discussed in chapter 3: The cluster of properties defining this role comprises features such as being liquid, being transparent, flowing in our rivers and lakes, etc. To get an idea of the properties which our expressions are connected with, one only has to look at how hypothetical scenarios are described in thought experiments: The descriptions usually characterize the same kinds of ordinary macroscopic features of our environment.[121]

Now if a hypothetical scenario is described to which a 'target expression', such as 'free action', 'knowledge' or 'water' applies, then these hypothetical circumstances represent sufficient conditions for the applicability of the relevant term. Importantly, the target expressions themselves are usually not used in the descriptions of the scenarios. So if the judgments about the applicability of a concept in the hypothetical circumstances at hand are a priori, one can conclude that there are a priori connections between the expressions used to describe the scenario and the target concept. Of course, the expressions used in such descriptions themselves are also associated with properties. What can be said about these? One can give a simple, and trivial, answer: 'liquid' is associated with being liquid, 'transparent' with being transparent etc. (An analogous kind of answer could already have been given to the question with what kinds of properties the target expressions are associated.) But at least in many cases, it is possible to say more. In particular, one can often give descriptions of scenarios to which the terms used in the descriptions of the original scenarios apply, without using those terms in the new descriptions. To give an example, it seems possible to specify the behavior of a substance, including its dispositions to behave in counterfactual circumstances, and conclude from this specification that the substance is liquid. Likewise, it

---

equivalent predicates which refer to different properties. This may seem counterintuitive, however. I will say more about this issue at the beginning of 6.2.

[121] But of course, these ordinary features are often combined in extraordinary ways. It is common for thought experiments to depict circumstances where, for instance, correlations between properties which are very robust in the actual world fail to hold.

seems plausible that one could give a description of a scenario from which one can derive that there is water present, without using terms such as 'liquid' or 'transparent', but even more basic expressions.

Of course, one cannot go on like this forever. At some point, one will be unable to give an even more basic description of the situation.[122] And we already know that the most basic description which we are able to give is not PQTI, or another canonical description of a possible world, since we do not possess the relevant concepts. However, the scrutability thesis is only committed to the claim that if we were in possession of the concepts used in the canonical description of a scenario (and were ideally rational), then we could determine the extension of each of our concepts on the basis of such a description.

I think we have already taken some important steps towards a clarification of the relation between the scrutability thesis, as expressed by (CJ++), and the idea that expressions are associated with an a priori semantic value: The basic idea is that expressions are a priori associated with properties, understood as putative features of the world, which determine their extension with respect to every world considered as actual. Very often, these properties can be expressed by more basic expressions. The residual question which I am going to address in the following is how one gets from there to the thesis that there is one basic vocabulary which can be used to describe the complete character of whole worlds and, crucially, that the extension of each of our concepts and thoughts can be derived from such a description.

Let me start by noting that according to two-dimensionalism, the hypothetical scenarios we evaluate on the basis of the associated properties correspond to possible worlds – unless their description implies an a priori contradiction. This is guaranteed by the thesis of metaphysical plenitude, which was introduced in chapter 2:

[122] Or one will land in a circle, in case the terms replacing the original ones were not really more basic after all.

Metaphysical Plenitude: For all (sentences) S, if S is epistemically possible, there is a centered metaphysically possible world that verifies S.

A consequence of this thesis is that everything is a priori scrutable from a canonical description of a scenario which exhaustively specifies the distribution of fundamental properties, if supplemented by a) a clause stating that the description thus given is complete and b) indexical information.[123] To see this, suppose that some fact F is not a priori scrutable from a canonical description D_n which specifies the distribution of all the fundamental properties in w_n. Then 'D_n & ~F' is epistemically possible. Given metaphysical plenitude, there is thus a metaphysically possible world verifying 'D_n & ~F'. And since canonical descriptions of scenarios only involve semantically neutral expressions, this entails that D_n & ~F is metaphysically possible.[124] Consequently, F does not supervene on the properties specified in D_n, which means that if, as we assumed, F is really a fact in w_n, then D_n cannot specify the distribution of all the fundamental properties in w_n, violating our initial assumption. It follows that any truth which is not scrutable from D_n and which thus violates the scrutability thesis also violates metaphysical plenitude.

Plausibly, PQTI is such a description which exhaustively specifies the distribution of fundamental properties in the (actual) world. It is also

[123] Regarding the need for additional indexical information, cf. chapter 2. The clause which says that the description is complete is necessary inter alia to enable the scrutability of negative truths. For example, if the description only mentions tables, then only the addition of that clause makes 'there are no chairs' scrutable.

[124] If there were semantically non-neutral expressions in D_n, however, then it could happen that a sentence of the form 'D_n & F' is an a posteriori necessity akin to 'Water is H_2O'. The negation of the latter sentence is epistemically possible and (in accordance with metaphysical plenitude) it is also verified by a metaphysically possible world. However, that world should more adequately be described as one where the watery stuff is not H_2O (cf. chapter 2). The requirement that only neutral expressions are used in a canonical description suffices to avoid these problems. I should note that, since it sometimes requires empirical information to determine whether an expression is semantically neutral (cf. 5.2.3), it makes sense to use only expressions which are also epistemically transparent.

plausible that the vocabulary in which PQTI is couched is limited. This fact, together with the foregoing considerations, explains why Chalmers and Jackson consider PQTI as a suitable basis from which all other truths are a priori scrutable.

I think we are now in a position to complete the general picture concerning the connection between the scrutability thesis – in particular (CJ++) – and the thesis that expressions are associated with an a priori semantic value. A key idea of two-dimensionalism is that linguistic expressions are a priori associated with properties. These properties determine the expressions' extensions with respect to every world considered as actual. Supplemented by metaphysical plenitude, this entails that all truths are a priori scrutable from a description of a world which specifies the distribution of fundamental properties in that world in a semantically neutral vocabulary. One can also argue in the opposite direction: If linguistic expressions are not a priori associated with properties which determine an expression's extension across possible worlds considered as actual – either because there are no associated properties, or because the associated properties are not a priori associated, or because they do not determine the extension with respect to every possible world – then (CJ++) is false. From this we get, by modus tollens: If (CJ++) is true, then linguistic expressions are associated with primary intensions.

Of course the fact remains that our actual judgments which are (supposedly) enabled by the grasp of primary intensions are not based on PQTI or any other complete description of a world; and it is undeniable that there are many other differences between actual judgments and those involved in the scrutability thesis.[125] But as was just explained, despite these differences the scrutability thesis is closely linked with the thesis that there is an a priori semantic value.

[125] As I said above, I will discuss potential problems arising from these differences in 6.2.

6.1.2 A case for (CJ)

If one could establish (CJ++), one would have thereby shown that linguistic expressions are associated with primary intensions. In 4.1, I did give an argument in support of primary intensions which made use of that connection. That argument, which aimed to establish a version of Jackson's (epistemic) thesis that we need primary intensions to determine our subject matter, took (CJ) as a premise, which is a version of the scrutability thesis seemingly less ambitious than (CJ++):

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension.

As I mentioned there, this thesis is accepted even by a number of critics of two-dimensionalism. But although (CJ) indeed appears to be much weaker than (CJ++) which says that the derivability from the information given is a priori and applies to all scenarios, I argued that if one accepts the former thesis, then it is not easy to reject the latter one either. It might therefore be wiser for an opponent of two-dimensionalism to reject (CJ). In any case, the tenability of (CJ) is of crucial importance for two-dimensionalism. In the following, I will thus have a closer look at the thesis. There are a number of considerations which have been invoked or which could be invoked to support the scrutability thesis and which I will thus discuss in the following, along with various objections against it.[126] It has to be noted that it is hardly possible to show conclusively that the scrutability thesis is correct, given the scope of the thesis. Surely I do not aim to offer such a proof here. I will try to make it plausible, however, that the reasons to endorse the thesis outweigh the reasons to reject it.

[126] A much more detailed defense of various scrutability theses can be found in Chalmers forthcoming.

6.1.2.1 Argument from metaphysical plenitude

As I noted above, the thesis of metaphysical plenitude entails a version of (CJ) in which the relevant 'sufficient empirical information' consists in information about the distribution of fundamental properties. One could therefore try to back up (CJ) by making a case for metaphysical plenitude.
In fact, metaphysical plenitude is much stronger than (CJ): According to the plenitude thesis, any metaphysical entailment relation is also an a priori entailment relation, at least where semantically neutral expressions are involved; moreover, it is a thesis about all epistemic possibilities and therefore also concerns many non-actual scenarios. The thesis is therefore sufficient to establish (CJ++). Metaphysical plenitude is surely a bold thesis. Nevertheless, I think it can be defended on independent grounds. In chapter 2 I already showed that the classical a posteriori necessities, which have convinced most philosophers that there is a gap between the epistemic and the metaphysical modality, can be accommodated within the two-dimensionalist framework. Accordingly, counterexamples to metaphysical plenitude have to be of a completely different kind than these Kripke/Putnam-style cases. They would have to be what Chalmers calls 'strong necessities', i.e. a posteriori truths which are verified by every metaphysically possible world. It is safe to say that so far, no clear example of such a strong necessity has been presented. (I will discuss a number of putative examples which have been invoked briefly below.)
In my view, the main motivation for assuming that epistemic and metaphysical modalities can come apart, besides the Kripkean cases, is to be found in the context of the debate over physicalism; at least, this is where the topic has been discussed most extensively: While there are a number of reasons to endorse physicalism, phenomenal consciousness presents a problem to such a view. For, although a physicalist has to hold that all facts are determined by physical facts, it is hard to see how phenomenal facts could be derived from physical facts. To put it differently, it seems epistemically possible that there is a world which is physically completely identical to ours, while differing in its phenomenal

features (cf. Chalmers 1996, 93ff.). If one accepts this, one either has to give up physicalism or reject metaphysical plenitude, and many philosophers have chosen the latter.
While I see the physicalist's dilemma, I do not think that it is a good strategy to reject metaphysical plenitude just for the sake of being able to stick to one's metaphysical commitments. To support her case, a physicalist should thus either argue that the plenitude thesis fails in other areas as well or try to give an explanation why the psycho-physical case is somehow exceptional. A popular way of implementing the latter strategy has consisted in an appeal to special features of phenomenal concepts in order to explain the epistemic gap between the phenomenal and the physical, while insisting that there is no corresponding ontological gap (cf. e.g. Hill & McLaughlin 1999; Tye 1999).[127] I should note that such a rejection of metaphysical plenitude need not threaten the general project of vindicating conceptual analysis on the basis of two-dimensionalism, as long as the exceptions to plenitude are local and explicable (cf. also 2.3).

6.1.2.2 Arguments from the scrutability of specific kinds of facts

Another way to support (CJ) is to start from a specific description of the actual world and then to argue that certain kinds of facts are derivable from this description. For instance, in *Conceptual Analysis and Reductive Explanation* (2001), Chalmers and Jackson try to make it plausible that ordinary macroscopic facts, such as 'Grass is green' or maybe 'There are more tables than people in Zilshausen', are scrutable from PQTI: For example, P specifies complete information about the space-time positions, velocities and masses of microphysical entities as well as their dispositions to behave in various circumstances.
From this information, one can derive facts concerning the 'structure and dynamics' of macroscopic systems, i.e. concerning their speed, location, shape and their dispositions. This way, one already gets a great number of

[127] Cf. Chalmers 2007 for reasons to doubt that this strategy can succeed.

macroscopic truths. If one adds to this the information from Q and indexical information, then it is supposed to be plausible that one will be in a position to derive all ordinary macroscopic truths.
I have not spelled out all the details of Chalmers and Jackson's argument. However, it is based on such rather general considerations throughout. And while the argument has some intuitive force, I find it hard to evaluate whether macroscopic truths are really scrutable from PQTI as long as one does not provide more detail as to how such a derivation is supposed to proceed.
Since it is hardly possible to provide a detailed analysis of the steps of such a derivation for each kind of macroscopic truth, it seems sensible to try and offer examples for how particular kinds of facts can be derived from more basic information. One such example is discussed by Andrea Wille (cf. Wille ms.). She addresses the question whether the disposition of water to form drops can be derived from a quantum-chemical description of H_2O molecules. Wille does not arrive at a clear conclusion, however: Although she herself provides such a derivation, she also notes that the theoretical models she relies on involve some approximations; it is not clear whether these could be replaced by models which are completely adequate. Most notably, there are problems concerning the transition from a quantum-mechanical level to the level of chemical properties.

If successful, arguments of the type just sketched would show that some kinds of truths are scrutable from more basic kinds of truths. Furthermore, since the two kinds of truths are expressed by quite different kinds of terms, this also makes it plausible that there are conceptual connections between different kinds of vocabularies, provided that the derivation does not rely on additional empirical background information (cf. chapter 4). However, an opponent could concede that ordinary macroscopic truths are scrutable from more fundamental information, but deny that the same holds for all other kinds of truths. And indeed, there are many kinds of truths which seem to pose a more serious challenge to the scrutability thesis. Therefore, rather than showing how some kinds of facts are thus scrutable, one could

try to defend the thesis by dispelling doubts about the scrutability of such problem cases. I will have a look at this strategy in the following.

6.1.2.3 Arguments from the absence of clear counterexamples

Recall that the scrutability thesis I want to defend states that all truths can be derived from fundamental truths (plus indexical information and a 'that's all' clause, cf. above). A problem case would thus have to involve a non-fundamental fact which is not derivable from the fundamental facts. This means that a critic of (CJ) who tries to establish a counterexample will have to argue for three things: Firstly, that there really is a fact of the matter; secondly, that this fact is not fundamental; and thirdly, that it is not scrutable from fundamental facts even for an ideally rational thinker.[128] Chalmers has discussed a great number of potential counterexamples to scrutability (cf. e.g. Chalmers 2010, ch. 6; Chalmers forthcoming, ch. 6): phenomenal truths, complex mathematical truths, moral truths, metaphysical truths, modal truths, intentional truths, etc. For reasons of space, I cannot reproduce the details of his discussion here, let alone add something substantial to it. While I agree with Chalmers that there are no clear examples which satisfy all three of the criteria just outlined, I nevertheless believe that there are cases which do raise some doubts about the scrutability thesis. To illustrate this, let me bring in a potential problem case which so far (to my knowledge) has not been discussed in the debate about inscrutable truths and strong necessities: the problem of personal identity. In the literature on personal identity, one can find discussions of a huge number of thought experiments. The hypothetical scenarios involved are typically quite exotic. For example, Derek Parfit invites us to imagine what would happen if some device destroyed our body completely, just to create a molecule for molecule identical copy of ours elsewhere: Would this person still be me (cf. Parfit 1984, 199f.)? In many such cases, we are

[128] Note that since metaphysical plenitude entails (CJ), any such counterexample to (CJ) will also reveal a strong necessity.

at loss what to say. For instance, even if we do have the intuition that we would survive such a procedure (maybe under the influence of various science-fiction stories where similar procedures are considered as means of long-distance travel, rather than of murder), the case can easily be modified such that it becomes even harder to evaluate: What if my body is not destroyed before the copy is created? And what if the original body dies shortly afterwards? Furthermore, even in those cases where we have a clear intuition about whether personal identity is given, it always seems as though the opposite is nevertheless conceivable. Even if one thinks, say, that one will survive the kind of 'teletransportation' just described, it does not seem contradictory to think that one might not.[129] And if so, then both options are epistemically possible.

Prima facie, this is a problem for the scrutability thesis because it suggests that facts about personal identity are not derivable from other kinds of facts. So how could one try to resolve this problem? I see the following three ways for an adherent of the scrutability thesis to respond to the challenge raised by the problem of personal identity:

First of all, one could argue that such cases would be conclusively evaluable if we were given complete information about the relevant scenario. According to this line of response, the inconclusiveness of such judgments does not threaten the scrutability thesis because they are based on incomplete information about the relevant cases.

It is admittedly obvious that we do not base our judgments on complete information about hypothetical scenarios, at least if one assumes that these scenarios correspond to possible worlds. It might thus be true that we thereby lack information which is relevant to the question at hand. Yet, it is hard to see how we could give a conclusive verdict even if we were given complete physical (and phenomenal) information. What information exactly would we need in order to tell whether we would still be the same person after the procedure?

[129] Cf. also Williams 1970 for a thought experiment where both outcomes seem possible.

The second way to defend the scrutability thesis is to hold that complete physical (and phenomenal) information is indeed insufficient to derive facts about personal identity, but only because physical (and phenomenal) facts do not determine facts about personal identity. According to such a 'further fact view', facts about personal identity are either themselves fundamental or they depend on other fundamental facts beyond physical (and phenomenal) ones (cf. e.g. Chisholm 1976; Shoemaker & Swinburne 1984; Lowe 1996). It is true that if the further fact view were correct, the cases at hand would pose no threat to the scrutability thesis or to metaphysical plenitude. But many philosophers do not like such a decidedly anti-naturalist account, for metaphysical, but also for epistemic reasons – for how are we supposed to recognize these additional fundamental facts?
The third line of response available to a proponent of the scrutability thesis is to argue that at least in many cases, there are no facts about the identity of persons and consequently, there are no truths to be derived from a description of such cases. The downside of this view is that it is highly revisionary. Common sense tells me that either I do survive or not, there does not seem to be any room for indeterminacy.

It thus appears that none of the available attempts to defend the scrutability thesis against the problem raised by the thought experiments on personal identity is entirely satisfactory. Nevertheless, it is far from clear that one should abandon the scrutability thesis in the light of these considerations. For consider what the resulting view would involve: One would have to say that although facts about the identity of a person are determined by physical (and phenomenal) facts, there is no way for us even in principle to come to know the supervenient facts based on knowledge of the subvenient facts. But then it becomes dubious how we come to know such facts at all. For, it is hardly plausible to assume that we have some kind of direct insight into the essence of a person. And if we concede that there is no way to know such facts, then our position towards someone who claims that facts about personal identity are sui generis, and especially towards someone who says that there are no such facts, no longer seems particularly strong.

Similar considerations apply in the case of other potential counterexamples to scrutability and metaphysical plenitude: If we insist that there are facts which are not scrutable from those facts we take to be fundamental, then it becomes unclear what reasons we have to assume that these inscrutable facts exist and that they supervene on those considered as fundamental.[130]
It has to be conceded that there are (putative) kinds of facts which do raise problems for the view that everything is derivable from a description of the distribution of fundamental properties. On the other hand, given the state of our knowledge and understanding of the world it is hardly surprising that there are truths concerning which we have (yet) no idea how they could be derived from fundamental truths. For this reason, our inability to derive such facts from lower-level facts need not be taken to indicate that they are inscrutable in principle. Let me thus conclude by noting that there are at least no definite counterexamples to (CJ) and that it is furthermore not obvious that we could ever have good reasons to believe in non-fundamental truths which are not scrutable from fundamental ones.

6.1.2.4 Arguments from the epistemic indispensability of scrutability

Finally, let me give two closely related arguments for the claim that many kinds of truths have to be derivable from more basic kinds of truths. Both of them tie in with the observation just made that inscrutable truths raise serious epistemic problems. According to the first of the arguments, if some set of facts cannot be derived from another, more basic set of facts, then the former facts cannot be reductively explained by the latter ones, either. The second argument is broader in scope. It aims to show that if we did not have the ability to derive many kinds of truths from more basic kinds of truths, then we would not be able to gain knowledge of the higher-

[130] The situation is slightly different in the case of phenomenal consciousness, since it is plausible that we can gain knowledge of phenomenal truths even if they are not scrutable from physical truths. On the other hand, the option to treat such facts as fundamental seems at least slightly less unattractive than in the case of many of the other kinds of facts mentioned above.

order truths at all. Both of these arguments thus point out that failures of scrutability would lead to undesirable epistemic consequences. And both of them aim to establish not just that it has to be possible to derive some kinds of facts from others in principle, i.e. under ideal cognitive conditions, but that we need to have the actual ability to do this.

Let me start with the argument for the requirement of scrutability in reductive explanations. In a reductive explanation, one aims to explain some higher-level phenomenon, such as the disposition of sugar to dissolve in water, by means of features of lower-level phenomena, such as the molecular structure of sugar. If one succeeds, one has not only explained the macroscopic phenomenon, but one has also shown that the explanandum is nothing over and above the features invoked in the explanans, i.e. one has also given a metaphysical reduction of the higher-level phenomenon. Now it seems clear that for the lower-level facts to explain the higher-level facts, the former ones must make it intelligible why the latter ones obtain, i.e., it must be intelligible why precisely these higher-level facts obtain and not different ones. In the light of these considerations, it is natural to think that for a reductive explanation to be successful, it must be possible to come to know the higher-level facts based on knowledge of the lower-level facts.

This would mean that in any reductive explanation, some kinds of facts are derivable from more basic kinds of facts. The dispute between Chalmers and Jackson on one side and Block and Stalnaker on the other which was mentioned in my discussion of the transition from (CJ+) to (CJ++) in 4.2 is situated precisely in this context. All of them seem to agree that a reductive explanation requires that the facts comprising the explanans are scrutable from the facts in the explanandum. Their disagreement is over the question whether the derivability must be a priori. However, that issue is irrelevant to the current argument, which only aims to support (CJ).

The upshot of the foregoing considerations is that if we were not able to derive higher-level truths from lower-level truths, then there could be no reductive explanations. The extent to which one thinks that higher-level phenomena can be reductively explained thus involves a commitment

concerning the extent to which such derivations are possible. But of course, not everybody is a friend of reductive explanations. Someone who does not believe that (m)any phenomena can be reductively explained will hardly be moved by the argument just presented. Let me thus give another argument in support of (CJ) which is considerably larger in scope.

In chapter 4 I said that the scrutability thesis can be taken as a generalization of our everyday ability to make judgments in the face of evidence. I also suggested that if one denies that we have such an ability, this will lead to general skepticism. The argument for (CJ) which I now want to present rests upon this idea. More specifically, its crucial point is that many kinds of truths would have to remain inaccessible to us if they were not derivable from other kinds of truths. To see this, consider our basic sources of evidence: Plausibly, these are just sense perception, introspection, and perhaps also rational intuition. Let me put the latter faculty aside here, though, since if it exists, then it is an a priori source of evidence anyway. If the list I just gave is in fact complete, this entails that everything we know ultimately has to be based on one of these sources, i.e. everything we know must be knowable on the basis of what these sources deliver. So what do they deliver, i.e. what are the contents of their immediate deliverances?

Introspection only gives us direct information about ourselves – our phenomenal states, our propositional attitudes, etc. The contents of perceptual states are more diverse. Sense perception gives us information about a huge variety of features of our surroundings. It is of course a difficult question as to how the contents of perception are to be construed; so it is hard to say what exactly is immediately given to us via perception. Some say that these contents comprise information about higher-order properties like being a table or a car, for example (cf. e.g. Siewert 1998; Siegel 2010). Others hold that perception only gives direct information about features like color, shape, size and distance to oneself (cf. e.g. Dretske 1995; Tye 1995). From this information, higher-order properties will then be inferred. But even if the former of these views is closer to the

truth, it still seems clear that large parts of our knowledge concern properties which cannot be represented by our perceptual states. There is political knowledge – for example knowledge about whether some state or other is democratic or not –, economic knowledge, moral knowledge, knowledge about the future and the past, modal knowledge, etc. Even if one thinks that some of these propositions can be among the contents of our perceptual or introspective states, it is hardly plausible to claim that of all them can. Consequently, if we have knowledge of such truths, then we have come to know them either by inferring them from those truths which were presented to us via our basic sources of evidence or by testimony from others who have done so. This means that many kinds of truths must be scrutable from those kinds of truths presented to us via perception and introspection.

Note that I am not arguing against empiricism here; I am not even arguing that the derivation of the relevant higher-level facts from those given by perception and introspection has to be a priori. All that the current argument is committed to is the claim that much of our knowledge is not immediately presented to us by our basic sources of evidence and therefore has to be derivable in some way from the contents which these sources deliver.

To conclude: I just discussed a number of arguments for (CJ). The first of these arguments was based on the thesis of metaphysical plenitude. Then we saw that there are good reasons, both direct and indirect ones, to think that at least many kinds of truths are derivable from more basic kinds of truths. For example, it seems plausible, though it is not entirely clear, that many (if not all) ordinary macroscopic truths can be derived from basic physical and, if necessary, phenomenal truths. Furthermore, although I conceded that there are some potential counterexamples to the thesis which are not easy to rebut, I also showed that the acceptance of any such counterexample would raise serious epistemic and metaphysical problems. Finally, I gave two related arguments which were designed to show that scrutability plays an important role in reductive explanations and in our

epistemic practice more generally. These arguments do not show that all truths have to be scrutable from fundamental truths. However, they strongly suggest that many kinds of truths are derivable from more basic kinds of truths.
(CJ) is certainly a very ambitious thesis. I already said at the outset that it will hardly be possible to prove that it is true. But still, I think that on balance the considerations in its favor weigh more heavily than those which speak against it.

6.2 Semantic idealizations and epistemic reality

In the following, I will deal with a number of issues arising from the idealizations involved in two-dimensionalism in general and in the scrutability theses in particular. I will first, i.e. in 6.2.1, discuss the question whether primary intensions are insufficiently fine-grained, because they fail to take phenomena arising from the cognitive limitations of actual subjects into account. I will discuss two ways to allow for more fine-grained distinctions and argue that none of them provides a completely satisfactory account of the phenomena in question. Furthermore, I will show that the connection between primary intensions and ideal rationality is essential for the purposes of conceptual analysis.
In 6.2.2, I will be concerned with judgments about hypothetical scenarios in our philosophical practice. I will first identify ways in which their reliability could be called into question. It will emerge that the most promising way to do this is to argue that these judgments rely on merely empirical associations which may thus fail to hold in the hypothetical scenarios considered. I will introduce Hare's two-level model of moral thinking, which offers an attractive basis for arguing that our intuitive moral judgments draw on empirical correlations. Since similar accounts could be applied to epistemology and possibly to other areas as well, the two-level model might therefore raise a challenge to our judgments about hypothetical scenarios more generally. I will argue, however, that it is at least unclear whether the two-level model provides an adequate account of

our judgments in moral, epistemic and other matters. Furthermore, and crucially, I will show that even though the model suggests that our intuitions about hypothetical scenarios are unreliable, it requires that we are able to evaluate hypothetical scenarios in other ways. I will then argue that since it is plausible that these other ways include a priori methods, the two-level model does not pose a threat to the viability of conceptual analysis after all.

Finally, I will consider how we could try to approach those ideally rational judgments involved in the scrutability theses in our epistemic practice. I will argue that there are several ways to evaluate hypothetical scenarios apart from relying on our intuitions about the scenarios themselves.

6.2.1 Are primary intensions too coarse-grained?

In chapter 2 I said that two-dimensionalism stands in the tradition of Frege's theory of meaning. It thereby inherits a number of advantages compared to other accounts which posit less fine-grained semantic values. Take for instance a straightforwardly extensional theory of meaning. Such a theory entails that expressions which have no reference have no meaning and that expressions which have the same extension are synonymous. But this seems false. Expressions such as 'unicorn' and 'perpetuum mobile' are hardly meaningless, and it seems equally obvious that 'Superman' and 'Clark Kent' differ in meaning. Furthermore, it is a natural condition for synonymy that one should be able to substitute synonymous expressions for each other without thereby changing the truth-values of the sentences in which they occur. Yet famously, there are contexts where the substitution of co-extensional expressions is problematic, for example in sentences describing propositional attitudes: While 'Lois Lane believes that Clark Kent has poor eyesight' is arguably true, 'Lois Lane believes that Superman has poor eyesight' seems false.

Typical intensional theories can account for many of these problem cases. For example, since there could have been unicorns, 'unicorn' does have an intension. But these theories cannot account for all of them. For if one

follows Kripke and accepts that names like ‘Clark Kent’ and (let’s say) ‘Superman’ refer rigidly, then if they co-refer, they do so necessarily. Accordingly, ‘Clark Kent’ and ‘Superman’ have the same intension.

By positing two kinds of intensions, the two-dimensionalist framework allows for even more fine-grained distinctions than other intensional accounts. For example, the thought that Clark Kent has poor eyesight may have the same secondary intension as the thought that Superman has poor eyesight; however, they differ in their primary intensions.[131] Likewise, two-dimensionalism can account for the intuitive difference in meaning between pairs of expressions like ‘water’ and ‘H_2O’. However, there are problem cases which two-dimensionalism seems unable to resolve: Assume that Lois Lane believes that $2^7 = 128$, but for some reason fails to believe that knowledge implies truth. This does not even seem possible according to the two-dimensionalist framework. This is because any a priori truth expressed in epistemically transparent vocabulary has a necessary primary and a necessary secondary intension; any such falsehood has a necessarily false primary and secondary intension. Two-dimensionalism is therefore unable to distinguish these truths or falsehoods. Yet, it is obvious that ‘$2^7 = 128$’ and ‘knowledge implies truth’ are no synonyms, just as it is obvious that someone who believes an a priori necessary truth does not believe all a priori necessary truths, etc. Note that Frege’s own theory handles such cases easily. According to Frege, the sense of an expression is constitutively linked to its cognitive value (cf. 1.2). And surely, to recognize an a priori (necessary) truth can be of cognitive value; moreover, not all a priori (necessary) truths have the same cognitive value. In Frege’s account it is thus possible to make seemingly useful distinctions which are neglected by the two-dimensionalist framework.

One of the motivations for positing primary intensions is to take a subject’s epistemic perspective into account. It is thus surely legitimate to ask for what reason it restricts itself to modelling the perspective of ideal thinkers, thereby leaving aside important phenomena which arise from our cognitive

[131] Two-dimensionalism also allows distinguishing expressions, concepts or thoughts which are a priori equivalent, but not (metaphysically) modally equivalent.

limitations. Not least, it is crucial to take these cognitive limitations into account in order to understand how conceptual analysis proceeds. For instance, often hypothetical scenarios are invoked in order to show that a proposed definition is inadequate, like the famous Gettier cases in the debate on the analysis of knowledge: These scenarios aim to demonstrate that knowledge is not justified true belief. Now if it is a priori that knowledge is not justified true belief, then the primary intension of 'knowledge = justified true belief' is necessarily false; there is thus no scenario which verifies 'knowledge = justified true belief'. Within the two-dimensionalist framework, it is therefore hard to see how people could believe that knowledge is justified true belief. Relatedly, two-dimensionalism leaves the fact that something important was learned from Gettier's scenarios inexplicable – after all, no possibility was excluded and thus no information gained when we came to believe that knowledge is not justified true belief.

The question is therefore whether it is possible to accommodate these data and thus to allow for distinctions among a priori truths and falsehoods in a way which is compatible with the two-dimensional framework. One possible way to do this is to understand primary intensions as structured entities, along the lines of Lewis and Max Cresswell (cf. Lewis 1972; Cresswell 1985).[132] The primary intension of a complex expression can then be conceived as a complex, consisting of the primary intensions of its parts. In this way, different a priori truths can be distinguished and it becomes intelligible how Lois Lane can believe that $2^7 = 128$ while failing to believe that knowledge implies truth. It has to be noted, however, that the structure of the primary intension of a sentence does not generally reflect its cognitive value. The structure of the intension of a very simple truth may be even more complex than that of one which is very hard to recognize – one should not suspect semantic complexity to match epistemic complexity.

[132] Cf. also Chalmers 2011b, where Chalmers specifically proposes to understand primary intensions as being structured.

Another way to introduce more fine-grained distinctions which is seemingly more natural than the one just sketched is to modify the notion of epistemic possibility. Since it seems strange to say that a person can believe something which is not even epistemically possible, one could say that there are different versions of epistemic possibility only one of which is tied to ideal rational capacities. A thought may thus be called epistemically possible if it cannot be ruled out by reasoning of a certain sort, i.e. by reasoning which may only be backed up by rather limited rational capacities.[133] This would make it possible to explain how a subject can believe that knowledge is justified true belief without rendering her completely irrational. Given the strength of one's condition on the required reasoning, 'knowledge is justified true belief' may well count as epistemically possible while for instance '2 + 2 = 5' does not.

There are countless ways to spell out such a version of epistemic possibility; it is doubtful whether there is a single best way to do this. What kind of reasoning one deems appropriate to define the epistemic modality will presumably depend, inter alia, on one's explanatory purposes. But I will not discuss this issue any further because it is anyway not clear that a semantic theory should link semantic values to some kind of non-ideal reasoning. I already mentioned in chapter 6 that there should always be a normative component to such a theory. To require anything less than ideal rationality would mean to water down the inherent normativity. Here is an example to illustrate why this would be problematic: It has proved very useful to model the content of a sentence or thought via its truth conditions. But there are no conditions under which knowledge is justified true belief or Fermat's last theorem is false, however difficult it may be to find out that this is so. This suggests that the only way to link truth conditions with epistemic possibilities is to require that a sentence or thought is epistemically possible only if it cannot be ruled out by ideal reasoning.

The foregoing considerations suffice to show that it is at least arguable whether semantic values should accommodate phenomena arising from non-ideal rationality. The point I just made about modelling content via

[133] Such non-ideal kinds of epistemic modality are discussed in Chalmers 2011a, 102ff.

truth conditions also suggests that it would be problematic to base conceptual analysis on a theory which draws on a more fine-grained notion of epistemic modality. This is because any such notion has no chance of satisfying metaphysical plenitude. The fact that a sentence or thought cannot be ruled out by some non-ideal type of reasoning can never guarantee that it is metaphysically possible. Accordingly, if one wants to draw substantial philosophical conclusions from an analysis of the primary intension of an expression, concept or thought, then it is essential that primary intensions are based on nothing short of ideal epistemic possibility. To conclude: Depending on one's purposes, it may be useful to introduce additional semantic values which are suited to accommodate phenomena arising from our limited rational capacities. There are a number of ways to do this which are in principle compatible with two-dimensionalism. However, as we saw there are good reasons to leave a semantic value which is tied to ideal rationality in place; if one aims to make two-dimensionalism as the basis of conceptual analysis, then such a semantic value is even indispensable.

6.2.2 Scrutability for real subjects

As was already stressed a couple of times, the scrutability thesis is not only a crucial component of two-dimensionalism, it could also play an important role in a vindication of conceptual analysis. In particular, it could provide the theoretical basis for the evaluation of hypothetical scenarios which are often invoked in thought experiments. Nevertheless, even if – as I argued above – the scrutability thesis is correct, this does not ensure that our judgments about hypothetical cases are reliable. Our actual judgments can therefore still go wrong, either because we are not ideally rational or because we have incomplete information about the scenario in question. In what follows, I will discuss the relation between those judgments under ideal conditions relevant to the scrutability thesis and the judgments we actually make in our epistemic practice. I will first deal with some doubts which could be raised concerning the reliability of our judgments. Then I

will address the question of whether and how our less-than-ideal judgments about hypothetical scenarios might be improved upon.

In recent years, the reliability of our intuitive judgments about hypothetical scenarios has been the subject of a number of empirical studies. The results obtained by experimental philosophy are not always comforting for proponents of conceptual analysis. It was found that there is often considerable intersubjective divergence in the judgments depending on factors such as the subjects' cultural background (cf. Weinberg, Nichols & Stich 2001; Machery et al. 2004), socio-economic status (cf. Weinberg, Nichols & Stich 2001), and even the order in which the cases are presented (cf. Swain, Alexander & Weinberg 2008). Some experimental philosophers have argued that these studies show that our intuitions about hypothetical cases are generally unreliable, which would mean that a substantial part of traditional philosophical methodology is flawed (cf. e.g. Weinberg 2007; Mallon et al. 2009).

This conclusion seems hardly justified, though. There are admittedly many ways in which our judgments can go wrong: We can fail to understand the case correctly, pay insufficient attention to some relevant detail, be unable to think through all of its consequences, fall prey to some bias, etc. But in this respect, our judgments about hypothetical scenarios are no different from those in any other area. The mere fact that our judgments are error-prone does not usually lead us to abolish those judgments, so why should this be different in the case of our intuitions about hypothetical scenarios? Moreover, thought experiments simply draw on our ability to correctly apply our concepts. Since all of our judgments depend on this general ability, we cannot discard it completely if we want to avoid general skepticism.

It follows that if experimental philosophers or any other critics of traditional philosophical methodology aim to discredit the appeal to thought experiments, they have to achieve two things: Firstly, they need to find a way to restrict their target. I.e., they need to specify criteria for determining which kinds of judgments are under attack and which are not. And secondly, they have to provide reasons for thinking that the kinds of

judgments thus specified are unreliable, or at least considerably less reliable than other kinds of judgments we make in our epistemic practice.

Weinberg et al. claim that judgments about 'remote' cases which 'lie beyond our linguistic practice' are particularly unreliable (cf. Weinberg et al. ms.). It is not clear, however, how this is to be understood (cf. also Kipper 2010). A case can be said to lie beyond our linguistic practice simply because we do not usually talk about such cases. But this does not mean that we cannot talk about them or that we cannot reliably apply our concepts to them. On the other hand, that a scenario is remote could also mean that it differs significantly from the actual world. Accordingly, a different line of argument which one could extract from the remark of Weinberg and his co-authors is that our judgments are unreliable if they are about remote possible worlds. One should note that since many of the hypothetical scenarios which are invoked in philosophical practice are quite mundane, that argument cannot discredit the method of thought-experimentation as a whole. But still, it is undeniable that a large number of thought experiments are about rather far-fetched cases. Moreover, it is unclear whether philosophical methodology in general, and in particular conceptual analysis, can do without an appeal to remote cases. How should one, for example, determine necessary and sufficient conditions for the applicability of an expression if one is only able to evaluate nearby cases? So if we were not able to make reliable judgments about remote possible worlds, this would surely be a problem for conceptual analysis. But just like above, it is not clear why we should believe this. For why should the mere fact that a scenario is remote keep us from making reliable judgments about it? On the face of it, embedding a Gettier case into some strange science fiction scenario will not make it more difficult to assess.

Things may not be quite that simple, however. For, the idea that our judgments about remote scenarios are just as reliable as those about everyday cases can only be upheld if one assumes that they are a priori. Let me elaborate: Suppose our judgments about whether some category C is instantiated are based on an (implicit) association of C with feature F. Suppose further that F is only empirically correlated with C. Our everyday

judgments about C could then be highly reliable, depending on the strength and stability of the correlation between C and F and our ability to track F-ness. But our judgments about hypothetical scenarios would not necessarily be trustworthy because the correlation between C and F may fail to hold in these scenarios. This would not imply that all of our judgments about hypothetical scenarios are suspect, nor that all judgments about remote scenarios are. But still, generally it would mean that the more remote a scenario is, the more prone we are to err about it.

Consequently, by holding that our intuitions about cases are based on empirical information, a critic of thought experiments can put our judgments about remote hypothetical scenarios into doubt while leaving our everyday judgments untouched. One may wonder, however, whether this is still a live option at this point. For did it not transpire in chapter 4 that if we are able to make such judgments, then we are also able to do so a priori? But notice that this is not quite the thesis which is at issue here. In chapter 4 I argued that the judgments in question can be justified a priori in principle, i.e. given ideal rationality. This is compatible with the claim that our actual judgments about hypothetical scenarios are only empirically justified or even with the claim that we are only able to base them on empirical information, given the limited cognitive abilities we have. So whether our actual judgments are a priori or not still needs to be discussed.

Let me first note that in order to defend the reliability of judgments of the form 'A → B', one need not demonstrate that the judgment is justified in a way which is independent of sense experience. It suffices to show that the entailment from A to B holds across all possible scenarios.[134] Likewise, if a subject determines whether A is present by tracking feature F, then one has to make sure that F is correlated with A across all scenarios. In this case, F would be an a priori associated property in Jackson's sense. Now recall that in chapter 5, I argued that there is a way of telling whether a given property is a priori associated or not: I proposed that if a certain state of affairs is conceivable, then its negation is not a priori. This entails that if 'A & ~F' is

[134] In this case, one would have even ensured that A → B is even strongly a priori (cf. 2.1.2).

conceivable, then F is not a priori associated with A. Conversely, if 'A & ~F' is inconceivable, then F is a priori associated with A. I argued for this thesis by holding that it provides the best explanation for differences in our judgments about hypothetical scenarios: The fact that we can easily conceive of cats not being animals yet are unable to conceive of a subject who knows a falsehood can be best explained by the assumption that being an animal is only a posteriori associated with 'cat', while truth is a priori associated with 'knowledge'.

One could thus apply the heuristics proposed in the previous chapter to the issue at hand: Take a judgment which is based on the description of a hypothetical scenario in a thought experiment: '$D_{TE} \rightarrow A$'. If 'D_{TE} & ~A' is inconceivable, then the conditions specified in the antecedent epistemically necessitate the consequent; if 'D_{TE} & ~A' is conceivable, then they do not necessitate it – which means that the judgment might not be trustworthy.

In fact, one could argue that the latter proviso is unnecessary because subjects apply stricter epistemic standards when evaluating hypothetical scenarios than in actual circumstances. To illustrate: In our everyday lives, we rarely hesitate to identify persons we are closely familiar with on the basis of their visual appearance. But suppose one invokes a hypothetical case where the subject is visited by a person who looks exactly like her mother and asks the subject whether the person thus described is (or would be) her mother. It is likely that she will deny that there is a clear answer to this question. I therefore think that it is not implausible that we make judgments of the form '$D_{TE} \rightarrow A$' only if we take 'D_{TE} & ~A' to be inconceivable, i.e. only if we think that '$D_{TE} \rightarrow A$' is epistemically necessary. In any case, if it does happen that we judge '$D_{TE} \rightarrow A$' while still finding 'D_{TE} & ~A' conceivable, then it is plausible that the judgment relies on empirical background information. If one is able to identify such a piece of information, call this E, then one can check whether 'D_{TE} & E & ~A' is conceivable, and so on, until one has found antecedent conditions which do epistemically necessitate the consequent.[135]

[135] Note that this is possible because any actual description of a hypothetical scenario will be incomplete.

All that is required for these heuristics to work is that we are able to tell whether a given state of affairs is conceivable or not. This is not a trivial claim. For, if there is to be a connection with epistemic modality, which is itself an idealized notion, the relevant notion of conceivability must be tied to ideal rationality as well.[136] A subject may thus think that some epistemic impossibility is conceivable, for example because she overlooks some deep inconsistency. Nevertheless, I do not see a principled reason to be skeptical regarding our ability to determine whether something is conceivable, provided that we are sufficiently thoughtful.

To determine whether something is inconceivable is a bit trickier. Suppose we want to find out whether '$D_{TE} \rightarrow A$' is epistemically necessary. It seems that in order to show that 'D_{TE} & ~A' is inconceivable, one would have to make sure that it is not verified by any possible scenario. In many cases, this is not difficult: We often find something inconceivable simply because we realize that it is contradictory. But other inconceivabilities are not so easy to detect. It is obvious that one cannot run through every possible scenario to check whether it verifies 'D_{TE} & ~A'. So maybe the following, more feasible procedure is also appropriate: If after a reasonable amount of time and despite serious effort one has failed to come up with a scenario which verifies 'D_{TE} & ~A', then one is justified in believing that it is indeed inconceivable. This presupposes, of course, that we are reasonably good at devising such scenarios. But the fact that counterexamples to proposed definitions are often found quickly within the philosophical community lends at least some support to this assumption.

If the foregoing considerations were correct, then one can draw the following two conclusions: Firstly, we are able to determine if a judgment about a hypothetical case is not epistemically necessary and therefore at least potentially problematic. Secondly, it is possible to avoid such problem cases by specifying the relevant scenarios further. One crucial premise in my argumentation was the claim that the (in)conceivability of a proposition

[136] It can therefore be understood along the lines of Chalmers' 'ideal conceivability' (cf. Chalmers 2002a).

reflects its epistemic (im)possibility. I therefore think a critic of the method of thought-experimentation should reject precisely this claim. This would mean that whether something is conceivable, or deemed to be conceivable, can also be dependent on implicit or explicit empirical background assumptions.[137] Given my discussion in chapter 5, where I pointed out that what we find conceivable does not seem to be influenced by any kinds of empirical beliefs, the critic of thought experiments would need to provide some reasons to support such a claim.

Richard Hare attempts to provide such reasons for thinking that our intuitive moral judgments about hypothetical scenarios are unreliable. Since his general strategy is very interesting and might be applied to other areas as well, I think it will be worthwhile to have a closer look at Hare's account.

6.2.2.1 The two-level model

A central characteristic of Hare's theory is his distinction between two levels of moral thinking, the intuitive level and the level of critical thinking (cf. e.g. Hare 1982). Most of our everyday moral judgments are intuitive. These intuitions are first of all shaped by our moral education, in which we are taught what Hare calls 'prima facie rules' such as the rule not to lie or not to harm innocent persons. One probably should not require, however, that (all of) these rules are explicitly taught to us. At least some of them may just implicitly underlie the way we are taught to distinguish, say, right actions from wrong ones. For the same reason, a subject need not be able to spell out the prima facie rules which shape her intuitive judgments; and one should not understand these judgments as the result of the conscious application of a rule – otherwise, they could hardly be called intuitive. The second level of moral thinking works quite differently. A subject who

[137] Alternatively, an opponent could hold that the fact that judgments about hypothetical scenarios are (sometimes) empirically based is independent of whether or not the relevant scenarios are insufficiently described.

makes moral judgments exclusively on the basis of critical thinking[138] considers an action to be morally good, or required, if and only if it maximizes the satisfaction of the preferences of those affected, without exception. Some of the actions recommended by an ideal critical moral thinker will infringe the prima facie rules from the intuitive level; for example when the action which maximizes the expected satisfaction of the preferences of all affected involves harming an innocent person. But it is crucial for Hare's view that in such cases, the moral verdict arrived at by proper critical thinking is always correct. There are thus cases in which the prima facie rules on which we rely in our everyday moral judgments lead us astray.

This does not mean, however, that these rules are not useful or that our moral intuitions are generally unreliable. First of all, if we always had to determine the expected satisfaction of preferences for each alternative before acting, the likely result would be complete paralysis. The prima facie rules are thus an important means for saving cognitive resources. Moreover, they can be of great cognitive value in the sense that their application leads to the production of true moral beliefs in a highly reliable way. For example, it is very plausible that an action which involves doing harm to an innocent person will rarely maximize the expected satisfaction of preferences. One cannot expect such rules which are supposed to be easy to memorize and easy to apply to get every possible case right. But this is not necessary. A rule is of epistemic value if its application leads to true moral beliefs with sufficient reliability in those kinds of circumstances we actually encounter.

This latter point is crucial for Hare's critique of the use of thought experiments: The prima facie rules which shape our moral intuitions are designed to guide our everyday moral judgments. It is thus to be expected that they will fail frequently if applied to far-fetched cases. And this is precisely the reason why Hare thinks that thought experiments in ethics which invoke remote hypothetical scenarios are flawed (cf. Hare 1982, 182, 194).

[138] Hare calls such a subject an 'archangel' (Hare 1982, 44ff.).

It is evident that Hare's position is in line with the type of critical view I outlined above: The prima facie rules which reliably guide our moral judgments are not suitable for the evaluation of remote scenarios. One could say that these rules are accurate because they rely on robust empirical, though not epistemically necessary, correlations, such as the correlation between being a morally bad action and being an action which involves harming an innocent person.[139] This interpretation is supported by Hare's claim that moral intuitions are no 'linguistic' intuitions (cf. Hare 1982, 10ff.). It is natural to go one step further and propose that the prima facie rules which are imparted on us in our moral education at least sometimes determine what we find conceivable. For example, it might be that we find it inconceivable that killing an innocent person can be the right thing to do because such an action violates these rules so blatantly.

Hare thus offers an explanation for why we should not trust thought experiments in ethics in particular. However, as I already indicated above, one might be tempted to apply a 'two levels model' to other areas as well, for example to epistemology. To see how this could work, consider the following two well-known thought experiments, both of which are supposed to speak against reliabilist accounts of justification. Let me start with the so-called 'new evil demon problem' (cf. Cohen & Lehrer 1983): Suppose there is a subject who is deceived by an evil demon. Although it seems to her that she perceives an external world, and although her perceptual experiences are as clear and coherent as ours, in fact all of these experiences are illusory. In such a case, the subject would have no perceptual knowledge because all of her perceptual beliefs would be false. But many think that intuitively, she would still have justification for her perceptual beliefs. At the same time, these beliefs have been produced by an unreliable process. The thought experiment is thus supposed to demonstrate that the reliability of the belief-forming process is not a necessary condition for a belief to be justified.

[139] Since Hare himself is a moral anti-realist and non-cognitivist, he would not take these to be correlations between moral and non-moral *properties*.

By contrast, BonJour's thought experiment about the clairvoyant Norman aims to show that reliability is not a sufficient condition for justification (cf. BonJour 1985, 41): Norman occasionally finds himself having beliefs about the current location of the president of the United States. He has no evidence that these beliefs are correct or that he has any supernatural abilities. But in fact, Norman is a highly reliable clairvoyant; his beliefs about the president's location are usually perfectly accurate. Now BonJour claims that Norman's current belief that the president is in New York, in spite of having been produced by a highly reliable process, is not justified.

Reliabilists have responded to the challenge raised by these thought experiments in a number of ways. The relevant kind of response in our context involves denying the correctness of the intuitions in question. This claim could be supported by reasons akin to those Hare gave to undermine the reliability of moral intuitions. Let us assume that there are two levels of epistemic thinking: the intuitive level and the critical level. On the critical level, the question of whether a belief is justified is answered only by determining whether the process by which the belief was produced is reliable. However, in our epistemic practice this is not always feasible. For this reason, such judgments are often made on the intuitive level. But how exactly is this intuitive level to be understood? Recall that according to Hare, moral intuitions are shaped by prima facie rules which are acquired in a subject's moral education. One could thus hold that we rely on similar kinds of rules in judging whether a belief-forming process confers justification or not. It is admittedly rather implausible that there is such a thing as an 'epistemic education', but this need not be a problem because there are a number of alternative ways in which the rules underlying the intuitive level of epistemic thinking may be acquired. Furthermore, just like in the moral case, one should not require that our judgments about justification result from a conscious application of these rules, or even that they can be spelled out by a subject.

The relevant rules may come in different forms. There may for instance be rules which explicitly refer to familiar types of processes, such as 'A belief which is based on wishful thinking is unjustified' or 'A belief which is

based on sufficiently clear and distinct sense perception is, absent defeaters, justified'. In fact, the view resulting from such a construal of the intuitive level is quite close to the one advocated by Alvin Goldman in *Epistemic Folkways and Scientific Epistemology* (1993). There, Goldman argues that our intuitive judgments about hypothetical cases are based on stored lists of epistemically 'good' belief-forming processes like sight, memory or reasoning in 'approved' ways and bad ones, such as guessing or ignoring contrary evidence (cf. Goldman 1993, 275f.). The items on this list have been categorized either on the basis of a subject's own past experiences or of testimony from others.

I think there should also be other kinds of rules, inter alia to enable a subject to evaluate novel types of belief-forming processes, for example a rule to the effect that there should be some intelligible connection between the relevant process and the truth-maker of the proposition believed (at least when contingent propositions are concerned). But the details do not matter too much here. What is more important is the general idea that, just like in the moral domain, our intuitive judgments are correct most of the times because the rules underlying them rely on reasonably stable empirical correlations. For example in those circumstances we usually encounter, a belief which is based on wishful thinking is plausibly never justified, so a rule to that effect seems very useful. According to the reliabilist, this is simply because wishful thinking is a very unreliable process. Furthermore, she will argue that because there are possible circumstances in which wishful thinking is a reliable process, there are also circumstances in which a belief which was produced by wishful thinking is justified. If this seems inconceivable to us, then it is due to the fact that our judgments about possible cases are shaped, or at least influenced, by rules which are mostly accurate in actual circumstances, but which can fail in remote hypothetical ones. A reliabilist who adopts this line of argument could thus claim that the rules underlying the intuitive level of epistemic thinking exist precisely because they are good indicators for the reliability of a belief-forming process and thereby for the justifiedness of the corresponding belief(s).

The reliabilist's response to the problem raised by the two thought experiments outlined above is now obvious: She will say that our intuitions about the scenarios in question go wrong because they rely on rules which are not suitable to be applied to such cases. The fact that we consider a belief which is based on clear and distinct (and coherent) sense perception as justified can ultimately be traced back to the fact that such sense perception is reliable. But if we make intuitive judgments about scenarios where it is not, we are likely to be misled. Likewise, since clairvoyance is known to be unreliable, it is reasonable to assume that there is a corresponding rule; for this reason, our intuition tells us that a belief based on clairvoyance is not justified. But since the case of Norman is precisely one where clairvoyance is reliable, our intuitions about it go wrong.

There are two questions which need to be addressed here: The first is whether the two-level model just introduced gives an adequate account of our judgments about actual and hypothetical cases in ethics, epistemology or elsewhere. The second question is what implications it would have for conceptual analysis if the two-level model applied to some domain or other. In order to evaluate the challenge raised by the two-level model, it will be important to keep the present dialectical situation in mind. Let me therefore give a brief recap of the current state of the discussion, before turning to the two questions just identified.

In chapter 4 and in 6.1, I argued for (CJ++) which states that we are able to determine the extension of an expression with respect to any hypothetical scenario a priori if we are given a canonical description of that scenario. However, (CJ++) involves some idealizations. So if the thesis, and thus two-dimensionalism itself, is to serve as a basis for conceptual analysis as a philosophical method, then we need to assume that our actual judgments about hypothetical scenarios are, at least by and large, reliable as well. The most promising way to put this assumption into doubt is to claim that our judgments about possible cases, including our judgments about what is conceivable, depend on empirical background beliefs. For if this were so, then our judgments about remote possible cases would go systematically

wrong with respect to those worlds where the relevant empirical correlations fail to hold. Hare's two-level model of moral thinking can be understood as a version of such a view. Furthermore, the same kind of model might be applied to epistemology and maybe to other areas as well. It therefore seems as though the two-level model does pose a potential threat to conceptual analysis. Let me thus turn to the question of how plausible it is as an account of our judgments in various areas.

As I mentioned above, a proponent of such a model faces two challenges: Firstly, we saw above that at least many of our conceivability judgments and of our judgments about hypothetical cases in general are not dependent on empirical beliefs. So if one argues that our judgments about some subject matter work differently, one should offer an explanation why these particular judgments are exceptional. Secondly, from what I said so far it does not become apparent why one should believe that an action is good if and only if it maximizes the expected satisfaction of preferences of those affected, or that a belief is justified if and only if it was produced by a reliable process. A proponent of a two-level model therefore has to find a way to establish the principles which underlie the critical level.

On the first challenge: In chapter 4 I argued that our conceivability judgments are independent of empirical background beliefs. One example I gave was the fact that we can conceive of cats being robots despite our well-entrenched belief that they are animals. So why should the case of moral judgments and/or judgments about justification be so different? More generally, a proponent of a two-level model has to assume that we confuse mere epistemic criteria for the presence of a certain category with semantic or metaphysical criteria. But again, we usually do not seem prone to do this. Even though, for instance, we recognize many kinds of things by sight, we do not tend to think that something which looks like an X is or would be an X in all possible cases. So again, if one claims that our judgments in some area are confused in this way, then this calls out for an explanation. Nevertheless, Hare does not offer such an explanation, which renders his account somewhat incomplete.

It is striking that both of the domains just discussed are normative domains. This fact might be used as a starting point in an attempted explanation. But I am at loss to offer a reason why we are supposed to be prone to confuse epistemic criteria for semantic ones when evaluating normative issues. Of course, the explanation required might be a purely psychological one, so one could argue that it is not appropriate to expect the proponent of a two-level model to provide one herself. Nevertheless, the fact remains that as long as there is no such explanation in sight, there is a notable gap in her account.

The second challenge to the two-level model is even more pressing. It is obvious that a two-level model can only be established if there are sufficiently strong grounds for accepting the rules which supposedly govern the critical level. In fact, Hare goes at great length to show that an ideal moral thinker – an archangel – would always strive to maximize the expected satisfaction of preferences. Hare's line of reasoning has two steps: First he argues that moral judgments are prescriptive, universalizable and have overriding power[140]. Then he tries to show that from these three features, his version of utilitarianism can be derived.

Hare's arguments for these two steps are debatable. But the potential flaws in his argumentation will not concern me here. The point I want to make concerns the basis of the three features of moral judgments which Hare identifies. According to him, these features stem from logical properties of certain moral words (like 'should' and 'must') which figure in the relevant judgments (cf. e.g. Hare 1991, 11). It is surely no stretch to say that he identifies the relevant features of moral judgments by way of conceptual analysis. This is an important insight because it shows that even if it turns out that our intuitions about (remote) possible cases (in a certain domain) are unreliable, this need not be worrisome for conceptual analysis. For note that on Hare's account, it is possible to determine the truth-value of a moral

[140] This latter principle states that when a reason to act which arises from a moral judgment conflicts with one arising from a judgment about a different domain – for example an aesthetic one – the moral reasons will always be stronger than the non-moral ones, i.e. they will override them.

statement even with respect to the remotest possible worlds, by applying the principles underlying the critical level of moral thinking. And importantly, these principles can be determined with the help of conceptual analysis. So there are, fortunately, other ways to do conceptual analysis than by invoking hypothetical scenarios.[141]

I conclude that a two-level model can only pose a serious problem to conceptual analysis if the principles constituting the critical level are established in a way which is incompatible with this method. Of course, there are a number of philosophers who hold such a view. Kornblith, for instance, claims that his favored theory of knowledge and justification is supported by empirical evidence. I find the idea that philosophical theories can be established by purely empirical means implausible; at some point in our inquiry, we have to rely on our conceptual competence. But note that even if Kornblith was right in thinking that we do not, this would be insufficient to repudiate conceptual analysis. In the worst case, it would show that conceptual analysis is a redundant method. What is currently at issue, however, is whether it is inapplicable to some area or other. In order to demonstrate this, one would have to show that the principles constituting the critical level cannot be justified on the basis of conceptual analysis. Kornblith, along with other naturalists, does in fact hold that philosophical theories cannot be established by a priori means. However, another qualification is called for here: Naturalists usually say that a priori knowledge cannot be had even in principle. But such a position is no longer an option at this point of the discussion, since I already argued in chapter 4 and in 6.1 that if we had ideal rational capacities, we would be able to determine the extension of an expression with respect to every possible world a priori. The question we are dealing with at this point is whether we also have this ability in practice. What the opponent of conceptual analysis has to argue here is therefore this: Even though the principles which constitute the critical level of moral/epistemic/… thinking can be established a priori in principle, in our epistemic practice we have to rely on empirical evidence. But such a view does not seem particularly

[141] I will say more about these other ways later on.

appealing, because it involves a rather skeptical attitude towards our rational capacities. This skepticism appears even less well-founded once one realizes that there are several ways to do conceptual analysis and that we thus need not rely on our intuitions about hypothetical cases – as was shown above. And in any case, the two-level model does not do anything to support such a skeptical view.

To sum up: The two-level model provides an interesting framework for understanding the way in which we arrive at judgments about moral, epistemic and maybe also other kinds of issues. It thereby challenges the reliability of our intuitions in particular about remote scenarios. There are two main problems for such a model which are yet to be solved: Firstly, it remains to be explained why we are liable to confuse epistemic criteria for semantic ones, given that we do not tend to do so in other areas. And secondly, the model remains unfounded unless the principles on the critical level have been established.
At this point, it becomes apparent that the two-level model can only undermine conceptual analysis if one holds that in our epistemic practice, these principles can be justified exclusively by means other than conceptual analysis, even though they can be justified a priori in principle. But this particular view is not supported by the two-level model itself, nor are there any other reasons to endorse it in sight. I thus conclude that while the two-level model may put our intuitions about remote possible cases into doubt, it should nevertheless not be considered as a threat to the method of conceptual analysis.

6.2.2.2 Approaching ideal judgments

Finally, let me say a few words about how one might go about trying to approach the ideal judgments involved in the scrutability thesis in our epistemic practice. Recall that according to (CJ++), we are able to determine the extension of an expression with respect to any hypothetical scenario a priori. The underlying idea is that these judgments mirror what

we associate with an expression. In turn, the associated properties determine the expression's extension. Nevertheless, it is our ideal judgments which determine an expression's extension; our actual judgments about hypothetical cases can go wrong in a number of ways. This raises the question of how we find out what we would say if we were ideally rational.

Suppose that we are confronted with a hypothetical scenario and are asked to decide whether a specific term is applicable to it or not. How are we to proceed if we want to do anything we can to get the answer right? It surely helps to make sure that we have completely understood the description of the scenario and do not miss any relevant details. Furthermore, it will often be necessary to draw some relevant conceptual, logical or other kinds of a priori inferences from this description. In the majority of cases, if we have taken all of this into account we can be confident that the resulting judgment is correct. But some cases are harder to evaluate. So what should we do if we are not sure what to say even after careful scrutiny or if we have reason to distrust our intuition about the case? We could have an even closer look at the scenario, think harder about our judgment and hope that we are thus able to sort out any residual doubts. But this will not always help.

A different route to determining the extension of an expression with respect to a scenario is this: First one identifies the expression's application conditions, then one checks whether these application conditions are met in the scenario at hand. This is essentially the way in which Hare proceeds to validate moral judgments. As I noted above, he believes that preference utilitarianism follows from logical properties of moral terms. Once this has been established, one can determine whether an action is good by checking whether it maximizes the expected satisfaction of the preferences of those affected. The critical step here, at least from a philosophical perspective, is obviously the first one. Unfortunately, Hare does not say very much about how he managed to identify the relevant logical properties. Presumably, it just seems intuitively obvious to him that moral judgments have the

features in question.[142] Those intuitions would be about the necessary conditions for the applicability of moral terms, for instance 'If two situations are identical in their universal descriptive properties, then we should make identical moral judgments about them as well'. It is indeed plausible that we often have intuitions of this kind: 'If something is known, then it is true', 'If something is a robot, then it is an artifact', etc. We also have, which is now rather trivial, intuitions about sufficient conditions for the applicability of a term, for example 'If something is known, then it is true' as a sufficient condition for truth. Then there are intuitions about necessary and sufficient conditions, one might call these intensional intuitions, such as 'Something is a bachelor if and only if it is an unmarried man'.

Skepticism about these kinds of intuitions aside, it is thus possible to evaluate a hypothetical scenario without relying on our intuitions about this scenario itself. And I think there are further ways to do this. Sometimes, the best way to evaluate a possible case is by looking at other scenarios. For example, people sometimes misinterpret possessive pronouns. Upon hearing someone say 'This is my cat', they may respond 'It isn't yours; a cat cannot be anyone's property'. In such cases, it usually suffices to point out that it is perfectly normal to say 'my sister', 'my mother' or 'my boyfriend' to convince one's interlocutor that possessive pronouns, in spite of their name, do not indicate possession or ownership. Here is another simple example: Is the sentence 'I went out and had fun' false if the fun occurred before one went out? Someone inclined to affirm this will probably change her mind when one presents her with a sentence like 'I bought a boat and ate a pizza – but I don't recall in which temporal order'.

There are thus a number of methods for evaluating hypothetical scenarios. This raises the question of how conflicts are to be resolved when these methods lead to different verdicts about a possible case. Often, we let our intuition about the case itself overrule a judgment arrived at in a different way. Classic examples are the Gettier cases. The traditional analysis of

[142] Note that Hare does not repudiate intuitions in general. The intuitions in question might thus be trustworthy because they are 'linguistic' intuitions.

knowledge which had arguably been supported by intensional intuitions and the fact that it is in accord with a large number of cases was overturned by our intuitions about two hypothetical scenarios. This seems to be a common pattern in the philosophical practice: A univocal intuition about a case is often taken as a decisive counterexample for a proposed analysis. However, our intuitions about a scenario are not always taken as the ultimate verdict, as is witnessed by the examples I gave above concerning the usage of possessive pronouns and 'and'. For, notice that in these cases, there is no direct conflict between our intuitions about hypothetical scenarios. Rather, the judgments about the original scenarios are retracted because they are incompatible with a unitary treatment of the semantics of the expressions in question.[143]

It is important to note that in most cases, the conflicting intuitions do not have to be weighted arbitrarily or on the basis of further intuitions. Often, they can be explained and integrated into a broader account; in other cases, they can be explained away. Take again the example from above: It is plausible that in a conjunction involving two events, the temporal order in which the events occur is implicated by the order of the conjuncts. Since pragmatic implicatures can be easily confused with semantic implications, it is natural to suppose that the sentence in question is mistakenly considered as false because it is infelicitous. As I noted above, this reading is also supported by the fact that it allows for a unitary reading of the semantics of 'and': If we held that the given sentence is false and adhered to our intuitions about other cases as well, we would be committed to saying that the temporal order is sometimes entailed and sometimes not. Now consider by contrast the Gettier cases: The intuition that the subjects would not have knowledge is clearly not epistemically basic or inexplicable. Rather, what all of the counterexamples to the analysis of knowledge as justified true belief have in common is that the subject arrives at a true belief by luck. So there is a definite pattern behind our intuitive judgments about the Gettier cases. Furthermore, the distinction

[143] For a rather extreme view regarding the importance of such considerations, cf. Weatherson 2003.

between a justified true belief which depends on a lucky accident and one which does not is plausibly epistemically relevant (cf. also Jackson 1998a, 36).
These examples show that we often implicitly assume – rightly, I think – that the application conditions of our expressions are based on intelligible patterns, and that they mark distinctions which are worth making or are at least not completely arbitrary. This background assumption enables us to increase the range of criteria for evaluating hypothetical scenarios.
Determining an expression's extension with respect to a possible world can thus be a process akin to theory building: There are a large number of potentially relevant data to be considered and appropriately weighted. The goal is to develop an account which makes the best sense of all of these data (cf. also Jackson 1998a, 36). In general, our intuitions about the possible cases in question and about necessary and/or sufficient conditions for the expression's applicability will be given most weight in such a process, since the default assumption should be that these intuitions stem from our conceptual competence.

This chapter was mainly dedicated to various issues concerning the role of scrutability theses both for two-dimensionalism and for conceptual analysis. In 6.1, I first tried to clarify the relation between (CJ++) and the idea that linguistic expressions are associated with an a priori accessible semantic value. In order to do so, I proposed a slightly more substantial understanding of the associated properties than the one I had assumed so far: I argued that they should be understood as real or at least possible features of the world, such that a term is applicable to a situation if and only if that associated feature is instantiated. These associated properties enable a subject to determine the expression's extension on the basis of a description of a scenario which does not contain that expression. If one combines this idea with the thesis of metaphysical plenitude, one arrives at the claim that knowledge of the distribution of fundamental properties enables a subject to determine the extension of any given expression, on the basis of the associated properties. I concluded that even though there is no

actual subject who possesses the required cognitive capacities and the concepts which would be necessary to describe the distribution of fundamental properties, the relation between the thesis that expressions are a priori associated with properties and the scrutability thesis is nevertheless evident.

Next, I discussed arguments for and against (CJ), which is a comparably weak version of the scrutability thesis. It transpired that there are a number of good reasons to endorse the thesis, but also some potential counterexamples. In my view, the most powerful of these considerations in favor of (CJ) are epistemic in nature: If one denies outright that we have the ability to determine the extensions of our expressions if given sufficient information about the world, one faces global skepticism. Likewise, if one makes the weaker claim that we do not have this ability when a certain kind of expression is concerned, i.e. in a specific domain, then it will not be easy to avoid skepticism about this particular domain. Therefore, accepting one or more of the potential counterexamples to scrutability would have its costs.

Recall that in chapter 4, I had argued that if one presupposes (the prima facie plausible) (CJ), then a convincing case can be made for (CJ++), which suffices to establish the central two-dimensionalist thesis that every linguistic expression is associated with a primary intension. By providing support to (CJ), 6.1 thus completes my case for (CJ++) and thereby for two-dimensionalism.

In 6.2, I discussed the role of idealizations in the scrutability theses and their relation to our epistemic practice. I first addressed, in 6.2.1, the question of whether primary intensions are insufficiently fine-grained because they rely on an idealized notion of epistemic possibility, thus leaving no room for a subject's limited rational abilities. There are ways to allow for more fine-grained distinctions, either by introducing structured intensions or by watering down the notion of epistemic modality. I also noted, however, that it is far from clear whether it is a requirement for a semantic theory to take non-ideal reasoning into account. In particular, any

version of two-dimensionalism which is based on a non-ideal version of epistemic modality will fail to satisfy the thesis of metaphysical plenitude and thus be only of limited use for conceptual analysis.

Section 6.2.2 was dedicated to the question of how reliable our actual judgments about hypothetical scenarios are. Since it is neither very attractive nor plausible to challenge the reliability of all of our judgments about cases, a critic will have to find a way to restrict the target of her attack. A promising way to do this is provided by Hare's two-level model of moral thinking which might also be applied to other areas. According to the two-level model, our intuitions about hypothetical scenarios are not a priori, but rather based on empirically shaped 'prima facie rules'. These rules may be highly reliable when applied to actual cases, but fail frequently with respect to remote cases. It is not clear how plausible the two-level model really is as an account of our judgments about moral, epistemic or other kinds of matters. But in any case, we saw that such a theory need not be a threat to conceptual analysis. The only available ways to undermine conceptual analysis on the basis of such a model would involve holding that it is in principle possible to make such judgments reliably and a priori, but that we are not able to do so in our epistemic practice. But as I argued above, this kind of view is not particularly attractive.

Finally, I discussed how one might improve on our prima facie verdicts about possible cases. I argued that the best way to evaluate a hypothetical scenario is not always to consider that particular scenario. Sometimes, it is necessary to engage in a process similar to theory-building to develop an account which integrates extensional intuitions and intuitions about the relevant expression's application conditions in the simplest and most coherent way.

7 The trouble with definitions and the aims of conceptual analysis

Up to this point, a considerable part of this work has been dedicated to scrutability – the thesis that if we are given sufficient, nontrivial information about the character of a world, we are able to determine the extensions of our expressions. The following version of the thesis was mainly defended in the first part of chapter 6:

(CJ) If a subject possesses a concept and has unimpaired rational processes, then sufficient empirical information about the actual world puts a subject in a position to identify the concept's extension.

Before that, in 4.1, I argued for the conditional claim that if one endorses (CJ), then one should also accept (CJ++):

(CJ++) If a subject possesses a concept and has unimpaired rational processes, then sufficient information about any given scenario puts a subject in a position to identify the concept's extension with respect to that scenario a priori.

Since these theses involve a number of idealizations, I also discussed, in the second part of chapter 6, to what extent we are able to make such judgments in practice. My verdict was overall rather optimistic, since it turned out that there are at least no good reasons to be skeptical about our ability to evaluate hypothetical scenarios.

One obvious reason for dedicating so much space to a defense of scrutability is that if one manages to establish (CJ++), one has thereby shown that linguistic expressions are associated with an a priori accessible semantic value. The relevance of this result for the prospects of conceptual analysis should be obvious. A second, more specific reason for making scrutability a central theme is that it is ideally suited to provide the foundations for the way conceptual analysis is usually pursued, namely via thought experiments, i.e. the construction and evaluation of hypothetical scenarios.

These observations leave open, however, what kinds of philosophical insights one may hope to gain on the basis of our grasp of primary intensions and our associated ability to evaluate hypothetical cases. This chapter will therefore be dedicated to what conceptual analysis might aim at.

A common source of skepticism towards conceptual analysis is the absence of successful analyses, as witnessed for example by the following remark of Timothy Williamson:

> Attempts to analyse the concepts *means* and *causes*, for example, have been no more successful than attempts to analyse the concept *knows*, succumbing to the same pattern of counterexamples and epicycles. The analysing concept does not merely fail to be the same as the concept to be analysed; it fails even to provide a necessary and sufficient condition for the latter. The pursuit of analyses is a degenerating research programme. (Williamson 2000, 31)

Chalmers and Jackson's response to this kind of objection is based precisely on considerations concerning scrutability. They stress that proposed definitions are usually found to be inadequate because they fall prey to counterexamples, i.e. hypothetical (or sometimes actual) cases with respect to which the purported equivalence between analysandum and analysans fails to hold. But to classify a case as a counterexample for a proposed analysis is simply to apply the general ability which is postulated in the scrutability theses. Consequently, the rejection of a proposed analysis due to a counterexample is itself based on conceptual analysis (cf. Chalmers & Jackson 2001, 320ff.).

I agree with Chalmers and Jackson that the (alleged) fact that successful analyses of philosophically relevant expressions are nowhere to be found cannot in itself speak against conceptual analysis, for the reasons they mention. However, their defense of conceptual analysis only applies to the *process* of doing conceptual analysis (cf. chapter 1), since evaluating whether a given expression applies to a hypothetical scenario or not is hardly an end in itself. Any assessment of the value of conceptual analysis

which does not also consider its potential aims and the prospects for achieving them would thus be incomplete.
In this chapter, I will therefore outline a number of goals which people have aimed (or might aim) at by doing conceptual analysis. In doing so, I will also try to clarify whether these ways of applying conceptual analysis to gain philosophical insights depend on explicit analyses, i.e. definitions, and if so, how. The second part of the chapter will deal with the question of whether it is realistic to expect that we can provide such definitions.

7.1 The aims of conceptual analysis

7.1.1 Definitions – complete, partial, and absent

Conceptual analysis as it has traditionally often been conceived and as it still is often conceived does aim at explicit analyses. Typically, such analyses provide necessary and sufficient conditions for the applicability of an expression in the form of a definition, as in 'bachelor $=_{df}$ unmarried male adult human' or 'necessarily: x is a bachelor iff x is an unmarried male adult human'. This fits well with a conception of philosophy according to which it is (often) occupied with 'What is X?' questions. Here, the idea is that an analysis of the concept 'X' will also provide insight into the nature of X. But as should have become clear from preceding chapters, conceptual analysis, understood as an a priori enterprise, can only reveal the nature / metaphysical essence of X if the relevant term 'X' is epistemically transparent: For other kinds of expressions, the metaphysical application conditions cannot be determined a priori.
In a case where we are interested in the nature of the referent of an expression which is not epistemically transparent, conceptual analysis can obviously not do the whole job. It may still have some role to play, though. Since an expression's primary intension determines its secondary intension (in a context), such an investigation should at least in some sense rely on our grasp of the relevant concept (cf. also my discussion on the role of primary intensions in determining the subject matter in 4.2).

Nevertheless, it would be unreasonable to think that in order to reveal the nature of X in a case where 'X' is epistemically opaque, we have to provide an explicit analysis of the expression. Lavoisier surely did not have an analysis of the term 'water' at hand when he discovered that water is H_2O. An implicit grasp of the concept (together with accurate sense perception) was sufficient for him to know that he was experimenting with samples of water.
Things are different if 'X' is epistemically transparent. In such a case, conceptual analysis alone can reveal X's nature or essence (leaving aside worries about a possible divergence between a thing's necessary properties and its essence, cf. Fine 1994). But here, rather trivially, one does require an explicit analysis of 'X'.
In the remainder of this section, I will be concerned with the question of how conceptual analysis can be useful where it yields only partial definitions or no explicit analyses at all.

For many purposes, it can already be helpful to have a partial definition of an expression. This can mean, for instance, that one has managed to identify either one or more necessary conditions or just sufficient conditions for the applicability of the expression. Let me consider the latter case first:
Take a description of a thought experiment D_{TE}. If one (correctly) judges D_{TE} to imply X, then one has already identified sufficient conditions for the presence of X – if 'X' is semantically neutral, then these are both epistemically and metaphysically sufficient conditions. Since the description of a thought experiment will usually comprise a large number of sentences invoking various kinds of facts, such sufficient conditions are not the sort of thing that usually enters into an analysis. Nevertheless, identifying them may provide important insights, for example into the relation between different kinds of properties. To illustrate: Take the phenomenal fact P. If a physicalist could show that P is implied by a description of a scenario which only invokes physical (or, say, functional) facts, this would be a huge success. To take this idea one step further: If P

is a conjunction of all phenomenal facts and D_{Phy} a (partial) description of the physical state of the world, then many would say that showing that D_{Phy} implies P amounts to establishing physicalism. This might even be so if there are no necessary physical conditions for the obtaining of any phenomenal facts to be had – in the case that the phenomenal facts could have also been realized by non-physical facts.
Accordingly, being able to identify sufficient conditions for the applicability of an expression, in whatever format, can be of major philosophical importance. This is not to say that finding less inclusive sufficient conditions is not generally preferable: For example, it would be good to know not only that phenomenal facts are entailed by physical facts, but which physical facts they are entailed by.[144] In any case, we have thus already determined one potential use of conceptual analysis which does not depend on (complete) explicit analyses.

Similar considerations apply to cases where we can give necessary conditions for the applicability of an expression: To start with, when one has identified sufficient conditions for ~X, then one can trivially extract necessary conditions for X. For example, when we have a description of a hypothetical scenario D_{TE} to which 'X' does not apply, then $\sim D_{TE}$ is a necessary condition for X. Such kinds of necessary conditions will typically not be very useful. However, this is not (merely) due to $\sim D_{TE}$ yielding a negative condition, as can be seen from the fact that many would consider it as an important insight that for a belief to count as knowledge, it must *not* be based on a falsehood. Rather, the problem is that D_{TE} contains so many facts that little information is provided by the recognition that their conjunction has to be absent for 'X' to be applicable.
In many cases, however, it is possible to determine informative and positive necessary conditions. Take the case of knowledge, which is generally considered as a standard example for the failure to provide

[144] This may well be a merely theoretical distinction. It is plausible that in practice, it is not possible to show that D_{Phy} entails P without being more explicit about the kinds of physical facts which do so.

explicit analyses.[145] There is widespread consensus even among those who say that the term 'knowledge' cannot be defined that truth, and being believed, are necessary conditions for a proposition to be known. In the case where the expression in question is epistemically transparent, determining necessary application conditions can help to reveal various kinds of necessary truths, even in the absence of complete analyses. Where opaque expressions are concerned, being equipped with (epistemically) necessary conditions will still allow one to identify inferential connections between our concepts.

I just discussed how conceptual analysis can be valuable where it only delivers partial analyses. Let me end this section by briefly addressing how conceptual analysis could contribute to philosophical progress in the absence even of such partial analyses:
According to ordinary language philosophers, many philosophical problems can be resolved by considering the usage of relevant linguistic expressions. This kind of analysis need not lead to an explicit definition at any stage. As many of its advocates have an agenda that is quite alien to the account I have tried to defend, I do not want to delve too deeply into the details of ordinary language analysis. But still, there are ways to use conceptual analysis which are in the spirit of ordinary language philosophy and which are clearly compatible with two-dimensionalism. For example, one could examine the usage of an expression to resolve disputes which are merely verbal: If we discover that your evaluation of some key scenarios involving an expression which is relevant to our disagreement differs from mine, even if both of us have fully understood the case and thought carefully about it, this may suggest that the disagreement is just verbal. This kind of conceptual analysis only needs to assume scrutability and does not have to rely on any kind of explicit analysis.

[145] But see my discussion in 7.2 below.

7.1.2 Reductive explanations

According to Joseph Levine, Chalmers and Jackson, and also Jaegwon Kim, conceptual analysis plays a crucial role in reductive explanations (cf. Levine 1993; Chalmers & Jackson 2001; Kim 2005, 111ff.). The role of a priori scrutability in reductive explanations, in which higher-level phenomena are explained by invoking lower-level facts, was already briefly discussed in the previous chapter. There I also mentioned that Block and Stalnaker reject Chalmers and Jackson's claim that a reductive explanation requires that the higher-level facts are a priori entailed by the lower-level facts.

But assume that both the explanandum and the explanans are couched in epistemically transparent terms. Then, given what I have argued in the preceding chapters, it is clear that for a phenomenon to be reductively explainable, it has to be a priori entailed by the facts in the explanans. This is because the higher-level phenomenon has to supervene on the relevant low-level facts, i.e. it has to be metaphysically determined by them. For otherwise, there could be variation in the higher-level facts which is independent of what happens on the lower level, and thus there would at least be some aspect of the higher-level phenomenon which cannot be explained by these lower-level facts. And if, as we have assumed, only epistemically transparent expressions are used, metaphysical supervenience coincides with a priori supervenience.

Let me bring in another plausible constraint on reductive explanations: Given the lower-level facts, it must be intelligible/transparent to us why a particular higher-level fact obtains (cf. also chapter 6). To illustrate: Assume that L_1 is a lower-level state and H_1 is the corresponding higher-level state, i.e., the higher-level state which obtains when the system in question is in L_1. Now suppose further that from our (epistemic) perspective, L_1 seems as compatible with H_1 as with H_2 or H_3. In other words, L_1 makes the occurrence of H_1 just as likely as that of H_2 or H_3. In such a case, it hardly makes sense to say that L_1 can explain the presence of H_1.

If one adds together what has just been argued, then it is natural to postulate the following two constraints on a successful reductive explanation of H-facts on L-facts: a) the H-facts must be a priori entailed by the L-facts; b) it has to be shown that a) obtains.[146]

Nevertheless, both of these constraints are contestable. Take a) first. The argument for a priori entailment was based on the premise that both the explanans and the explanandum are couched in epistemically transparent expressions. With respect to the explanans, this assumption seems reasonable: It would be strange to give an explanation using terms whose metaphysical application conditions are not transparent to us. However, it is hard to see why one would not want to explain phenomena which are described in opaque terms. As an example, take the disposition of water to form drops which was already mentioned in the previous chapter. One might want to explain this feature by drawing on facts about the behavior of aggregates of H_2O molecules in certain conditions. But, since 'water' is epistemically opaque, it is not a priori that if H_2O molecules behave in such-and-such a way, then water behaves in such-and-such a way. To draw this inference, one at least needs the additional empirical premise that water is H_2O.

There is a natural response to this objection: The fact that water is H_2O is itself a priori entailed by low-level features. There is thus no need to invoke an additional empirical premise. But I think this response fails. I do agree that water's being H_2O can be a priori derived from low-level facts – in fact this is just an instance of the a priori scrutability thesis (CJ++) (or simply of (CJ_{ap}) which says that all truths about the actual world are a priori entailed by its canonical description). However, I do not believe that the identity of water and H_2O can be a priori derived from the facts invoked in the explanans of the relevant reductive explanation. For consider what one would need in order to draw such an inference:

[146] Levine, Chalmers and Jackson, and Kim clearly endorse a). It is less clear whether they endorse b) as well. But in any case, their model of reductive explanation involves not only a priori derivability, but substantial parts of actual a priori derivations as well.

In chapter 3 I argued that it is a priori that water, if it exists, is the substance which plays a certain theoretical role, including its floating in our oceans and lakes, being drinkable, sometimes falling from the sky in drop shape and sometimes in flake shape, freezing at 0° Celsius, etc. Now consider the vast amount of information one would need to determine that this theoretical role is played by H_2O: Information about the distribution of H_2O molecules on our planet and in its atmosphere, their behavior in various kinds of circumstances, their effects on our organisms, etc. It would be absurd to demand that a reductive explanation of the disposition of water to form drops has to invoke all of these facts.

One can of course insist that reductive explanation does require completely a priori entailments. One would then have to say that in the example I gave, one has really only explained the disposition of H_2O molecules to form drops. But I do not see an independent reason for this claim. It seems perfectly natural to say that one has given a reductive explanation of a macroscopic feature of water, even if the explanation is not a priori through and through. We can be very confident that water is indeed H_2O, so it does not do any harm to use this piece of empirical knowledge as a premise in the reductive explanation.

If one accepts this result, one is forced to reject constraint b) as well – if the entailment from explanans to explanandum does not have to be purely a priori, then one cannot require that it is shown that there is such an entailment, either. Moreover, I would suspect that even in cases where the explanans does entail the explanandum a priori, it will often be practically impossible or at least excessively difficult to draw all the relevant a priori inferences. In many such cases, it may do, for instance, to work with approximate models or to invoke a well-established empirical premise at some point or other. a) and b) are thus arguably too strong as constraints on reductive explanations. Nevertheless, the point I made above about the need for intelligible/transparent connections between explanans and explanandum remains valid. And in general, the connection between an L-fact and an H-fact will be intelligible if one can derive the H-fact from the L-fact without additional empirical premises. This suggests that one should

be parsimonious in invoking such empirical premises. In judging the adequacy of the explanation, one should also take into account at which part of the explanation these premises are invoked and which form they take. For example, it would seem inadequate if one had to postulate mediating 'bridging principles' at an important stage of the explanation which themselves remain unexplained.

Which kind of inquiry one should reserve the label 'reductive explanation' for is arguably not crucial. I already noted that if a phenomenon can be reductively explained (in a neutral sense of the term) by certain lower-level features, then either the properties of the higher-level phenomenon are a priori entailed by the lower-level features – in a case where only epistemically transparent expressions are used – or they are a priori entailed from a broader basis. Furthermore, in the latter case there is always an explanandum nearby which is in fact a priori entailed by the explanandum: In the example I gave above, for example, this would be the disposition of H_2O to form drops. It is evident that if one could show how the facts in the explanandum are a priori entailed by those in the explanans, this would provide an excellent explanation. Given this, there is no reason why one should not take this ideal case as a model of explanation and try to come as close to it as possible.

In the following, I will have a closer look at how the type of a priori reductive explanation advocated by Levine and others is supposed to proceed. A key question which I will try to answer is whether such reductive explanations require explicit analyses, and if so, to what extent.

One reason why people have been hesitant to concede that conceptual analysis plays an important role in reductive explanations is their general skepticism regarding the possibility of giving explicit analyses (cf. e.g. Block & Stalnaker 1999). In their joint paper, Chalmers and Jackson address this worry (cf. Chalmers & Jackson 2001). They point out that their claim is only that the facts in the explanandum are a priori entailed by those in the explanans. This is compatible with the absence of a definition of the explanandum, just like it is possible to derive facts about knowledge from a

description of a hypothetical case (for instance a Gettier scenario) in the absence of a definition of 'knowledge'. Chalmers and Jackson's case for the importance of conceptual analysis in reductive explanations is thus based on scrutability, rather than on explicit analyses. On the face of it, the ability mentioned by a thesis such as (CJ++) seems tailor-made for a reductive explanation: Take a higher-level fact H and a lower-level fact L. If it can be shown that H is a priori entailed by L, then this offers a perfect explanation for the occurrence of H in the sense outlined above: Given L, it becomes intelligible why H occurred, or even had to occur. Does this mean that reductive explanations can just rely on scrutability and do not need any kind of explicit analyses? As I will show in the following, this is far from clear.

The way I put things so far might convey the impression that the process of giving a reductive explanation is completely bottom-up. But this cannot be correct. First of all, it would be absurd to think that we start from complete low-level information and then reason our way upwards, deriving one higher-level truth after the other until we stumble upon our H. Furthermore, and somewhat less trivially, one does not typically aim at a reductive explanation of a singular fact, event or state. Rather, reductive explanations target wholesale phenomena. That is to say, firstly, that one deals with a type of fact (event, state) which usually has a huge number of instances. (Arguably, one may at least occasionally want to explain a range of merely possible instances as well.) Secondly, the target of the explanation will typically comprise a number of different configurations or states at the macro-level.

Consequently, in order to give a reductive explanation it will not suffice to derive one higher-level fact from a lower-level fact – if only because one has thereby at best explained one kind of high-level configuration. But needless to say, we cannot derive every single instance of the relevant higher-level fact(s) from lower-level facts, either. So if the whole process was indeed bottom-up and we had to rely on our ability to derive higher-level facts from lower-level facts alone, we would not be able to explain the whole phenomenon.

If one takes the scrutability thesis and our corresponding ability as the sole basis of reductive explanation, one thus seems at a loss to say how such an explanation is supposed to work. For this reason, one will at least have to make an additional assumption to the effect that we have some kind of prior grasp of the explanandum, beyond our ability to recognize its instances if confronted with lower-level information. It is thus no accident that when people outline how a reductive explanation which is based on a priori entailments proceeds, they usually start with a definition of the explanandum (cf. e.g. Levine 1993, 131). Let me briefly sketch how such a reductive explanation could proceed:

Suppose one aims to give a reductive explanation of heat. In the first step, one will provide an analysis of 'heat', in the sense of explicating its primary intension. Plausibly, this will mean identifying the causal role associated with the term: Heat is the phenomenon which makes ice melt and water boil, which brings about tissue damage, causes gases to expand, etc.

In the second step, one identifies the occupant of this causal role: Mean molecular kinetic energy (or in the cases I mentioned (relatively) high mean molecular kinetic energy) is what does all these things. The determination of the occupant of the theoretical role is enabled by the fact that once we suspect that the relevant effects might be related to the energy-level of molecules, we can reason from facts in the explanans to what we actually observe at the macro-level. Here is how this might look, in very crude form: When the mean molecular kinetic energy of H_2O molecules reaches a certain level, the ordering of the molecules in the crystalline structure (of ice) will collapse and the aggregate phase of the sample will change from solid to liquid. If one increases the energy level even further, one thereby increases the likelihood of the molecules to escape the liquid and thereby the liquid's vapor pressure, up to the point where the vapor pressure will equal the pressure of the surrounding environment and thus vapor bubbles will start to form in the liquid. When proteins in our skin interact with high-level particles, they will disintegrate, resulting in tissue-damage. And when the molecules in a gas gain

molecular kinetic energy, the mean distance between them will increase which means that the gas expands.[147]

We just saw that if one starts from a definition of the explanandum, one can at least tell an intelligible story of how the envisaged explanation could proceed. This does not mean that reductive explanations have to begin with an analysis of the phenomenon to be explained, however. It may usually be sufficient to give an approximate characterization of the explanandum, or even just to rely on our implicit grasp of it. And arguably, something less than a complete definition typically has to be sufficient. For, it is very likely that for most phenomena which have been reductively explained, no analysis which holds up to philosophical standards – such as, say, a definition which is not prone to counterexamples – has been presented.

Accordingly, I do not want to deny that it is possible to provide reductive explanations in the absence of perfect definitions. Yet, let me note once more that such an explanation depends on more than bottom-up scrutability; we also need the ability to overlook what the phenomenon in question comprises. In particular, we need to be able to assess not only where to look, but also when we have explained everything which needed to be explained. The latter point can be illustrated by drawing an analogy with what happened in the case of knowledge. Based on their understanding of the term, people thought for a long time that having a justified true belief is really all there is to knowing something. But they missed a relevant aspect of the phenomenon, even after considerable reflection about their subject matter. Something similar could happen in the case of reductive explanations. Unless one has a complete (and correct) definition of the explanandum, it might happen that one believes to have provided a reductive explanation even though there is an important dimension to it which is yet to be accounted for.

David Lewis' account of pain might serve as an example (cf. Lewis 1983): In Lewis' view, pain can be defined in functional terms.[148] Let us thus say

[147] If one adds to this a third step in which heat is identified with mean molecular kinetic energy, then the whole process follows quite closely the general model of reduction in the Canberra Plan, which will be discussed in the next section.

that pain is the phenomenon which causes avoidance behavior, which brings people to utter things like 'Ouch!' or 'That hurts', which makes them consult a doctor or take a pill, etc. If this definition is correct, then it is plausible that pain can be reductively explained in purely physical terms, since the relevant functional roles are plausibly played by broadly physical features. However, critics have objected that Lewis' account of pain leaves out a crucial aspect of pain, namely the distinctive phenomenal quality associated with it. They say that since there cannot be a definition of pain in functional terms, there is no way to give a reductive explanation of pain in physical terms (cf. e.g. Levine 1993; Chalmers 1996; Kim 2005). If these critics are correct, an analysis in the spirit of Lewis might still capture many or even most important aspects of pain and thus allow a reductive explanation of these aspects in physical terms. Nevertheless, the overall explanatory project will be doomed to failure.

The example suggests that even if in many cases, a "rough-and-ready analysis" of the explanandum (Chalmers 1996, 43) or even our implicit grasp of it turns out to be sufficient for giving a reductive explanation of a phenomenon, this method is not fool-proof: Unless one has a complete and correct definition at hand, it can happen that one misses relevant aspects which need to be explained.

7.1.3 The Canberra Plan

The Canberra Plan is intimately linked with a specific account of the objectives of metaphysics. Before I turn to a discussion of the Canberra Plan itself, let me thus outline this account which has been most prominently advocated by Lewis and Jackson.

In Jackson's words, 'serious metaphysics' does not occupy itself with enumerating everything there is in the sense of drawing up a huge list of

[148] With the addition that pain only needs to play this functional role usually in members of the relevant species, in order to account for occurrences of pain which do not play the functional role ('mad pain') and states which have this role but are not pain ('Martian pain').

things which includes black holes, electrons and nucleic acid, as well as spoons, potato chips and billiard cues. Rather, it seeks to compile a much more parsimonious account in which only the fundamental building blocks of the world are mentioned (cf. Jackson 1998a, ch. 1). On Jackson's own account, for instance, this will be a description in terms of a completed physics.

However, this idea immediately raises a problem. In the more parsimonious fundamental description of the world, spoons, potato chips or billiard cues will hardly appear – after all, the whole idea of doing 'serious metaphysics' was to provide a shorter list. But at the same time, the list should still be comprehensive. So does this mean that one has to say that there are no shoes, potato chips or billiard cues? This would obviously be quite an unhappy result. The solution to the problem is based on the idea that many kinds of facts which are not explicitly mentioned in the fundamental description of the world nevertheless appear implicitly in it. The facts which appear implicitly in the description are those which are metaphysically entailed by it, i.e. those which supervene on the distribution of fundamental properties.

At this point, another important task of metaphysics becomes apparent: It aims to show how the fundamental description of the world makes the non-fundamental facts true. Metaphysics thus deals with 'location problems' (this is again Jackson's expression), i.e. it tries to locate higher-level phenomena like spoons and potato chips or, more realistically, knowledge, meaning and actions in the basic physical reality. Given my discussions in the preceding part of this work, it should not come as a surprise that conceptual analysis plays an essential role in this kind of enterprise. The key idea here is that one aims to show that higher-order facts are metaphysically entailed by the fundamental facts by pointing out how they are a priori entailed by them. The theoretical underpinning for this kind of procedure is provided by (CJ++), and more generally by the thesis of metaphysical plenitude: Provided that the fundamental description of the world is couched in purely epistemically neutral vocabulary, every fact

which is epistemically entailed by this description is also a priori entailed by it and vice versa.[149]

Just like in the case of reductive explanations, the question arises how this objective is supposed to be achieved. As I already noted in the previous section, one cannot just start from a complete fundamental description and hope to derive specific higher-order truths from it. This may be the reason why the starting point of the Canberra Plan is at the top level: Its first step consists in an analysis of the relevant higher-level phenomenon. In what follows, I will set forth the specifics of the Canberra Plan in some more detail.

The first step of the Canberra Plan is based on Lewis' account of defining theoretical terms which is in turn based on the work of Frank Ramsey and Carnap (cf. Lewis 1970): Suppose you have a theory T about a specific domain. T comprises some theoretical terms $t_1, \ldots, t_n$. Of course, there are also many other kinds of terms which figure in T. Lewis calls these the O-terms (here, the 'O' stands for 'Other'). One may assume that the theoretical terms are newly introduced with the theory and derive their interpretation only from their relation to the O-terms as specified in T. Now the theoretical terms are replaced by free variables, thus yielding $T(x_1, \ldots, x_n)$. Lewis calls this the realization formula of T: Any n-tuple of entities which satisfies the formula realizes T.

The *Ramsey sentence* of T says that T has such a realization: $\exists x_1, \ldots, x_n\ T(x_1, \ldots, x_n)$. And its *Carnap sentence* says that if T is realized, then the theory's postulate is satisfied: $\exists x_1, \ldots, x_n\ T(x_1, \ldots, x_n) \rightarrow T(t_1, \ldots, t_n)$.

Importantly, the Carnap sentence is supposed to be analytic, i.e. it is analytic that if the realization formula of T is satisfied, then the theory is correct. According to Lewis, the same goes for the inverse of the Carnap sentence: $\sim\exists x_1, \ldots, x_n\ T(x_1, \ldots, x_n) \rightarrow \sim T(t_1, \ldots, t_n)$. This implies that the theory T, i.e. its postulate, is a priori (analytically) equivalent to its Ramsey sentence. Note also that none of the T-terms figures in the Ramsey

[149] As was already noted a couple of times, to be able to derive all truths one will additionally need indexical information and a clause which states that the description is complete.

sentence. The Ramsey sentence thus provides a way of dispensing with the T-terms; it is surely no stretch to say that it thus provides an analysis of the T-terms.

The fact that the theoretical terms are replaced by variables suggests that Lewis' model only applies to names. However, in his view predicates are eligible as well; they simply have to be treated as names of properties. For example, if 'green' is the relevant T-term, then 'x is green' is rephrased as something like 'x has greenness', where 'has' is an O-term. Furthermore, Lewis' theory places no constraints on what one can consider as a theoretical term in a given context. Thus in principle, any kind of term or set of terms can be subject to this kind of analysis. For this reason, Lewis' account of defining theoretical terms can serve as the basis of a comprehensive metaphysical project of reduction.

Here is how this project is supposed to proceed: One starts with an (alleged) higher-order phenomenon. In a first step, one constructs a Ramsey sentence, with those terms which are relevant for characterizing the phenomenon figuring as T-terms. One thereby gets an account of the relations of that phenomenon to other kinds of features, as described by the O-terms; call these relations the T-role. Then, one goes about determining what, if anything, in the fundamental description satisfies the T-role. If one succeeds, one has thereby located the higher-level phenomenon within the fundamental description of the world. Given this, one can also reason upward from the fundamental description to arbitrary truths about the phenomenon in question, in accordance with (CJ++).

Let me flesh out this highly abstract picture with the aid of an example. Suppose the higher-order phenomenon to be reduced is water. 'Water' can thus be considered as the sole theoretical term. The Ramsey sentence will roughly say something like this: There is an x such that x actually stands in an acquaintance relation to me, is clear and drinkable, sometimes falls from the sky in drop shape, expands when it freezes, …; call this the water role.[150] Now one tries to determine what plays this theoretical role: Since,

[150] Cf. 3.1.1 for more details.

as we all know, H_2O plays this theoretical role, one can conclude that water is H_2O. [151]

In the following, I will discuss a couple of issues concerning the commitments of the Canberra Plan, its viability and its potential merits for philosophy. I will first, i.e. in 7.1.3.1, say a few words to clarify how Ramsey sentences are built and how this relates to the problem of ensuring unique reference. Then, in 7.1.3.2, I will discuss two worries regarding the scope of the Canberra Plan and finally, in 7.1.3.3, two further worries concerning its practical feasibility.

7.1.3.1 Ramsey sentences, primary intensions and unique reference

Let me start by looking in a bit more detail at how the relevant Ramsey sentences are supposed to be constructed. In *Psychological and Theoretical Identifications*, Lewis expresses the view that one has to start with a collection of platitudes about the subject matter at hand (cf. Lewis 1972, 256).[152] Jackson occasionally comments in the same vein, for instance in the context of a reduction of moral properties (cf. e.g. Jackson 1998a, 130). The example of water I gave above seems to confirm this idea. The Ramsey sentence I sketched there mainly (but really only mainly) looks like a collection of platitudes about water. One should not assume, however, that such a procedure is generally adequate. To see this, recall that the Ramsey sentence is supposed to provide a definition of the term(s) in question. In effect, it should thus spell out its/their primary intension(s). But as I argued most extensively in chapter 3, one important lesson to be drawn from Kripke's arguments against descriptivism is that the properties associated

[151] H_2O is most likely not included in a fundamental description of the world, but it is at least sufficiently low-level to illustrate the general schema. The underlying problem that there can still be many levels between what is said in the Ramsey sentence and in the fundamental description will be addressed in 7.1.3.3.

[152] The subject matter which he was concerned with there was folk psychology. He later changed his mind, though, cf. Lewis 1994, 216.

with an expression do not always correspond to what readily comes to a subject's mind when she comes across a token of this expression.
From what I just said it can be concluded that the question of how one should go about constructing Ramsey sentences amounts to the question of how one can provide definitions. I will leave a discussion of this difficult issue to 7.2.

It has been argued that Ramsey-style analyses are liable to the so-called 'permutation problem' (cf. Smith 1994, 48ff.): Suppose that the network of relations between T-terms and O-terms which we manage to identify is not sufficiently tight. Then it can happen that the resulting Ramsey sentence is satisfied by many (tuples of) entities; it does not permit us to uniquely locate the target phenomenon in the fundamental description.
The permutation problem is closely related to Kripke's argument from Ignorance against descriptivism which was discussed in chapter 3. According to this argument, the reference of names cannot be determined by speaker associations because these associations are often too unspecific to yield a unique referent. The connection between the argument from Ignorance and the permutation problem is due to the fact that, as I just argued, the Ramsey sentence is supposed to reflect the primary intension(s) of the expression(s) in question, in other words the associated properties. The reply to Kripke's argument I gave in chapter 3 and my general defense of the idea that speaker associations determine reference with respect to every world considered as actual throughout much of this work thus apply to the permutation problem as well. Consequently, if the Ramsey sentence does not suffice for determining the reference of the T-term(s), this can only be due to the fact that it does not capture its/their primary intension(s) completely.
Nevertheless, it can of course happen that even a complete Ramsey sentence does not uniquely determine an entity or tuple of entities. We hardly have an a priori guarantee that, for example, our 'water theory' picks out exactly one substance. For this reason, the problem of what should be done when there is more than one realizer or tuple of realizers cannot be

completely avoided. When Lewis introduced his account of defining theoretical terms (in *How to Define Theoretical Terms*), he argued that it is part of a theory that its postulate is satisfied by a unique tuple of entities (cf. Lewis 1970, 432ff.). Accordingly, if there is more than one tuple of entities which satisfies the postulate, this means that the theory is false and thus the T-terms fail to refer. In later work, Lewis seemed less sure how to deal with such cases, wavering between reference failure and indeterminacy (cf. Lewis 1999, 347). Wolfgang Schwarz argues, however, that the latter choice is clearly the appropriate one (cf. Schwarz 2009, 221). While I tend to agree with Schwarz, I do not think that this choice should be made on the level of our meta-theory. Rather, it should be decided by what we associate with the term or terms in question. There is, for instance, nothing which prohibits using a term in a way such that it refers only if it refers uniquely. In such a case, the relevant condition will be part of the term's primary intension and thus a clause to that effect will appear in its Ramsey sentence, as in 'There is exactly one x such that …'.

7.1.3.2 The scope of the Canberra Plan

It is clear that one cannot 'ramsify away' all kinds of terms. One will always need some set of O-terms which define the T-terms.[153] Since the general project is to locate higher-level phenomena in a fundamental description of the world, it is natural to think that the ultimately remaining O-terms will be those which are required to describe the fundamental properties and their distribution.[154] However, Schwarz argues that there are terms which clearly seem unnecessary in a fundamental description, but which are nevertheless not eligible for a Ramsey-style analysis (cf. Schwarz 2009, 225ff.): Recall that in Lewis' theory, predicates are also replaced by variables in a Ramsey sentence – after having been rephrased

[153] In Lewis 1984, he defends a view which can be interpreted as leaving only one such O-term, namely 'naturalness'.

[154] Chalmers forthcoming discusses in detail what kind of vocabulary will have to be considered as primitive.

as names for properties. But as Schwarz notes, there does not have to be an entity, such as a property, corresponding to a given predicate. An example he gives is causation: In Lewis' own theory of causation, absences can be causes. Accordingly, there is no causation relation, since absences are hardly eligible as relata.[155] If one constructs a Ramsey sentence for 'causes', one will thus find that there is no entity which satisfies it. However, it would clearly be wrong to conclude from this that Lewis should be an eliminativist about causation. For in his theory, facts about causation are made true by lower-level facts. Consequently, Lewis' method of defining theoretical terms cannot be applied to every kind of expression.
It is not obvious that this result is detrimental to the overall reduction project, however. One could still apply a very similar procedure in order to locate causation facts in a fundamental description: One starts with a conceptual analysis, for instance by systematizing the way the term 'causes' is applied to various hypothetical scenarios. Ideally, one will end up with an explicit analysis, yielding a way to express facts about causation without mentioning the term 'causes'. Then one can go about determining what in our fundamental description makes these facts true.

The next worry concerning the scope of the Canberra Plan is also due to Schwarz (cf. Schwarz 2009, 222–224). To elucidate what he aims at, let me remind you of a feature of many expressions which I discussed in chapter 5: There I argued that many of our terms are functional terms. Their purpose is not to pick out a specific individual, but rather to characterize a function, such that anything which has that function falls under the term. Now, Schwarz claims that for such expressions (he talks more generally about role terms), the second and the third step of the Canberra Plan, i.e. those steps in which the higher-order phenomenon is located in the fundamental description, are not applicable or at least irrelevant. Take his example of the property of being a clock: Suppose we have a Ramsey sentence for that property. According to the Canberra Plan, we are now supposed to locate it in our fundamental picture of the world. But it is

[155] This is remarked by Lewis himself, for instance in Lewis 2004.

highly implausible that the various things which count as clocks have any kind of, say, microphysical commonality. We could describe the microphysical state of every existing clock and say that the property we are after is the disjunction of all of these microphysical states. However, that will not work. For what if someone would build another kind of clock? There are arguably even possible worlds containing clocks which are made of non-physical stuff. Schwarz concludes that the property we are after is just a (higher-level) functional property. Once we have the Ramsey sentence which spells out this function, we already have everything we could be looking for. It is thus neither possible nor necessary to follow up with a microphysical reduction.

While I agree with much of Schwarz's reasoning here, I do not agree with the conclusions he draws. First of all, it is true that we will not be able to give a microphysical reduction of the property of being a clock. But why should this be necessary, within Lewis' own account, for a successful reduction? For note that other possible worlds are part of his fundamental ontology as well. Lewis could thus say that the property of being a clock can be reduced to the set of microphysical or otherwise fundamental configurations constituting clocks across all possible worlds. This is in line with Lewis' general account of properties which he construes as sets of possible worlds.

Even more importantly, Schwarz' conclusion that the only step of interest in such cases is the first one of developing the Ramsey sentence is clearly exaggerated. Take a term which is mentioned by him as well, 'freedom'. Arguably, the term satisfies Schwarz' criteria for role terms: Free actions or choices (if there are any) will hardly share any microphysical commonality. And building the disjunction of microphysical realizations will not help either, since there could be different microphysical, or even non-physical, configurations which also constitute free actions or choices. Nevertheless, we might be interested in determining whether freedom does have a realizer on the level of fundamental properties, in order to find out whether there are free actions or choices. Alternatively, we might want to know whether it has, or could have, a physical realizer. Similar considerations

apply to many other philosophically relevant terms which are role terms in Schwarz' sense as well, for example 'belief', 'pain', 'good', 'true', etc. Sometimes we are interested in knowing whether a higher-level role is realized by the fundamental properties, in order to know that the phenomenon in question exists. Alternatively, we may want to know whether it could be realized by what we take to be the fundamental properties. At other times, we are interested in whether the phenomenon could be realized by specific kinds of fundamental properties, for example by physical properties – either to locate the higher-level phenomenon, or to eliminate it, or to expand the repertoire of fundamental properties we posit. Accordingly, even in cases where we are dealing with role terms, the task of relating the corresponding properties with fundamental properties (in accordance with the Canberra Plan) will often be of philosophical interest.

7.1.3.3 The practicability of the Canberra Plan

The idea of locating ordinary macroscopic phenomena in a description of the world which only spells out the distribution of its fundamental properties should strike one as extremely ambitious. It is far from obvious how such a project should proceed and whether it can be successfully carried out. In the following, I will therefore identify the requirements of the Canberra Plan and evaluate its practical feasibility.

To begin with, the Canberra Plan relies on explicit analyses in the form of Ramsey sentences. One may wonder whether these analyses have to be complete and completely accurate, like in the case of reductive analyses discussed above. Plausibly, one will reach a verdict here which is similar to the one I reached there: In some cases, reasonably good though imperfect analyses will be sufficient. But in many cases, details clearly matter. To take just one example, consider Lewis' reduction of dispositions to counterfactual conditionals (cf. Lewis 1997). There is widespread agreement that there is a strong link between dispositions and counterfactuals. But many people think that there are counterexamples to each of the analyses Lewis has proposed; and this in turn has led many to

believe that dispositions cannot be reduced to counterfactual conditionals at all. It is thus plausible that successful reductions in line with the Canberra Plan will often require explicit, complete and accurate analyses.

If one has constructed the relevant Ramsey sentence, the next task is to use this analysis to locate the target phenomenon in a fundamental picture of the world. It is not obvious how this can be done. For a start, an opponent of the Canberra Plan might claim that since we do not have a description of the distribution of fundamental properties at hand, we can hardly locate anything in it. In principle, that objection is surely correct. But at least in most cases we do not need such a complete description in order to solve a location problem. What we do usually need, though, is a rough grasp of what the fundamental properties will be like. Here, we have to rely on the assumption that our best scientific and/or philosophical theories are at least on the right track. If this assumption turns out to be wrong, then many of our putative reductions will be inaccurate. But this is a problem which cannot be avoided when one is concerned with metaphysics, so it is hardly a specific problem for the Canberra Plan.

Next, assume that we do have both a Ramsey-style analysis of our target phenomenon and a sufficient understanding of the fundamental properties. If we are lucky, then the O-terms will be part of the vocabulary used in the fundamental description of the world. In this case, the reduction is already complete. But since this is not to be expected, we are still confronted with a location problem: The problem of relating facts expressed by O-terms to those expressed by the fundamental vocabulary. In many cases, that problem will still not be trivial. To give an example: Lewis analyzes meanings in terms of linguistic conventions. Prima facie, it seems no more obvious how linguistic conventions relate to the fundamental properties than how meanings do. It is therefore reasonable to introduce a few more intermediate steps. Here is a crude sketch of what such a reduction could look like within Lewis' framework: Conventions are (very roughly) reduced to dispositions of speakers[156] which are then analyzed via

[156] A bit more precisely, they are reduced to dispositions and intentional states of speakers. Since on Lewis' functionalist account, intentional states are themselves, by

counterfactual conditionals. Counterfactual conditionals are analyzed in terms of relations between possible worlds; these (the relations, not the worlds!) are supposed to be reducible to qualitative similarities in the distribution of fundamental properties of these worlds and laws of nature. Laws of nature are in turn reduced to regularities among the fundamental (in particular, microphysical) properties. At each of these steps, one will require an analysis; each time, some of the former O-terms become T-terms and are thus defined away in the new analysis. And once again, when the reduction is complete, one can reason bottom-up, from the distribution of fundamental properties to the laws of nature to … to meaning facts.

The Canberra Plan, as it is paradigmatically carried out by Lewis, is certainly a fascinating project. The idea to reduce everything there is to a limited basis of fundamental facts, moreover in such a way that it becomes a priori intelligible how the higher-order facts are grounded in the fundamental ones, seems highly appealing. At the same time, as I already mentioned above, the project is extremely ambitious and therefore also highly vulnerable. In the example I just gave concerning a possible reduction of meaning facts to fundamental properties, each of the steps involved is as intricate as it is controversial.
Nonetheless, analyses in the spirit of the Canberra Plan can be useful even for those who do not have such far-reaching reductive ambitions. If, for instance, one managed to establish just one of the steps I sketched of a Lewisian reduction of meaning, this would already be a major success. Generally speaking, one does not always have to reduce the target phenomenon to fundamental properties. One may just aim at showing that facts of a domain X are reducible to facts of domain Y, or just to facts outside of domain X (for instance in the case of moral or generally normative facts). Or, maybe even more modestly, one just tries to show that all the facts about a specific domain can be expressed without using the

---

and large, reducible to dispositions, the statement that conventions are reduced to speaker dispositions is nevertheless approximately adequate.

vocabulary which is considered to be characteristic for that domain, in order to dispense with these expressions.
The Canberra Plan thus offers an attractive model of how conceptual analysis could be done in philosophical practice. However, just like a number of other varieties of conceptual analysis which I have sketched in this chapter, it has to rely at least to some extent on the possibility of providing explicit analyses. In the following, I will turn to a discussion of the problems and prospects of this project.

7.2 The trouble with definitions

In the previous part of this chapter, I pointed out that conceptual analysis can be valuable even if it does not yield explicit analyses. However, in my discussion it also transpired that many ways in which conceptual analysis can be applied to gain philosophical insights do involve explicit analyses at one stage or other. For example, they are crucially involved in an attempt to reveal the essences of philosophically relevant categories, in the Canberra Plan of locating higher-level phenomena among the fundamental properties, and, as I have argued, they are at least sometimes also required in reductive explanations.
This dependence on explicit analyses will presumably seem worrisome to many. For it is often held that the history of attempted definitions of philosophical terms is a history of failure. It should be well worth considering whether the prospects of the project of providing definitions are really that bleak, given its implications for the scope and the potential value of conceptual analysis. The rest of this chapter will therefore be dedicated to that question.

7.2.1 Adequacy conditions for definitions

In order to evaluate the prospects of finding definitions, one will first have to determine the conditions which an adequate definition has to meet. It seems natural to demand that the definiens has the same application

conditions as the definiendum, where application conditions are here to be understood as primary application conditions. That is to say, the primary intensions need to be identical. However, one could argue that this requirement is a bit too strict. One might for example hold that it is even a virtue of a definition if it reduces vagueness. In such a case the definiens has a definite extension in at least one scenario which is borderline according to the term to be defined. It may thus make sense to loosen the criterion just proposed slightly and require instead that the definiens should have the same primary application conditions with respect to cases where the definiendum has a definite extension.

Even that much is not uncontroversial. There may be cases where the expression to be analyzed needs to be slightly 'adjusted' for some reason or other. For example, the term in question might be inherently confused or just contingently empty. Alternatively, it may turn out that as it is, the term is of no particular philosophical use for other kinds of reasons. In each of these cases, it could be useful to offer an analysis of the original term with at least slightly different application conditions. The analysis might thereby yield an expression which is not confused (and thus not necessarily empty), not empty or just of greater philosophical interest for another reason. Such an approach seems well in line with Lewis' views and also with Jackson's. Jackson explicitly says that often, conceptual analysis will involve a slight departure from the folk concept at hand (cf. Jackson 1998a, 44–46). As an example, he discusses what it takes for an action to be free: Jackson believes that (our folk concept of) a free action is strictly speaking not compatible with determinism:

> What compatibilist arguments show, or so it seems to me, is not that free action as understood by the folk is compatible with determinism, but that free action on a conception near enough to the folk's to be regarded as a natural extension of it, and which does the theoretical job we folk give the concept of free action in adjudicating questions of moral responsibility and punishment, and in governing our attitudes to the questions of those around us, is compatible with determinism. (Jackson 1998a, 44f.)

Accordingly, if the compatibilist offers an analysis of 'free action', the analysans will have application conditions which differ at least slightly from those of the analysandum.

In my view, such an adjustment of the expression to be analyzed is not unproblematic. A skeptic concerning free action will say that if Jackson's assessment of the situation is accurate, then she wins the debate. For in this case, it may well be true that there are actions that fall under a notion which is in some sense similar to our notion of free action, but nevertheless, there are no free actions.[157] As another example, consider the question of whether there are moral properties. John Mackie argues that there are no such things as categorical imperatives, i.e. there are no judgments which in themselves give us a reason to act / motivate us to act. His reason for holding this view is that he thinks there are no properties which are intrinsically motivational. Furthermore, Mackie claims that our moral discourse presupposes that there are such 'objectively prescriptive' properties. For this reason, he considers this discourse to be fundamentally flawed (cf. Mackie 1977): It presupposes the existence of moral properties which do not, or even cannot, exist.

Now suppose one goes about defining moral terms, proceeding as follows: First one slightly adjusts our folk terms such that it is no longer a necessary condition that knowledge of their extension in itself motivates a subject to act. Then one develops a definition in terms of, say, naturalistically acceptable properties. It should be obvious that one has not thereby shown that there are naturalistically acceptable moral properties. Quite the opposite: If our moral terms are such that their applicability requires the existence of intrinsically motivating properties and if moreover there are no such properties, then, rather trivially, our moral terms are empty, regardless

[157] Since our world is most likely not deterministic, this does not follow immediately. However, it is hard to see how the falsity of determinism is supposed to help the realist about free actions, so the skeptic's conclusion may nevertheless be justified. In this case, one could say that if Jackson's assessment is accurate, then there are necessarily no free actions.

of whether there are non-empty terms in the vicinity which might even be philosophically useful.

This does not mean, however, that I consider Jackson's view to be fundamentally flawed. I agree with him that the concepts we find in a natural language may not always pick out philosophically interesting categories or, more generally, that they may not always be perfectly suitable for philosophical purposes. In many such cases, it can be useful to replace a concept with one which has similar application conditions: The successor may well help to clarify important philosophical questions, for instance. The examples I just gave show, however, that one has to be careful in drawing philosophical conclusions from such analyses. At bottom, this is because philosophical questions are phrased in terms of a natural language; these folk terms thus define our subject matter (cf. 4.1).

It is natural to liken the kind of 'adjusting analysis' proposed by Jackson to Carnapian explications (cf. Carnap 1947/1956). Jackson's remarks do suggest that there are contexts in which explications are sufficient for our purposes. Given what I have just argued, however, it is important to distinguish these from actual definitions. Let me thus extract the following minimal condition of adequacy from the previous considerations: In a successful definition, the application conditions of the definiens have to be similar enough to those of the definiendum to not beg any philosophically relevant questions which might come up.

Let me add the following, rather trivial condition of adequacy for definitions to the one just identified: They must not be of the form 'a = a', i.e. the definiens must use different expressions than the definiendum. Ideally, the terms used in the definiens or the features expressed by them will in some sense be more basic, but this should not be considered necessary. Often, definitions are also expected to be short, as in 'bachelor $=_{df}$ unmarried male adult human' or 'human $=_{df}$ rational animal'. However, I see no reason to take this as a necessary condition, either. One could require that definitions are of manageable length. However, depending on the context (and maybe a subject's cognitive capacities), analyses which

are much more extensive than anything philosophers have so far come up with can still be manageable.

I conclude that the task we are confronted with in constructing a definition of a given term amounts to that of finding a string of other words of manageable length yielding an expression with at least close to identical primary application conditions.

7.2.2 Objections to the eligibility of definitions

7.2.2.1 Objections from the relation between definiendum and everything else

In what follows, I will identify a number of necessary conditions for the feasibility of the task I just outlined. All of these conditions concern the relation of the definiendum with those expressions which could be used in the definiens, or alternatively with the facts expressed by them. I thereby provide skeptics of the project of giving definitions with targets for potential objections, since they could deny that these conditions are met by philosophically relevant expressions. My discussion will show, however, that there are reasons to believe that the conditions in question obtain in the case of most of the expressions with which we are concerned in philosophical practice.

Let me call the expression to be defined E and the properties denoted by it the E-properties. Now suppose that the E-properties do not supervene on any other properties, i.e. they are fundamental. In such a case, E can only be defined if it is epistemically opaque. Otherwise, it has to be considered as primitive. We have thus identified one necessary condition for the definability of E: E must not transparently denote properties which are metaphysically fundamental.
Next, assume that E-properties supervene on other kinds of properties for which, however, we do not have concepts. E-properties are thus not

metaphysically fundamental, but one could say that E is intentionally fundamental. With respect to the issue at hand, the implication is the same: Such a term would be primitive and thus unanalyzable. The second necessary condition for the definability of an expression is therefore this: The definiendum must supervene on properties to which we have intentional access, i.e. it must not be intentionally fundamental.

Now let me outline a third and final way in which E could be primitive: It could happen that we have intentional access to properties on which E-properties supervene, yet lack the linguistic means to express these properties. E-properties could, for example, be scrutable only from (hypothetical) evidence which cannot be expressed linguistically.

E could thus be undefinable because it is linguistically, or intentionally, or because it is metaphysically fundamental. It is to be expected that there are expressions of philosophical relevance which fall into one of these categories. But I do not believe that there is a problem for the general project of providing definitions to be found here, because the vast majority of these expressions are not primitive in any of these ways. For a start, it is safe to assume that few philosophical categories are metaphysically fundamental. It is furthermore plausible that in most cases, we are familiar with the kinds of properties on which they supervene, at least in the sense of having intentional access to them. We are thus only left with expressions which are linguistically fundamental despite being intentionally (and metaphysically) non-fundamental.

It is possible that there are such expressions. For example, we might be able to apply some terms only in response to sensory evidence which we are unable to express linguistically. But firstly, even in such cases it is not clear why we should assume that we are unable to express the features in question linguistically even in principle. Secondly, it seems far-fetched to think that many philosophical terms are such that their application is triggered (only) by non-linguistic evidence. Most philosophically relevant properties are not directly perceivable, let alone accessible only via perception.[158] Finally, the ubiquity of thought experiments in philosophical

[158] Exceptions are arguably provided by expressions related to conscious experience.

practice indicates that a great number of philosophical terms are not linguistically fundamental. For, in evaluating hypothetical scenarios, we apply the terms in question in response to (hypothetical) evidence which is phrased linguistically.

I just argued that the majority of philosophically relevant expressions are not linguistically fundamental. If an expression is thus non-fundamental, we can describe hypothetical scenarios to which it applies, without using the expression in the descriptions. Theoretically, one could use these descriptions to construct a Carnap sentence for our target expression, call it T, which may roughly look as follows:
$\exists x ((D_1 \rightarrow F(x)) \,\&\, (D_2 \rightarrow G(x)) \,\&\, \ldots)) \rightarrow T(x)$
The idea here is to determine for every scenario to which entities the expression in question applies and then to form a disjunction of these applications across the scenarios. However, in order to spell out the content of the target expression in such a way, one would have to invoke infinitely many hypothetical scenarios, resulting in a Carnap sentence of infinite length. Since this obviously exceeds the limits of manageability, an expression's being linguistically non-fundamental does not guarantee that it is definable: We could still be unable to spell out its content in a finite way. Let me give a positive account of how this could be the case: Since T is linguistically non-fundamental, we have a set of other terms (the O-terms) at our disposal by means of which we can express facts (the O-facts) on which the application of T supervenes a priori. That is to say, each configuration of O-facts determines a priori whether T applies or not, the application of T with respect to the various O-facts reflecting T's primary intension. Now assume that T is assigned to the configurations of O-facts completely at random. There is thus no pattern governing the applicability of T relative to the O-facts.[159] In such a case, there is no shorter way of expressing the application conditions of T by means of O-terms then in the form of a Ramsey sentence of the kind I just outlined. And thus, the existence of a pattern in the relation between the applicability of a term T

[159] Cf. Jackson 2005 for more on such patterns.

and the facts expressed by other terms is a necessary condition for T's definability.

It is hard to see, however, how this necessary condition could be turned into an objection against the project of providing definitions. It seems hopeless to argue that any terms which we would like to define exhibit such patternlessness. As Jackson notes, we could not possibly have learned to use such a term (cf. Jackson 2005, 128). For in such a case, knowledge of the term's extension with respect to an arbitrarily large number of scenarios would still be insufficient to tell whether it can be applied to any other scenario. Thus, if there was no relevant pattern, we would not be able to apply the term in question to a situation we are not already familiar with. It follows that there has to be a pattern in the application conditions of the terms we use, more specifically a pattern with respect to those properties in response to which we apply them. Let me add to this that since it is usually not overwhelmingly difficult to acquire and apply a term, this pattern will not be overwhelmingly complicated, either.

It is still a theoretical possibility that we are unable to express the pattern in a finite way. But while I see no way to prove that it is possible to do this, a denial of this claim seems excessively pessimistic with respect to the expressive powers of language: If an expression's application supervenes non-randomly on properties which we can express, then why should the pattern in the application conditions not be expressible as well?

In this section, I identified a number of necessary conditions for the definability of an expression. First of all, the definiendum must not be metaphysically, intentionally, or linguistically fundamental. I argued that the majority of philosophically relevant expressions do not fall into either of these categories. Then it turned out that there is another necessary condition on an expression's definability: There must be a pattern in the connection between the definiendum and other terms – which, however, cannot be seriously denied. I conclude that most philosophically relevant expressions meet the identified necessary conditions for their definability, which shows that considerations about the relation between the

definiendum and everything else provide no principled obstacles to the project of defining philosophical terms.

7.2.2.2 Objections from the format of concepts

It has been argued that the eligibility of definitions depends on the structure or the format of the concept in question. For this reason, (relatively) recent theories of concepts such as the prototype theory are often taken to speak against the prospects of conceptual analysis (cf. for example Kornblith 2007a, 41f.). In my view, our concepts come in a variety of formats. One therefore should not expect to find an answer to the question as to which of the theories in question is the correct one. Be that as it may, it still makes sense to discuss possible implications of the structure of a concept on the definability of a term which expresses it.

According to the classical theory of concepts, our concepts have a definitional structure in the sense of being composed of simpler concepts which provide necessary conditions for the application of the concept thus composed. It would obviously be convenient for conceptual analysis if the classical theory was correct. Laurence and Margolis even claim that conceptual analyses require that concepts have definitional structure (cf. Laurence & Margolis 2005/2011; cf. also Stich 1992; Ramsey 1992). And they go on to argue that since there are good reasons for rejecting the classical theory, the prospects of conceptual analysis are bleak. But this is not a compelling argument; in particular, its first premise seems clearly false: While I agree that only comparably few of our terms have definitional structure in any substantial sense, this is only a problem if one expects analyses to be very short or easy to come by. Definitions in the sense I outlined above only aim at capturing a term's application conditions, not its putative compositional structure.

The prototype theory provides a very different account of concepts (cf. e.g. Rosch & Mervis 1975). Its basic idea is that the application of a concept is dependent on a putative referent's similarity with a prototype. Let me spell

this out in a bit more detail: Suppose that a concept is associated with a set of features. That thing which displays these features most distinctively is the prototype of the category in question. Now, the concept applies to anything which has a sufficient number of the associated features, or alternatively to anything whose (weighted) sum of the degrees to which the features apply exceeds a certain value. It is often considered as a major advantage of the theory that it can account for so-called typicality effects: For example, chairs are considered by subjects as more typical pieces of furniture than refrigerators; the former are also more quickly and more accurately categorized as such than the latter. According to prototype theory, this is because chairs are closer to being a prototypical piece of furniture, i.e. they have more of the features which are relevant for membership in the category.

Typicality effects seem to be ubiquitous. Sharon Armstrong et al. have found that they can even be observed with respect to categories like 'even number' (cf. Armstrong, Gleitman & Gleitman 1983): For example, the number 8 is categorized as an even number more quickly than the number 34 and is also considered as a better example of an even number. However, while such effects may teach us something about the psychological mechanisms involved in categorizations, one should be careful to draw conclusions about the contents of the relevant concepts. The sole criterion for a number's being even is whether it is divisible by 2. It would be strange to hold that membership in the category depends on, say, similarity to the number 2.[160]

It has been denied that prototypical concepts can be defined because the relevant prototypical features do not yield necessary and sufficient conditions, for example by Stephen Stich (cf. Stich 1992, 249). In one sense, this is correct: For example, it may be that no single feature is necessary for the applicability of the concept. But in another sense, the prototypical features clearly do provide necessary and sufficient conditions. In order to avoid potential terminological confusions about what is meant

[160] Armstrong et al. also take such examples to show that there is no strong connection between a concept's exhibiting typicality effect and its having prototypical structure.

by necessary and sufficient conditions, let me try to phrase my response to the objection in more neutral terms: Prototype theorists do not deny that concepts have application conditions. I.e., they concede that there are conditions which have to be met by, say, an object or a situation to fall under the concept. Prototype theory even has something to say about the structure of these application conditions, namely that they are determined by a set of features which have to be present to a sufficient degree in a putative referent.[161]

Now, a definition simply aims at capturing these application conditions in an explicit linguistic format. I concede that doing so would often be an extremely tedious task. But I fail to see a principled reason why it should be impossible. One 'only' needs to identify the relevant features and their relative importance for the concept's applicability, which can be done in familiar ways – for example, by considering our judgments about hypothetical cases.[162] Let me add that it is plausible that when we are dealing with a prototype concept, we will usually be satisfied with an analysis which does less than specifying the concept's application conditions. Often, it will be sufficient to identify the features, or even only the kinds of features, which are relevant for the relevant term's applicability, without determining their relative importance in full detail.

Similar things could be said about other theories of concepts: As long as they grant that concepts do have application conditions, which is hard to deny, then there is no apparent reason why these application conditions should not be captured by using strings of words which do not involve the target terms themselves. At best (or worst), some of the formats in which concepts appear provide practical obstacles for specifying their application conditions.

---

[161] Cf. Jackson 1998a, 61 for a similar kind of response.

[162] Since cluster terms like 'water' can at least arguably be understood as prototypical, my discussion in chapter 3 bears some relevance to the question how this can be done.

7.2.3 The absence of successful definitions and some reasons for optimism

I just discussed a number of objections to the eligibility of definitions. It emerged that there are no principled reasons for thinking that definitions cannot be had. However, we found that there are a number of practical obstacles to the task of providing them. In the following, I will determine a couple of other factors which complicate the search for definitions and which may therefore help to explain why adequate definitions are so hard to come by. Against this backdrop, I will discuss to what extent the striking absence of successful definitions gives us reason to be skeptical with respect to the prospects of finding definitions and I will try to provide some reasons for optimism.

One merely practical reason why successful definitions are so rare is that philosophers usually aim at short definitions and rarely offer analyses which exceed a certain length. It is of course reasonable to prefer short definitions if they can be found – simplicity should be a virtue for definitions, too. But firstly, as I argued above one should not require them to be short. And secondly, there is no particular reason to expect that very short definitions are to be had –[163] this assumption seems to derive its popularity mainly from an adherence to the classical theory of concepts.

Another factor which might hinder the project of constructing definitions is that speakers may associate slightly different properties with an expression. In fact, for many concepts it would be rather surprising if it turned out that all competent speakers associate exactly the same primary intension with them (cf. also Jackson 2005, 135). Such variation in primary intensions can result in disagreement in the evaluation of one or more relevant hypothetical scenarios which can in turn lead to disagreement about the adequacy of a proposed definition. It can thus happen that a definition perfectly reflects the term's application conditions in a subject's idiolect and yet fails to be accepted by others. So what should be done in such a case? If it can be established that the disagreement about the adequacy of

[163] Jackson elaborates a bit more on this point in Jackson 2005, 132.

the definition is just verbal in this sense, one can either accept the definition, and maybe go on constructing definitions which reflect other people's concepts as well, or – which seems more reasonable – try to identify a core which is common to the concepts of all or at least most competent speakers. This latter proposal would inter alia be in line with Jackson's idea that conceptual analysis aims at analyses of our folk concepts. But either way, the fact that there can be mild intersubjective variation in the properties associated with a term does not amount to a principled obstacle to the task of providing definitions, either.

The main reason why many philosophers are skeptical with regard to the chances of providing definitions is that there are hardly any examples of adequate definitions which have been yielded so far. However, the preceding considerations have not only shown that there are no principled reasons to think that explicit analyses cannot be given, but also that there are a number of purely practical hindrances to that task. This result alleviates the objection from the absence of definitions. But needless to say, the only thing which will convince the skeptic is at least one clear-cut example of a waterproof analysis of a philosophically relevant term. Unfortunately, I cannot offer such an analysis here (although I will bring in an analysis which I think holds some promise). In the remaining part of this chapter I will, however, try to provide some support to the idea that the project of searching for definitions is a progressing one, rather than a degenerate one.

The objection from the absence of definitions should plausibly be construed as a kind of pessimistic induction: Since all the previous attempts to define key philosophical terms have failed, it is very likely that future attempts will fail as well. This thought is present in the comment of Williamson I quoted near the beginning of this chapter and also in the following remark of Chalmers and Jackson on the analysis of knowledge:

> On the absence of explicit analyses of knowledge: We take it that this is a reasonable conclusion to draw from four decades of failed attempts to produce explicit analyses. Certainly no explicit analysis has met with widespread

> approval, and proposed analyses are always confronted quickly by plausible counterexamples. (Chalmers & Jackson 2001, 321)

I am not sure about their claim that proposed definitions are *always* confronted by *plausible* counterexamples. But admittedly, counterexamples to such proposals are often found quickly. On the other hand, this fact also gives some reasons for optimism. For, it suggests that we are rather good at identifying weak spots in a proposed analysis, which is plausibly an important condition for making progress. And in many cases, such counterexamples do not just reveal yet another failure. Often, there is a lot to be learned from them. The failure of the standard analysis of knowledge presents a particularly vivid example for this: For it was quickly found that in all the counterexamples to the standard analysis, the subject's belief falls short of being knowledge because it owes its truth to luck in some way or other. And while a definition of the sort 'S knows that p $=_{df}$ S has the justified true belief that p the truth of which was not arrived at by luck' is surely unsatisfactory because it still needs to be specified what 'luck' exactly means, the insight does reveal that there is an important pattern which underlies the counterexamples. In my view, a lot of progress has been made in the attempt to capture this pattern in recent years.

One approach to explicating this pattern involves the notion of safety: A belief, for example, is safe if it could not have easily been false, that is if it is not only in fact true but also in all close worlds. Now a comparably recent idea to take this notion as the basis of an analysis of knowledge goes roughly as follows:

(MS) S knows that p $=_{df}$ S has the true belief that p which was produced by a safe process or method.[164]

Even though this analysis does get a significant number of problem cases right, it is, needless to say, far from being generally accepted. It is also true that a number of putative counterexamples to it have been proposed (cf. Comesaña 2005; Kemp 2009). But even so, there is still plenty of room for

[164] Here, (MS) stands for Method Safety. A proposal which can be construed along these lines was made by Mark Sainsbury (cf. Sainsbury 1997).

improving a safety-based account. Comesaña's and Kemp's cases are such that a method which (arguably) produces a knowledgeable belief is not safe because it could have easily been used in circumstances in which it would have been unreliable. At the same time, that method is highly reliable in the circumstances present during its actual application. One might thus propose to modify the analysis at issue slightly, yielding:

(MS*) S knows that p $=_{df}$ S has the true belief that p which was produced by a process or method which is generally reliable and safe in the circumstances in which it produced the belief.[165]

This formulation is still not perfect, because it seems to misclassify the subject's true belief that she sees a barn in Fake Barn County (cf. Goldman 1976, 772f.) as knowledge: For it would seem that it is a part of the circumstances in which the subject forms her belief that the light conditions are good, she is not distracted, and that *there is a real barn standing in front of her*. Given this, the belief conforms to (MS*).[166] Plausibly, the subject's belief in the fake barn scenario is not considered as knowledge because the recognition of barns via visual perception, while being generally reliable, is unreliable in the subject's current environment. So the above proposal can be repaired by adding a clause concerning the local reliability of the process or method involved, which is anyway an independently plausible condition for knowledge:[167]

(MS)** S knows that p $=_{df}$ S has the true belief that p which was produced by a process or method which is generally and locally reliable and safe in the circumstances in which it produced the belief.

This proposal handles a number of the most persistent problem cases for an analysis of knowledge, such as the original Gettier scenarios, various lottery cases, and – as we just saw – the fake barn scenario. Despite this, I

[165] Maybe surprisingly, the resulting view closely resembles Sosa's virtue-epistemologist account which he develops in Sosa 2007.

[166] Sosa is happy to accept this consequence of his account (cf. the previous footnote) and holds that the subject in the scenario does know that there is a barn in front of her.

[167] Here, local reliability should be understood not as a spatial, but as a modal notion.

am not claiming that I have thus presented the ultimate analysis of knowledge. It will probably turn out that this proposal still needs to be refined or that other accounts which do not rely on the notion of safety are more suitable. I mainly wanted to illustrate that there are approaches which integrate many lessons from counterexamples to previous analyses, which have already yielded analyses which are (at least in an intuitive sense) more adequate than the standard analysis of knowledge and which are eligible for further development. I therefore think there is hope that the program of constructing definitions is progressing.

This chapter was dedicated to a wide range of goals which people pursue when they practice conceptual analysis and to an assessment of the preconditions and the prospects for achieving these goals.
In the first part of the chapter, I discussed various ways in which the method and the aims of conceptual analysis have been understood: As a tool for revealing the essences of philosophically relevant categories, as an essential ingredient in reductive explanations, as a way of locating higher-level phenomena in a fundamental picture of the world, etc. It transpired that conceptual analysis can be a useful tool even when it fails to deliver explicit analyses, or does not even aim at delivering them. But a number of the ways of understanding conceptual analysis which I discussed do, or at least may, involve definitions at some stage or other.
The second part of the chapter was therefore meant to provide an assessment of the prospects of the project of constructing definitions of philosophically relevant terms. After determining adequacy conditions for definitions, I discussed potential obstacles to their achievability. Although there are a number of considerations which suggest that it will be difficult to develop definitions which are not susceptible to counterexamples, I found no principled reason for thinking that it is impossible. It is undeniable that the project of providing analyses has few spectacular successes to present as yet. Nevertheless, there is evidence that it is progressing.

8 Concluding remarks

In the previous chapters, I tried to show how two-dimensionalism can serve as a basis for the method of conceptual analysis. I also examined various other issues on which the practical value of conceptual analysis depends which go beyond the two-dimensionalist framework, sketched the aims and ambitions of conceptual analysis and discussed how they might be achieved. In what follows, I will give a brief recap of the results most relevant for the viability of conceptual analysis, connect a couple of loose ends and point out some questions which have been left open.

Two-dimensionalism holds, firstly, that our concepts are associated with substantial a priori information, and secondly, that there is a close connection between conceptual/epistemic and metaphysical modality. It therefore promises to play a major part in the endeavor of (re-)establishing conceptual analysis as a valuable philosophical method. Let me review how two-dimensionalism lives up to this promise, starting with the latter of the theses just mentioned:

The discussion showed that the standard cases of the necessary a posteriori can be accounted for within the two-dimensionalist framework. Contrary to what has often been claimed, these cases provide no obstacles to maintaining an intimate link between epistemic and metaphysical modality (cf. chapter 2). Counterexamples to metaphysical plenitude would therefore have a structure which is very different from the standard Kripkean cases. Although I argued that there are (mainly) epistemic reasons for thinking that there are no such strong necessities, I did not rule out their existence (cf. chapter 6). One might thus wonder whether they could pose a problem for conceptual analysis. For if it is such an important component in a defense of conceptual analysis to account for the classical a posteriori necessities, then it is natural to think that the possible existence of another kind of a posteriori necessity which cannot be accounted for is a reason to worry.

There are good reasons to reject this thought, though. What makes the Kripkean a posteriori necessities problematic in the first place is that they are quite frequent, that their metaphysical and epistemic status is highly compelling and that they are (seemingly) well in line with a causal theory of reference and with essentialism about individuals and natural kinds. In contrast, there is not a single uncontroversial example of a strong necessity and there is no independent motivation for their existence in sight, either. The most elaborate attempt to make their existence plausible is based on putative features of phenomenal concepts (cf. 6.1.2.1) and therefore highly limited in scope. If such a view turned out to be correct, it would nevertheless be unproblematic for the prospects of conceptual analysis in general. Accordingly, the existence of strong necessities would only raise a serious problem for conceptual analysis if they concerned a substantial number of expression types. Since there are no indications for this, a proponent of conceptual analysis need not be overly concerned by a possible failure of metaphysical plenitude.

Let me turn to the other main contribution of two-dimensionalism to the defense of conceptual analysis, namely its postulation of substantial a priori information which is derivable from our concepts: Two-dimensionalism claims that the implications of our concepts are substantial enough to make (CJ++) true, i.e. to determine their extension with respect to every world considered as actual. For a number of reasons, the resulting scrutability thesis can be considered as the central component of a defense of conceptual analysis. It therefore figured in a number of considerations throughout this book. It turned out that (CJ++) can even be defended with respect to natural kind terms and proper names, which are often held to enable no a priori inferences whatsoever (cf. chapter 3). Furthermore, the idea that our concepts are associated with primary intensions was confirmed by considerations about our ability to determine our subject matter and about communication (cf. chapter 4).

One may think that such a result calls out for an explanation: Why is it that we are able to determine the extension of every linguistic expression we understand with respect to hypothetical scenarios? I think two-

dimensionalism provides a natural answer to that question: Plausibly, the underlying idea of the two-dimensionalist account is a 'voluntaristic' picture of meaning: Since we are the ones who use language, we also decide, at least implicitly, what is represented by linguistic expressions. In other words: Speaker associations determine an expression's extension across the set of scenarios; this idea is most evident in Jackson's descriptivism. It is therefore no wonder that a speaker can make the relevant judgments.

While I agree with many aspects of the general picture just sketched, I think it is overly individualistic; this tendency is particularly salient in Chalmers' account. As we saw, the sole dependence of the meaning of an expression on a speaker's associations leads to problems in the context of deferential usage and incomplete understanding. I therefore proposed the following interpretation of the two-dimensionalist account: The properties a subject associates with a linguistic expression determine the content of the concept or thought which is expressed by an utterance of that expression. In a case where these properties are those associated by a significant proportion of speakers within the subject's linguistic community, it will typically be plausible to say that they also mirror the expression's meaning (cf. chapter 3).

The idea that the properties associated by a subject with an expression determine the content of the thought expressed may raise the following question, which I have not addressed in this book: How does a subject manage to associate expressions with properties in the first place? In other words, if primary intensions (primarily) reflect the content of mental states, then how is it that mental states come to have contents? It seems that there is a notable gap in the two-dimensionalist account: If the meaning of linguistic expressions is explained by the content of thoughts, then the explanation is incomplete unless one also provides an explanation of how thoughts represent.

Notice, however, that my goal in this book was not to explain how our expressions come to have their meanings, but rather to show that their meanings, i.e. their application conditions, can be appropriately modeled by

means of the two-dimensionalist framework. Furthermore, the problem just sketched is just a version of the general problem of intentionality, i.e. the problem of explaining how mental states manage to represent features of the world. This problem is indeed a difficult one, but it is a problem which concerns any kind of account, not just a two-dimensionalist one.
I should note that the latter point is contested. Stalnaker argues that two-dimensionalism involves a commitment to an essentially internalist account of intentionality. Furthermore, he believes that only an externalist account provides the means for solving the problem of intentionality (cf. Stalnaker 2006). I do not think that the latter point is correct. In my view, it is rather a merit of two-dimensionalism that it leaves room for a notion of narrow content. I nevertheless concede that the challenge raised by Stalnaker is a legitimate one and that two-dimensionalism is essentially incomplete unless it is combined with an account which addresses the problem of intentionality. This issue will have to be left to future research.

Two-dimensionalism can only provide the basis for a defense of conceptual analysis, the viability of which is dependent on a number of other issues. These issues were mainly discussed in the second part of this book. Let me briefly sum up the main results of this discussion:
It is plausible that the existence of primary intensions can only vindicate conceptual analysis as a philosophical enterprise if they are shared among the speakers of a linguistic community. Considerations about the role of primary intensions in communication suggest that while not all of our terms have commonly shared primary intensions, there are good reasons for thinking that primary intensions are usually shared among speakers, even in the case of epistemically opaque terms (cf. chapter 4). With respect to transparent terms, it is in any case very plausible that they are shared among competent speakers. Therefore, my discussion on transparency and opacity, which showed that a great number of our terms are indeed transparent, provided one more reason to be optimistic that primary intensions are frequently shared. Notice that the result of the discussion on epistemic transparency is important for yet another reason; at least

depending on one's goals, opaque expressions are less suitable for conceptual analysis than transparent ones. Nevertheless, the scope of conceptual analysis need not be limited to epistemically transparent expressions: I also sketched how uncovering the primary intensions of opaque terms can be useful as well (cf. chapter 5, chapter 7).

Two-dimensionalism undeniably involves a number of significant idealizations. These idealizations can be useful, and are maybe even required, in a systematic semantic account. But at the same time, their presence raises the question of how for instance the ideal judgments invoked in the scrutability thesis relate to those in our epistemic practice. Our actual judgments can certainly go wrong in many ways. I argued, however, that skepticism concerning our ability to determine the extension of our concepts with respect to hypothetical scenarios is unfounded, in particular since there exist several routes to achieving that aim. According to the resulting view, conceptual analysis can surely be tedious, but it is nevertheless feasible (cf. chapter 6).

Finally, a survey of a number of philosophical projects involving conceptual analysis showed that while there are applications of the method which do not rely on explicit analyses, for many purposes it is useful, or even necessary, to be able to provide definitions. Accordingly, even though this issue is not decisive for the tenability of conceptual analysis, it is still relevant to its scope and its utility. I therefore examined the prospects of the project of giving definitions, which is often regarded as a complete failure. I identified a number of practical impediments, but found no principled reasons to believe that they cannot be overcome. I concluded with some considerations which suggest that the general project, even though it has not yet brought any great successes, may still be progressing (cf. chapter 7).

References

Armstrong, Sharon; Gleitman, Lila and Gleitman, Henry (1983). What Some Concepts Might Not Be. *Cognition* 13(3): 263–308.

Arntzenius, Frank (2003). Some Problems for Conditionalization and Reflection. *Journal of Philosophy* 100(7): 356–370.

Barner, Alma (manuscript). *Kripke on Modal Error*.

Bealer, George (1998). Intuition and the Autonomy of Philosophy. In M. DePaul and W. Ramsey (eds.), *Rethinking Intuitions*. Lanham, MD: Rowman & Littlefield, 201–239.

———————. (2006). The Origins of Modal Error. *Dialectica* 58(1): 11–42.

Block, Ned and Stalnaker, Robert (1999). Conceptual Analysis, Dualism, and the Explanatory Gap. *Philosophical Review* 108(1): 1–46.

BonJour, Laurence (1985). *The Structure of Empirical Knowledge*. Cambridge, MA: Harvard University Press.

———————. (1998). *In Defense of Pure Reason*. Cambridge: Cambridge University Press.

Brown, James (1991). *The Laboratory of the Mind: Thought Experiments in the Natural Sciences*. London: Routledge.

Burge, Tyler (1979). Individualism and the Mental, *Midwest Studies in Philosophy* 4: 73–122.

Byrne, Alex and Pryor, James (2006). Bad Intensions. In M. García-Carpintero and J. Macià (eds.), *Two-Dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press, 38–54.

Carnap, Rudolf (1947/1956). *Meaning and Necessity*. Chicago, IL: The University of Chicago Press.

Castañeda, Hector-Neri (1967). Omniscience and Indexical Reference. *Journal of Philosophy* 64(7): 203–210.

Casullo, Albert (2003). *A Priori Justification*. Oxford: Oxford University Press.

————————. (2010). Knowledge and Modality. *Synthese* 172(3): 341–359.

Chalmers, David (1996). *The Conscious Mind. In Search of a Fundamental Theory*. Oxford: Oxford University Press.

————————. (2002a). Does Conceivability Entail Possibility? In T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*. Oxford: Oxford University Press, 145–200.

————————. (2002b). On Sense and Intension. *Philosophical Perspectives* 16: 135–182.

————————. (2002c). The Components of Content. In D. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press, 608–633.

————————. (2003). The Nature of Narrow Content. *Philosophical Issues* 13(1): 46–66.

————————. (2004). Epistemic Two-Dimensional Semantics. *Philosophical Studies* 118(1–2): 153–226.

————————. (2006). The Foundations of Two-Dimensional Semantics. In M. García-Carpintero and J. Macià (eds.), *Two-Dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press, 55–140.

————————. (2007). Phenomenal Concepts and the Explanatory Gap. In T. Alter and S. Walter (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford: Oxford University Press, 167–194.

————————. (2010). *The Character of Consciousness.* New York: Oxford University Press.

————————. (2011a). The Nature of Epistemic Space. In A. Egan and B. Weatherson (eds.), *Epistemic Modality*. Oxford: Oxford University Press, 60–107.

————————. (2011b). Propositions and Attitude Ascriptions: A Fregean Account. *Noûs* 45(4): 595–639.

————————. (forthcoming). *Constructing the World*. Oxford: Oxford University Press.

Chalmers, David and Jackson, Frank (2001). Conceptual Analysis and Reductive Explanation. *Philosophical Review* 110(3): 315–361.

Chappell, Vere (ed.) (1999). *Hobbes and Bramhall on Liberty and Necessity*. Cambridge: Cambridge University Press.

Chisholm, Roderick M. (1976). *Person and Object*. Chicago & La Salle, IL: Open Court.

———————. (1989). *Theory of Knowledge*. 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Cohen, Stewart and Lehrer, Keith (1983). Justification, Truth, and Knowledge. *Synthese* 55(2): 191–207.

Comesaña, Juan (2005). Unsafe Knowledge. *Synthese* 146(3): 395–404.

Cresswell, Max (1985). *Structured Meanings*. Cambridge, MA: MIT Press.

De Brabanter, Philippe; Nicolas, David; Stojanović, Isidora and Villanueva Fernández, Neftalí (2005). *Deferential Utterances*. Online publication. URL: http://jeannicod.ccsd.cnrs.fr/docs/00/05/36/20/PDF/ijn_00000575_00.pdf (last retrieved: 08/02/2012).

Devitt, Michael and Sterelny, Kim (1999). *Language and Reality: An Introduction to the Philosophy of Language*. 2nd ed. Cambridge, MA: MIT Press.

Dretske, Fred (1995). *Naturalizing the Mind*. Cambridge, MA: MIT Press.

Feyerabend, Paul (1962). Explanation, Reduction and Empiricism. In H. Feigl and G. Maxwell (eds.), *Scientific Explanation, Space, and Time*, Minnesota Studies in the Philosophy of Science, Volume III. Minneapolis, MN: University of Minneapolis Press, 28–97.

Field, Hartry (1996). The A Prioricity of Logic. *Proceedings of the Aristotelian Society* 96: 359–379.

Fine, Kit (1994). Essence and Modality. *Philosophical Perspectives* 8: 1–16.

Frankfurt, Harry (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66(23): 829–839.

Frege, Gottlob (1892/2002). Über Sinn und Bedeutung. In M. Textor (ed.), *Gottlob Frege. Funktion – Begriff – Bedeutung*, Göttingen: Vandenhoeck & Ruprecht, 23–46.

Gettier, Edmund (1963). Is Justified True Belief Knowledge? *Analysis* 23(6): 121–123.

Goldman, Alvin (1976). Discrimination and Perceptual Knowledge. *Journal of Philosophy* 73: 771–791.

————————. (1993). Epistemic Folkways and Scientific Epistemology. *Philosophical Issues* 3: 271–285.

————————. (2007). Philosophical Intuitions: Their Target, Their Source, and Their Epistemic Status. *Grazer Philosophische Studien* 74(1): 1–26.

Haas-Spohn, Ulrike (1997). The Context-Dependency of Natural Kind Terms. In W. Künne, M. Anduschus & A. Newen (eds.), *Direct Reference, Indexicality and Propositional Attitudes*, Stanford: CSLI-Publications, 333–349.

Hare, Richard (1982). *Moral Thinking: Its Levels, Methods and Point*. Oxford: Oxford University Press.

————————. (1991). *The Language of Morals.* Oxford: Oxford University Press.

Heck, Richard (1995). The Sense of Communication. *Mind* 104(413): 79–106.

Hill, Christopher and McLaughlin, Brian (1999). There are Fewer Things in Reality Than are Dreamt of in Chalmers's Philosophy. *Philosophy and Phenomenological Research* 59(2): 445–454.

Horvath, Joachim (2009). The Modal Argument for A Priori Justification. *Ratio* 22(2): 191–205.

Jackson, Frank (1998a). *From Metaphysics to Ethics. A Defence of Conceptual Analysis.* Oxford: Clarendon Press.

————————. (1998b). Reference and Description Revisited. *Philosophical Perspectives* 12, Language, Mind, and Ontology: 201–218.

————————. (2001). Locke-ing onto Content. In D.M. Walsh (ed.), *Naturalism, Evolution and Mind*. Cambridge: Cambridge University Press, 127–144.

———————. (2002). Language, Thought, and the Epistemic Theory of Vagueness. *Language & Communication* 22(3): 269–279.

———————. (2003). Representation and Narrow Belief. *Philosophical Perspectives* 13(1): 99–112.

———————. (2004). Why We Need A-Intensions. *Philosophical Studies* 118(1–2): 257–277.

———————. (2005). Ramsey Sentences and Avoiding the *Sui Generis*. In H. Lillehammer and D.H. Mellor (eds.), *Ramsey's Legacy*. Oxford: Clarendon Press, 123–136.

———————. (2007). On Not Forgetting the Epistemology of Names. *Grazer Philosophische Studien* 74(1): 239–250.

———————. (2010). *Language, Names, and Information*. Oxford: Wiley-Blackwell.

Jackson, Frank; Mason, Kelby and Stich, Steve (2009). Folk Psychology and Tacit Theories: A Correspondence between Frank Jackson, and Steve Stich and Kelby Mason. In D. Braddon-Mitchell and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism*. Cambridge, MA: MIT Press, 45–97.

Jones, Janine (2004). Illusory Possibilities and Imagining Counterparts. *Acta Analytica* 19(32): 19–43.

Kaplan, David (1989). Demonstratives. In J. Almog, J. Perry and H. Wettstein (eds.), *Themes from Kaplan*. Oxford: Oxford University Press, 481–563.

Kelp, Christoph (2009): Knowledge and Safety. *Journal of Philosophical Research* 34, 21–31.

Kim, Jaegwon (1988). What is Naturalized Epistemology? In J. Tomberlin (ed.), *Philosophical Perspectives* 2. Asascadero, CA: Ridgeview Publishing Co: 381–406.

———————. (2005). *Physicalism, or Something Near Enough*. Princeton, NJ: Princeton University Press.

Kipper, Jens (2010). Philosophers and Grammarians. *Philosophical Psychology* 23(4): 511–527.

Kornblith, Hilary (1980). Referring to Artifacts. *Philosophical Review* 89(1), 109–114.

————————. (1998). The Role of Intuition in Philosophical Inquiry. An Account with No Unnatural Ingredients. In M. DePaul and W. Ramsey (eds.), *Rethinking Intuition*. Lanham, MD: Rowman & Littlefield, 129–141.

————————. (2002). *Knowledge and its Place in Nature*. Oxford: Clarendon Press.

————————. (2007a). Naturalism and Intuitions. *Grazer Philosophische Studien* 74(1): 27–49.

————————. (2007b). How to Refer to Artifacts. In E. Margolis and S. Laurence (eds.), *Creations of the Mind: Essays on Artifacts and Their Representation.* Oxford: Oxford University Press, 138–149.

Kripke, Saul (1971). Identity and Necessity. In M. Munitz (ed.), *Identity and Individuation*. New York: New York University Press, 135–164.

————————. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kroon, Frederick (1983). The Problem of 'Jonah': How *Not* to Argue For the Causal Theory of Reference. *Philosophical Studies* 43(2): 281–299.

————————. (1987). Causal Descriptivism. *Australasian Journal of Philosophy* 65(1): 1–17.

————————. (1989). Circles and Fixed Points in Description Theories of Reference. *Noûs* 23(3): 373–382.

————————. (2004). A-Intensions and Communication. *Philosophical Studies* 118(1–2): 279–298.

————————. (2009) Names, Plans, and Descriptions. In D. Braddon-Mitchell and R. Nola (eds.), *Conceptual Analysis and Philosophical Naturalism.* Cambridge, MA: MIT Press, 139–158.

Kuhn, Thomas (1962). *The Structure of Scientific Revolutions*. Chicago, IL: University of Chicago Press.

Laurence, Stephen and Margolis, Eric (2003). Concepts and Conceptual Analysis. *Philosophy and Phenomenological Research* 67(2), 253–282.

———————. (2005/2011). Concepts. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford: The Metaphysics Research Lab Center for the Study of Language and Information, Stanford University. URL: http://plato.stanford.edu/entries/concepts (last retrieved: 08/02/2012).

Lehrer, Keith (1990). *Theory of Knowledge*. Boulder, CO: Westview Press.

Levine, Joseph (1993). On Leaving out What It's Like. In M. Davies and G.W. Humphries (eds.), *Consciousness. Psychological and Philosophical Essays.* Oxford: Blackwell, 121–136.

Lewis, David (1970). How to Define Theoretical Terms. *Journal of Philosophy* 67(13): 427–446.

———————. (1972). *Psychological and Theoretical Identifications*. In ibid., *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press, 248–261.

———————. (1979). Attitudes *De Dicto* and *De Se*. *Philosophical Review* 88(4): 513–543.

———————. (1983). Mad Pain and Martian Pain. In ibid., *Philosophical Papers 1*. Oxford: Oxford University Press: 122–130.

———————. (1984). Putnam's Paradox. *Australasian Journal of Philosophy* 62(3), 221–236.

———————. (1994). Reduction of Mind. In S. Guttenplan (ed.), *A Companion to Philosophy of Mind*. Oxford: Blackwell, 412–431.

———————. (1994). Finkish Dispositions. *The Philosophical Quarterly* 47: 143–158.

———————. (1999). Naming the Colours. In ibid., *Papers in Metaphysics and Epistemology.* Cambridge: Cambridge University Press, 332–358.

———————. (2004). Causation as Influence. In J. Collins, N. Hall, and L.A. Paul (eds.), *Causation and Counterfactuals*. Cambridge, MA: MIT Press, 75–106.

Lowe, E. Jonathan (1996). *Subjects of Experience*. Cambridge: Cambridge University Press.

Lycan, William (1988). *Judgement and Justification*. Cambrigde: Cambridge University Press.

————————. (2009). Serious Metaphysics: Frank Jackson's Defense of Conceptual Analysis. In I. Ravenscroft (ed.), *Minds, Ethics and Conditionals: Essays in Honour of Frank Jackson*. Oxford: Oxford University Press, 61–83.

Machery, Edouard; Mallon, Ron; Nichols, Shaun and Stich, Stephen (2004). Semantics, Cross-Cultural Style. *Cognition* 92: B1–B12.

Mackie, John (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Mallon, Ron; Machery, Edouard; Nichols, Shaun and Stich, Stephen P. (2009). Against Arguments from Reference. *Philosophy and Phenomenological Research* 79(2): 332–356.

Menzies, Peter (1998). Possibility and Conceivability: A Response-Dependent Account of Their Connections. In R. Casati (ed.), *European Review of Philosophy* 3: 255–277.

Millikan, Ruth (2005). *Language: A Biological Model*. Oxford: Oxford University Press.

————————. (2010). On Knowing the Meaning. With a Coda on Swampman. *Mind* 119(473): 43–81.

Nimtz, Christian (2004). Two-Dimensionalism and Natural Kind Terms. *Synthese* 138(1): 125–148.

————————. (2007). Kripke vs. Kripke – eine bescheidene Verteidigung der Kennzeichnungstheorie. In A. Rami and H. Wansing (eds.), *Referenz und Realität*. Paderborn: mentis, 99–121.

Ninan, Dilip (2010). De Se Attitudes: Ascription and Communication. *Philosophy Compass* 5(7): 551–567.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Clarendon Press.

Perry, John (1979). The Problem of the Essential Indexical. *Noûs* 13(1), 3–21.

Polger, Thomas (2008). H_2O, 'Water', and Transparent Reduction. *Erkenntnis* 69(1): 109–130.

Putnam, Hilary (1962). It Ain't Necessarily So. *Journal of Philosophy* 59(22): 658–671.

————————. (1970). Is Semantics Possible? *Metaphilosophy* 1(3): 187–201.

————————. (1975). The Meaning of 'Meaning'. In ibid., *Mind, Language and Reality – Philosophical Papers* 2. Cambridge: Cambridge University Press: 215–271.

————————. (1990). Is Water Necessarily H_2O? In ibid., *Realism With a Human Face*, ed. by James Conant. Cambridge, MA: Harvard University Press, 54–79.

Quine, Willard van Orman (1951). Two Dogmas of Empiricism. *Philosophical Review* 60(1): 20–43.

————————. (1969). Propositional Objects. In ibid., *Ontological Relativity and Other Essays*. New York: Columbia University Press, 139–160.

Ramsey, William (1992). Prototypes and Conceptual Analysis. *Topoi* 11(1): 59–70.

Rosch, Eleanor and Mervis, Catlin (1975). Family Resemblances: Studies in the Internal Structure of Categories. *Cognitive Psychology* 7(4): 573–605.

Sainsbury, Mark (1997). Easy Possibilities. *Philosophy and Phenomenological Research* 57(4), 90–919.

Salmon, Nathan (1981). *Reference and Essence.* Princeton, NJ: Princeton University Press.

Sayre-McCord, Geoffrey (1997). 'Good' on Twin Earth. *Philosophical Issues* 8: 267–292.

Schroeter, Laura (2006). Against *A Priori* Reductions. *The Philosophical Quarterly* 56(225): 562–586.

Schwarz, Wolfgang (2009). *David Lewis: Metaphysik und Analyse*. Paderborn: mentis.

Schwartz, Stephen (1977). Introduction, in S. Schwartz (ed.), *Naming, Necessity, and Natural Kinds*. Ithaca, NY: Cornell University Press, 13–41.

————————. (1978). Putnam on Artifacts. *Philosophical Review* 87(4): 566–574.

————————. (1980). Natural Kinds and Nominal Kinds. *Mind* 89(354): 182–195.

————————. (1983). Reply to Kornblith and Nelson. *Southern Journal of Philosophy* 21(3): 475–479.

Searle, John (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences* 3: 417–457.

Shoemaker, Sidney and Swinburne, Richard (1984). *Personal Identity*. Oxford: Blackwell.

Siegel, Susanna (2010). The Contents of Visual Experience. New York: OUP.

Siewert, Charles (1998). *The Significance of Consciousness*. Princeton: Princeton University Press.

Smith, Michael (1994). *The Moral Problem*. Oxford: Blackwell.

Soames, Scott (2002). *Beyond Rigidity: The Unfinished Semantic Agenda of Naming and Necessity*. Oxford/New York: Oxford University Press.

Soames, Scott (2004). Knowledge of Manifest Natural Kinds. *Facta Philosophica* 6: 159–181.

Sorensen, Roy (2001). *Vagueness and Contradiction*. Oxford: Oxford University Press.

Sosa, Ernest (2007). *A Virtue Epistemology. Apt Belief and Reflective Knowledge, Vol. 1*, Oxford: Oxford University Press.

Stalnaker, Robert (1978). Assertion. In P. Cole (ed.), *Syntax and Semantics 9: Pragmatics*. New York: Academic Press, 315–332.

————————. (1981). Indexical Belief. *Synthese* 49(1): 129–151.

————————. (2006). Assertion Revisited: On the Interpretation of Two-Dimensional Modal Semantics. In M. García-Carpintero and J.

Macià (eds.), *Two-Dimensional Semantics: Foundations and Applications*. Oxford: Oxford University Press, 293–309.

Stich, Stephen (1992). What is a Theory of Mental Representation? *Mind* 101(402): 243–261.

Stoljar, Daniel (2006). *Ignorance and Imagination*. Oxford: Oxford University Press.

Strawson, Patrick (1959). *Individuals*. London: Methuen.

Swain, Stacey; Alexander, Joshua and Weinberg, Joshua (2008). The Instability of Philosophical Intuitions. Running Hot and Cold on Truetemp. *Philosophy and Phenomenological Research* 76(1): 138–155.

Tidman, Paul (1994). Conceivability as a Test for Possibility. *American Philosophical Quarterly* 31(4): 297–309.

Titelbaum, Michael (2008). The Relevance of Self-locating Beliefs. *Philosophical Review* 117(4): 555–606.

Torre, Stephan (2010). Centered Assertion. *Philosophical Studies* 150(1): 97–114.

Tye, Michael (1995). *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

———————. (1999). Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion. *Mind* 108(432): 705–725.

Weatherson, Brian (2003). What Good Are Counterexamples? *Philosophical Studies* 115(1): 1–31.

———————. (2007). How to Challenge Intuitions Empirically Without Risking Skepticism. *Midwest Studies in Philosophy* 31(1): 318–343.

Weinberg, Jonathan; Nichols, Shaun and Stich, Stephen (2001). Normativity and Epistemic Intuitions. *Philosophical Topics* 29(1–2): 429–460.

Weinberg, Jonathan et al. (ms.). *Intuition and Calibration*.

Wiggins, David (1976). Locke, Butler and the Stream of Consciousness: and Men as a Natural Kind. In A.O. Rorty (ed.), *The Identities of Persons*. Berkeley, CA: University of California Press, 139–173.

Wille, Andrea (2010). *Apriori-Physikalismus? Reduktive Erklärung und begriffliche Ableitbarkeit*. Manuscript.

Williams, Bernard (1970). The Self and the Future. *The Philosophical Review* 79, 161–180.

Williamson, Timothy (1994). *Vagueness*. London: Routledge.

———————. (2000). *Knowledge and its Limits*. Oxford: Oxford University Press.

———————. (2007). *The Philosophy of Philosophy*. Malden, MA: Blackwell.

Yablo, Stephen (1993). Is Conceivability a Guide to Possibility? *Philosophy and Phenomenological Research* 53(1): 1–42.

———————. (2006). No Fool's Cold: Notes on Illusions of Possibility. In M. García-Carpintero and J. Macià (eds.), *Two-Dimensional Semantics: Foundations and Applications*. New York: Oxford University Press, 327–345.