Philosophical Naturalism
David Papineau

For Katy

This is a somewhat revised version of the book I published in 1993 with Blackwell. I have had thoughts of reworking the book for a prospective second edition. The following text incorporates revision made so far. For the most part, though, it is derived from the ultimate draft of the Blackwell book, though without any changes added in proof. The Bibliography is also currently missing. I apologize for the scruffiness of the text, which can be attributed to repeated copying from an electronic original composed two decades ago.

The main revision so far is Chapter 1, which has been substantially rewritten, originally for the Greek edition (2002); some consequent changes have also been made to chapter 3 (at the end of section 3.10).

DP
24 Aug 2013

Editor's note
Original document: "*philosophical_naturalism_for_website_aug_2013_org.docx*", downloaded from https://www.davidpapineau.co.uk/books-incl-online-philosophical-naturalism.html has hereby been edited. More than two thousand errors from repeated copying, typically extra spaces in words and typos, were corrected (there might still be more). Typology changes were made throughout the document. Furthermore, page numbers and table of contents were added.

Jakob H. Heidelberg, 11 Dec 2025 (no rights reserved)

CONTENTS

Contents 2

Introduction 5

Chapter 1 Supervenience and Identity 8

1.1 Introduction 8

1.2 Supervenience 8

1.3 More Preliminaries 10

1.4 The Completeness of Physics 12

1.5 The Manifestability Argument for Supervenience 12

1.6 Manisfestability is Not Enough 14

1.7 The Causal Argument for Physicalism 16

1.8 Generous Causation and Alternatives to Type Identity 18

1.9 The Completeness of Physics Defended 20

Extra References 23

Chapter 2 Reduction and Selection 26

2.1 Introduction 26

2.2 An Unexplained Coincidence 26

2.3 Laws in the Special Sciences 28

2.4 Functionalism and Dispositions 29

2.5 The Irreducibility of Ordinary Dispositions 31

2.6 The Meaning of Reduction 32

2.7 Teleology and Irreducibility 33

2.8 Selectional Explanations 35

Chapter 3 The Teleological Theory of Representation 40

3.1 Introduction 40

3.2 Functionalism and Representation 40

3.3 Broad Contents 42

3.4 Functional Falsity 43

3.5 The Reality of Beliefs and Desires 46

3.6 Truth as the Guarantee of Success 48

3.8 Alternative Accounts of Desire Satisfaction 54

3.9 Do We Need Reified Truth Conditions? 56

3.10 Broad Contents Revisited 60

3.11 Accidental Replicas 61

3.12 Empirical Evidence for the Teleological Theory 63
3.13 Verificationism Refuted 67
Chapter 4 Consciousness and the Antipathetic Fallacy 73
4.1 Introduction 73
4.2 What is it Like to be a Bat? 73
4.3 What Mary Didn't Know 75
4.4 Concepts of Experience 77
4.5 The Antipathetic Fallacy 79
4.6 Theories of Consciousness 82
4.7 Life and Consciousness 84
4.8 Consciousness is Vague 86
4.9 Consciousness is Anthropocentric 87
4.10 Pains, Shapes and Colours 89
Chapter 5 Reliabilism, Induction, and Scepticism 99
5.1 Introduction 99
5.2 Knowledge and the Project of Enquiry 99
5.3 Certainty and Reliability 100
5.4 Knowledge and Normativity 101
5.5 Rationales versus Intuitions 101
5.6 Knowledge and Percentages 103
5.8 The Attractions of Certainty 105
5.9 The Problem of Induction? What Problem? 106
5.10 Is Induction Reliable? 106
5.11 The Legitimacy of Normal Methods 108
5.12 Varieties of Circularity 108
5.13 Who Needs Persuading? 109
5.14 The Strategy Generalized 111
5.15 Non-Enumerative Induction 112
5.16 The Historical Contingency of Knowledge 115
5.17 The Pessimistic Meta-Induction 116
Chapter 6 Mathematics and Other Non-Natural Subjects 121
6.1 Introduction 121
6.2 Content and Knowledge 122
6.3 Anti-Realism versus Fictionalism for Mathematics 123

6.4 If-Thenism 124
6.5 Postulationism 125
6.6 Reductionism 127
6.7 Neo-Fregeanism 130
6.8 Mathematical Anti-Realism Rejected 132
6.9 The Putnam-Quine Defence of Mathematics 134
6.10 Field's Fictionalism 135
6.11 Morality and Modality 138
6.12 The Non-Doxastic Alternative 140
6.13 Why are Morality and Modality Non-Doxastic? 140
6.14 Why isn't Mathematics Non-Doxastic? 142
6.15 Fictionalism and Modality 144

INTRODUCTION

What is philosophical "naturalism"? The term is a familiar one nowadays, but there is little consensus on its meaning. For some philosophers, the defining characteristic of naturalism is the affirmation of a continuity between philosophy and empirical science. For others the rejection of dualism is the crucial requirement. Yet others view an externalist approach to epistemology as the essence of naturalism.

I shall not engage directly with this issue. It is essentially a terminological matter. The important question is which philosophical positions are right, not what to call them. I suspect that the main reason for the terminological unclarity is that nearly everybody nowadays wants to be a "naturalist", but the aspirants to the term nevertheless disagree widely on substantial questions of philosophical doctrine. The moral is that we should address the substantial philosophical issues first and worry about the terminology afterwards. Once we have worked out which commitments ought to be upheld by philosophers who aspire to "naturalism", then we can agree to use the term accordingly.

As it happens, I am in favour of a naturalist answer to all the general questions raised above, as will become clear from the arguments which follow: that is, I am against dualism and epistemological internalism, and in favour of the view that philosophy is continuous with empirical science. But there is a further strand to my naturalism, which I shall defend in the first two chapters, and which takes it beyond these general commitments. This further commitment is physicalism, the thesis that all natural phenomena are, in a sense to be made precise, physical.

At one time I intended to call this book "Philosophical Physicalism" rather than "Philosophical Naturalism", on the grounds that "physicalism" is a more informative term than "naturalism". But I decided against this for two reasons. First, the title "Physicalism" might have carried the suggestion that my philosophical stance is tied to the categories of current physical theory, whereas my position, as we shall see, is formulated, not in terms of current physics, but in terms of the science of whatever categories eventually turn out to be needed to explain the behaviour of matter. And, second, the last two-thirds of the book move away from the details of physicalism as such, and address issues about mind and knowledge that arise, not just for strict physicalists, but for anyone of more generally naturalist inclinations.

In detail the plan of the book is as follows. There are three sections -- Physicalism, Mind, Knowledge -- and each section contains two chapters.

In the first chapter I argue that physicalism is not a prejudice, but a consequence of some evident truths. The second chapter then argues that physicalism also requires reductionism, except about phenomena that are the products of selection processes.

In the second section I discuss mental representation and consciousness. Chapter 3 offers a detailed version of the teleological theory of mental representation, a theory which I have defended in previous writings. I explain the relationship between this theory and other views, and I defend it against various objections. Chapter 4 deals with consciousness. I argue that there is nothing in consciousness to threaten physicalism, and I try to unravel some of the reasons why consciousness has seemed to many philosophers to offer such a threat.

In the final section, on Knowledge, I offer a principled defence of a reliabilist theory of knowledge, a defence which shows how reliabilism can yield an adequate response to the problem of induct ion, and to sceptical arguments generally. This defence of reliabilism comprises chapter 5. Chapter 6 then addresses the special epistemological issues that arise for mathematical knowledge, and considers some comparisons between mathematical, moral, and modal knowledge.

I said above that my overall position will imply a naturalist stance on the issues of dualism, epistemological externalism, and the continuity of philosophy with the empirical sciences. The first two issues will be dealt with at length in what follows; in particular, the arguments of sections 1 and 2 will bear on dualism, and section 3 will be concerned with epistemology. But the third issue, the continuity of philosophy with empirical science, will not be explicitly discussed in the rest of the book. So let me conclude this introduction with some brief comments on this topic.

At one level, the continuity of philosophy and empirical science is uncontentious. Many philosophical problems arise because of apparent tensions or conflicts within the assumptions which empirical evidence recommends to us. The most obvious examples are issues in the philosophy of science, such as problems about the interpretation of quantum mechanics, or the asymmetry of time, or the logic of natural selection. But other less specialist philosophical questions, like the existence of free will, also arise because of difficulties raised by empirical assumptions (in particular, in this case, by assumptions about the extent to which human beings are subject to the same laws of nature as the rest of the world).

This is not to say that these philosophical issues are no different from the kinds of issues normally address ed by natural scientists. Philosophical problems are characterized by a special kind of difficulty, a difficulty which means that they cannot be solved, as scientific problems normally are, simply by the uncovering of further empirical evidence. Rather they require some conceptual unravelling, a careful unpicking of implicit ideas, often culminating in the rejection of assumptions we didn't realize we had. But, still, despite these differences, there is clearly a sense in which philosophical thinking of this kind is part and parcel of the construction of scientific theories. Even if there is no direct involvement with empirical evidence, the task of the philosophers is to bring coherence and order to the total set of assumptions we use to explain the empirical world.

The question at issue is whether all philosophical theorizing is of this kind. Naturalists will say that it is. Those with a more traditional attitude to philosophy will disagree. These traditionalists will allow, of course, that some philosophical problems, problems in applied philosophy, as it were, will fit the above account. But they will insist that when we turn to "first philosophy", to the investigation of such fund amental categories thought and knowledge, then philosophy must proceed independently of science.

Naturalists will respond that there is no reason to place even first philosophy outside science. They will point out that even the investigation of basic topics like thought and knowledge needs to start somewhere, with some assumptions about the nature of the human mind and its relation to the rest of reality. Without any assumptions to work from, investigation would be paralyzed. And the obvious strategy, naturalists will argue, is to begin with our empirically best-attested theories of the mind and its relation to reality, and use these as a framework within which to raise and resolve philosophical difficulties, in the way outlined above.

Traditionalists will counter that we are not entitled to any empirically-based assumptions until we have somehow established the legitimacy of empirical knowledge by independent means. Maybe, they will concede, we need some assumptions of some sort to start with. But, on pain of pre-empting important philosophical questions, they had better be assumptions we can establish by such arguably incontrovertible methods as introspection, conceptual analysis, or deduction, and not assumptions which rest on the all-too-questionable principles of empirical investigation.

This argument, that philosophy needs firmer foundations than those available within empirical science, has undoubtedly been of great influence on the modern Western conception of philosophical method. But it is important to realize that this argument itself derives from various specific philosophical assumptions, and is by no means a necessary consequence of the very idea of philosophical activity. In particular, as I shall show in chapter 5, this argument depends on the assumption that claims to knowledge need to be certain, in the sense that they should derive from methods that necessarily deliver truths. Once you accept this requirement on knowledge, then you will indeed demand that philosophical knowledge in particular should come from such arguably incontrovertible methods as introspection, conceptual analysis, and deduction; and the epistemological status of science will remain in question until such time as philosophy succeeds in showing how it too satisfies the demand of certainty.

On the other hand, if we reject the idea that knowledge demands certainty, as I shall urge in chapter 5, then this whole line of argument for first philosophy falls away. For, as I shall also show in chapter 5, the rejection of certainty removes the rationale for restricting our initial methods to introspection, analysis, and deduction, and therewith removes the rationale for eliminating scientific assumptions from the framework within which we do philosophy.

So the dialectical situation is as follows. If you hold that knowledge requires certainty, then you will hold that philosophy needs to come before science. If you reject this demand, as I shall in chapter 5, then you will have reason to regard philosophy as continuous with science. But there is also a prior procedural question, about which philosophical methodology should be used to address this issue: that is, when we address the issue of whether knowledge require certainty, should we do so within the constraints of first philosophy, or as a topic within a naturalized philosophy? When I turn to this topic of knowledge and certainty in chapter 5, I shall proceed in the latter way, and conduct my argument within the framework of various empirical assumptions about the nature and needs of human beings. My defence of this strategy is that the onus surely lies with those who want to exclude relevant and well-confirmed empirical claims from philosophical debate to provide some prior rationale for doing so. (If there are readers who find this unconvincing, I would ask them to wait until chapter 5 before passing final judgement; it is relevant that the empirical assumptions I use there are not esoteric discoveries of physiological theory, but mundane truisms about human capabilities.)

One last point about the relationship between philosophy and science. If we set philosophy within science, this does not mean that the epistemological status of science is not itself a proper topic for philosophical debate. Naturalism can perfectly well investigate the status of scientific knowledge, and indeed much of chapter 5 below will consist of just such an investigation. All that naturalism claims is that this investigation, like any other philosophical investigation, is best conducted with the framework of our empirical knowledge of the world.

CHAPTER 1 SUPERVENIENCE AND IDENTITY

1.1 Introduction

Like many other contemporary philosophers, I have strong physicalist intuitions. I am inclined to think that chemical phenomena, for example, are all at bottom physical, even though chemists do not describe those phenomena in physical terms. What is more, I am inclined to think the same about the phenomena studied by meteorology, biology, psychology, sociology and the other so-called "special sciences".

My aim in this initial chapter is to see how far such physicalist intuitions can be supported by serious arguments. This question is not as much discussed in the contemporary philosophical literature as it might be. Of course many philosophers with physicalist inclinations have formulated different possible versions of physicalism, and explored the relations between them. And many other philosophers, with opposed inclinations, have elaborated various non-physicalist views of psychology, biology, sociology, and other special phenomena. But for the most part neither party has paused to argue its case against the other. The friends of physicalism tend simply to start with their physicalist intuitions, and try to develop a theory which fits them. Their opponents dismiss those intuitions out of hand as symptoms of an overblown admiration for science.

Not all philosophers treat physicalism as beyond debate in this way. An increasing number of contemporary thinkers are coming to recognize that there are plenty of pertinent arguments that bear on the issue.[1] Dogmatic physicalists and anti-physicalists alike will do well to attend to these arguments. Anti-physicalists will discover that physicalism is supported by premises which are difficult to deny, even if you have little regard for science. And physicalists will find out why some versions of physicalism are defensible, while others are not.

1.2 Supervenience

Let me start by trying to be a bit more precise about what I mean by physicalism. One simple way of formulating physicalism would be to require that all special properties, like chemical, or biological, of psychological properties, should be identified as types with physical properties, in the way that the property of being hydrogen, say, can be identified with the physical property of having atoms with one proton and one electron. But while such "type identities" may be available within basic chemistry, they seem unlikely to characterize the other special sciences. In particular, it seems unlikely that psychological properties, such as being worried about the future, for example, can be identified with any specific physical properties, along the lines of having a certain arrangement of molecules in your head. It is surely implausible to suppose that all the different people who have ever been worried about the future must have some intra-cranial molecular property in common. And, if that is not implausible enough, what about the future brain-injured people who will have their damaged parts replaced by miracles of silicon-based micro-technology, or the hominid but silicon-based denizens of Proxima Centauri's third planet? Presumably they will be able to worry about the future too. But they can't possibly share molecular arrangements with the rest of us, given that we don't have any silicon in our brains.

Fortunately for physicalism, type identity is not the only way in which special properties can be viewed as essentially physical. An alternative way of formulating physicalism is in terms of the supervenience of the special on the physical. Supervenience on the physical means that two systems cannot differ chemically, or biologically, or psychologically, or whatever, without differing physically; or, to put it the other way round, if two systems are physically

identical, then they must also be chemically identical, biologically identical, psychologically identical, and so on.

The advantage of formulating physicalism in terms of supervenience is that, unlike type identity, this doesn't require that the same physical property must determine a given special property whenever it is instanced. My worrying about the future might involve one molecular arrangement, an arrangement such that that anybody who has it will be worrying about the future; your worrying about the future might be ensured by a different physical arrangement, but again one that suffices to determine that all its possessors are worrying about the future; future brain-damaged patients and Proxima Centaurians will have yet different such physical arrangements; and so on.

How satisfactory an explication of physicalism is the requirement of supervenience on the physical? I shall consider first whether supervenience is necessary for physicalism, second whether it is sufficient.

On the face of it, supervenience seems an obvious necessary condition for physicalism in any given area: if two chemical systems, say, can differ, even though they are physically identical, then it would seem to follow that they must contain something non-physical.

However, an immediate qualification is needed. Suppose two chemical samples are physically identical: they contain exactly the same molecules and have exactly the same internal structure. Nevertheless one may be heavier than the other, if one is one the earth and the other on the moon. So the heaviness of chemical systems does not supervene on their physical characteristics. Yet presumably we don't want on this account to regard physicalism as refuted by the heaviness of chemical samples. If anything supervenes on physical characteristics, surely heaviness does.

The obvious response to this problem is to note that heaviness is a relational property of chemical samples, depending not only on the intrinsic features of the sample, but also on the features of another system, namely, the surrounding gravitational field. Accordingly, we should modify the requirement of supervenience, for relational properties, so as to demand that such properties should supervene, not on the internal physical characteristics of the system at issue, but rather on those plus the physical characteristics of the relevant related system. If we do this, then the heaviness of chemical samples is no longer a counter-example to physicalism: for the heaviness of a chemical sample obviously does supervene on the internal physics of the sample plus the physics of the surrounding gravitational field. (Equivalently, if less naturally, we could say that the relational properties of a system were not really properties of that system as such, but only of some larger system incorporating the relevant related system, and then require that such relational properties supervene on the physical properties of the larger system.)[2]

Given this qualification about relational properties, I shall take it henceforth that supervenience is a necessary condition for physicalism. But is supervenience sufficient for physicalism? This is a rather more tricky issue. In outline, we can see how supervenience might suffice. Supervenience says that, if two systems are physically identical, then they must also be chemically identical, biologically identical, psychologically identical, and so on. That is, the shared physical features of these systems determine their special features. But how could this be so, if anything non-physical were required for those special features?

Some care is needed, however, to make this line of thought watertight. The issue depends on exactly how we understand supervenience, and in particular on how strongly we read the "determine" in "the shared physical features of these systems determine their special features". In due course we shall see that there is a weak reading of this "determine" on which supervenience clearly does not suffice for physicalism, and a stronger reading on which supervenience does provide a satisfactory characterization of physicalism. But let me not pause for these technicalities at this point. My primary interest in this chapter, as I said, is not with the characterization of physicalism as such, but with the possibility of arguments which support physicalist views. In line with this, it will make more sense for me to fill in the details of what I mean by physicalism once we have seen what arguments are available, rather than before.

Perhaps it will be helpful to be graphic for a moment. The world recognized by physicalism is at bottom a world consisting of physical facts, of particles and fields in motion through space. At this basic level all facts can be described by strictly physical terminology, like "mass", "energy", and "position". However, physicalism, as I am thinking of it, will also allow that we often use non-physical terminology, like "sulphuric acid", "thunderstorm", "elephant", and "thinking of the future", to group and categorize large-scale arrangements of physical facts. Moreover, physicalism allows that such special terminology isn't just a shorthand for complex physical properties: for, in those cases where type identity fails, special categories cannot even in principle be specified in physical terms. Nevertheless, physicalists will say, the instances of any such special kind will still just be complexes of physical stuff. For supervenience, in an appropriately strong sense, implies that nothing more is required for any special kind to be instanced than the physical facts should be thus-and-so. After all, if anything more were required, then presumably it would be possible for the special features of two systems to differ even though they were physically identical, which is just what supervenience rules out.

So far we have been concerned only with what physicalism says. It remains to consider whether we should believe it. In the rest of this chapter I shall argue that physicalism is strongly supported by an important feature of physical science, namely, the internal completeness of physics. However, before proceeding, it will be helpful briefly to consider a number of further preliminary points that may be worrying some readers.

1.3 More Preliminaries

1.3.1 Some of you may feel uneasy about my brisk dismissal of the possibility of type identities between physical categories and special categories. In particular, you may feel that if a special subject matter is scientific enough to contain projectible laws, then it would be surprising if its categories were not type identifiable with physical categories. For why should we expect special categories to conform to any stable regularities, if they are determined by different physical structures on different occasions?

I think this question points to a powerful, though not inescapable, argument for type identity, and shall devote the next chapter to it. In this chapter, however, I shall focus on the prior issue of whether we should accept physicalism, understood in terms of supervenience. Once we have decided this, we can then turn to the further issue of whether we should accept type identity as well. My eventual conclusion will be that type identity holds for some, but not all special sciences: more specifically it holds for those special sciences that lack a teleological underpinning.

1.3.2 In this chapter, and in much of the rest of this book, I shall speak as if our "common sense psychology", which attributes beliefs, desires and other familiar states to people, is a "special science". But this is of course a contentious assumption. Many philosophers view everyday psychology as somehow incommensurable with science, as offering a quite different kind of understanding from science. And other philosophers, while allowing that folk psychology may have pretensions to science, hold that it fails miserably to live up to them.

I intend to by-pass this issue in most of the rest of this book, by stipulating that, unless I say otherwise, my use of folk psychological talk is to be understood as a place-holder for the true special science of psychology. So philosophers who think that folk psychology is already a science can take my words at face value. On the other hand, those who think something different is needed for a genuine cognitive science should simply understand my psychological talk as referring indirectly to their own favoured cognitive states. There remain the pessimists who think that cognitive science of any kind is impossible, that there cannot be a theory of our cognitive workings that stands to our physics and physiology as meteorology, say, stands to the physics of the atmosphere. To these pessimists I simply concede that if their bet about the future of cognitive science is right, then a number of the issues I address in this book do not arise. (Though in fact the issue of this chapter, the relationship between the psychological and the physical3, arises not only for optimists who accept the possibility of a high-level psychological science, but also for those pessimists of a Wittgensteinian or Davidsonian bent who reject this possibility but nevertheless uphold everyday psychology as a respectable but non-scientific form of discourse. For they too need to consider the relationship between psychological states and brain states. It is only pessimists who take the eliminativist line and reject high-level psychological thinking of any kind who can avoid addressing the mind-body problem.)

My own view, for what it is worth, is that everyday psychology constitutes an impressive theory from a scientific point of view, capable of improvement and refinement, of course, and with a number of philosophically puzzling aspects, but certainly containing a great deal of predictive information, and quite probably giving some insight into the structure of our internal workings. I prefer to avoid, however, debates about whether its undoubted imperfections merely mean it is a somewhat inaccurate theory about real entities (like nineteenth-century atomic theory), or whether they make it a false theory about imaginary entities (like the eighteenth-century caloric theory of heat). This issue would be hard enough to resolve if we knew the whole psychological truth (though then it wouldn't matter very much). But, as it is, there are better things to think about.[4]

1.3.3 Among the ways in which psychology is philosophically puzzling is that it deals in propositional attitudes: its explanations invoke beliefs, desires and other states which represent things as being a certain way. (And we can expect the states of any future cognitive science to be similarly representational.) In chapter 3 below I shall address the topic of mental representation. At this stage we need only note that representation complicates the issue of the supervenience of the mental on the physical. For, as a number of writers have observed5, there are many plausible cases of two people having physically identical brains, and yet having propositional attitudes with different representational contents. These examples imply that psychological states individuated by representational content don't supervene on the physics of the brain.

Physicalists about psychology have two options here. They can argue that any such "broad" psychological state is really a kind of relational state, and that therefore, in the way indicated earlier in this chapter, physicalism only requires the supervenience of such states on the physical properties of some larger system which includes the individual's brain as a part. Alternatively, they can argue that such broad states are not really part of serious psychological theory, and therefore that their non-supervenience is not a problem for physicalism about serious psychology. In what follows I shall defend the former line, in particular in sections 1.5 and 1.7 below.

1.4 The Completeness of Physics

Now for the arguments in favour of physicalism. In what follows I shall consider two different such arguments. But both arguments will hinge on what I shall call "the completeness of physics". So in this section let me explain what I mean by this.

I take it that physics, unlike the other special sciences, is complete, in the sense that all physical events are either determined, or have their chances determined, by prior physical events according to physical laws. In other words, we need never look beyond the realm of the physical in order to identify a set of antecedents which fixes the chances of any subsequent physical occurrence. A purely physical specification, plus physical laws, will always suffice to tell us what is physically going to happen, insofar as that can be foretold at all.

Note that not all subject areas are complete in this way. For instance, meteorology is not complete. Some weather phenomena arise from antecedents which are not themselves weather phenomena. The beat of a butterfly's wing, students of chaos tell us, can play a part in determining next week's cyclone. Less exotically, psychology is obviously not complete, given that plenty of mental events result from non-mental ones, as when I sit on a drawing pin and feel a pain. But physics is special in this respect. If we take any physical result, and look back in time to see what gave rise to it, then, I say, prior physical factors will always suffice to give us as full an explanation of that result as is possible.[6]

I have stated the completeness of physics baldly, as something to which all will assent. But perhaps some readers will have doubts. I can imagine two possible sorts of worry. The first would be a general worry that the completeness of physics is an empirical claim and therefore inadmissible in a philosophical argument. I have nothing to say to this beyond the points made earlier in the introduction to this book. The second worry would be more specific: even if empirical claims are admissible in philosophy, is the completeness of physics really a well-supported empirical claim? In particular, what exactly does "physics" mean here? On some perfectly natural ways of reading this term, the completeness of "physics" seems false.[7] However, let me postpone discussion of this second worry to section 1.9 below. For the moment it will be helpful take the completeness of physics at face value and see what would follow if it were true. We will be better placed to evaluate queries about it when we see how it matters to physicalism.[8]

1.5 The Manifestability Argument for Supervenience

Consider now the following argument for the supervenience of psychology on physics.

Premise (1). According to the completeness of physics, the chances of physical consequences are fixed, once physical antecedents are given. So if two systems are physically identical and in the same physical contexts, they will issue in the same physical consequences or chances thereof.

Premise (2). Now add in the assumption, which I shall call the "manifestability of the mental", that if two systems are mentally different, then there must be some physical contexts in which this difference will display itself in differential physical consequences, or at least in differential chances for such consequences.

Conclusion. It follows that mental differences without physical differences are impossible. (1) tells us that physical identity guarantees identity of physical consequences or chances thereof. And (2) tells us that mental difference requires the possibility of different physical consequences or chances thereof. So physical identity rules out mental difference.[9]

The crucial idea here is that the completeness of physics leaves no room for mental differences, or any other differences, to make a difference to physical consequences, once physical antecedents are given. Physical categories by themselves always suffice to fix the chances of physical consequences, without the help of mental categories. So the only way for mental differences to be manifestable is for them to have different physical bases.[10]

The two premises to this argument are the completeness of physics and the manifestability of the mental. As I said, I shall come back to the completeness of physics at the end of the chapter. Here we need to consider the manifestability of the mental. The most obvious argument for this principle would be that mental differences must always be capable of showing themselves in differential behaviour: there certainly seems something initially odd about the idea of two people who are mentally different, yet behave in the same way in all physical contexts. (In this connection, note that the manifestability principle is not the strong requirement that every particular mental difference actually manifests itself in differential physical consequences; just the weaker assumption that, for any type of mental difference, there is some type of physical context in which that difference would be physically manifested.)

If this behavioural interpretation of the manifestability principle were acceptable, then a strong version of the supervenience of the psychological on the physical would follow, namely, the supervenience of psychological states on brain states. For we could run the argument as follows. Mental difference require behavioural differences. But behavioural difference are fixed specifically by prior brain states. So there can't be mental differences without brain state differences.[11]

However, there are good reasons for denying that all psychological states supervene on brain states I am thinking here of the kind of "broad" propositional attitudes mentioned in section 1.3.3. As we saw, the distinguishing characteristic of broad attitudes is that individuals with identical brains can fail to share them. So it follows from the argument in the last paragraph that a manifestability requirement in terms of behavioural displays is too strong a requirement for broad attitudes. And this is of course what we do find: differences in in broad attitude don't automatically display themselves in behavioural differences. To take a familiar example, consider Carl, who wants a glass of H2O, and Lrac, his physically identical Twin Earth counterpart, who wants a glass of XYZ. They have different broad attitudes. But their behaviour, in the sense of the physical movements of their bodies, will be the same in all physical contexts.

As I observed in 1.3.3, the failure of broad attitudes to supervene on brain states does not mean that physicalism is false. For if broad states are relational states, then it will suffice for

physicalism that they supervene on the physics of the individual-system-and-relevantly-related-systems, even if not on the physics of the individual system alone. So it remains possible that the general manifestability argument for supervenience might still establish this weaker kind of supervenience for broad beliefs, even if not supervenience on brain states. All we need is a weaker manifestability premise to the effect that differences in broad beliefs are somewhere manifested in physical consequences, even if they are not manifested in behavioural consequences.

In defence of this weaker version of the manifestability premise, note that a mental difference which was not physically manifestable in any way would be radically undetectable. We know that our sense organs work by physical interaction with the environment, as do the instruments and other aids by which we extend the power of our sense organs. So if two different mental states yielded exactly the same physical manifestations in all contexts, then there would seem no possibility of our ever finding out about their difference. Yet surely any real mental difference ought to be somehow detectable, even if not behaviourally.

To illustrate this point, note that even the broad mental difference between Carl and his identical doppelganger Lrac will be distinguished by some differential physical consequences. For this broad mental difference depends on the relational difference that, where Carl is surrounded by H2O, Lrac is surrounded by XYZ. And this difference in their environments will obviously produce some differential physical consequences by which we can distinguish the two cases.

I recognize that this defence of the manifestability requirement, and hence of supervenience, is less than fully principled. For one thing, it leaves it open for opponents of physicalism to object that it is possible that there be mental differences that are not in any way detectable by human beings. More pressingly, opponents of physicalism could also query whether our ability to detect mental differences always depends on physical interaction with our environments. Thus anti-physicalists might argue that our access to conscious mental states in particular is primarily via introspection, rather than via the normal five senses, and that there is no immediate reason to suppose that the deliverances of introspection are mediated by physical processes, however it may be with the other senses. This would then open the way for anti-physicalists to argue that conscious states might fail to supervene on physical goings-on, and so that conscious differences need not manifest themselves physically, and yet to hold that those differences could still be detectable, via introspection: for example, they could argue that you could be in just the same physical state at two different times, and yet know introspectively that you were in pain at one time and not at the other.

I shall not attempt to plug this particular gap in the manifestability argument, however. For there is a rather more basic flaw in the argument, to which I shall now turn. To deal with this more basic flaw, we will need to switch to a significantly different form of argument for physicalism. Moreover, this alternative form of argument will be immune to the anti-physicalist appeal to non-physical introspective powers.

1.6 Manisfestability is Not Enough

To understand the more basic flaw in the manifestability argument, recall how I earlier alluded to the possibility of different ways of understanding supervenience, depending on how strongly we read the "determine" in "physical features determine special features" (or, equivalently, on how strongly we read the "cannot" in "cannot differ in special features without differing physically").

A weak version of supervenience would understand these notions in terms of natural necessity: that is, it would take physical features to determine special features across all possibilities where we hold the actual laws of nature fixed; equivalently, it would say that any two systems governed by the actual laws of nature cannot be different in any special respects without differing physically.

A strong version of supervenience would do it in terms of "metaphysical" necessity rather than natural necessity: supervenience requires that physical nature determines special nature across all possible worlds whatsoever; no two possible systems of any kind can be different in some special respect without being physically different.

Now only the stronger of these versions of supervenience constitutes a plausible explication of physicalism. To see why, we need only consider epiphenomenalism, the doctrine in the philosophy of mind which holds that mental states "float above" the brain as distinct conscious phenomena, not responsible for any physical effects themselves, but nevertheless causally determined by the physics of the brain, and so incapable of varying without physical variation. Epiphenomenalism implies supervenience in the weak sense, since it implies that, if we hold all natural laws constant (including in particular the putative epiphenomenalist laws by which by physical brain states cause conscious states) then physical nature will determine mental nature: identical physical brain states, plus the laws according to which physical brain states cause conscious states, will ensure identical conscious states. But epiphenomenalism clearly isn't a physicalist doctrine, since it explicitly specifies that conscious mental properties are ontologically distinct from physical ones. So weak supervenience clearly does not suffice for physicalism.

However, the same objection does not apply if we equate physicalism with strong supervenience. For while epiphenomenalism does imply that the mental is fixed by the physical across all natural possibilities, it does not imply that such brain-mind determination holds across all possibilities, including those possible scenarios where these epiphenomenalist brain-mind laws break down. This is because epiphenomenalism insists that conscious properties are ontologically distinct from physical ones. So it implies that it is metaphysically possible, even if not "naturally" possible, for a creature physically just like me, say, to have different conscious states, or indeed to have no conscious states at all. While such a creature would violate epiphenomenalism's putative natural laws of mind-brain causation, and so fail to be "naturally" possible, epiphenomenalism allows that these laws themselves are not absolutely necessary, and so implies that such a creature is metaphysically possible. Conversely, the doctrine that such a creature is not metaphysically possible would be inconsistent with epiphenomenalism's distinction of conscious mind from physical brain, and so would constitute a plausible explication of genuine physicalism.

After all, strong supervenience says that it is metaphysically quite impossible for two beings to differ in some special property unless they differ physically. But how could this be so absolutely impossible, unless the special property was itself in some sense itself physical? If the special property weren't itself physical, then surely there would be metaphysical room, if not natural room, for it to float free of the physical realm, in the way the epiphenomenalist's conscious properties float free of physical properties in worlds with different brain-mind causal laws. So it looks as if strong supervenience that is, the denial of any metaphysical room for special properties to float free of physical ones will indeed ensure that special properties are physical.[12]

To return to the original issue of this section, the basic flaw in the manifestability argument for physicalism is that it only constitutes a good argument for weak supervenience, not for strong supervenience, and so fails to establish physicalism. The fault lies with the manifestability premise, that is, the premise according to which mental differences must manifest themselves in differential physical consequences. For any version of this premise strong enough to deliver genuine physicalism would blatantly beg the question against non-physicalist views like epiphenomenalism.

Consider what epiphenomenalists would say about the manifestability premise. They would happily allow that mental differences will display themselves in differential physical consequences as long as the laws by which brain states cause conscious states are held constant: given these laws, then different conscious states must have been caused by different physical states, and we can expect these physical differences to have different physical consequences. But epiphenomenalists will point out that there is no need to expect this manifestability requirement to hold up across all possible worlds, including worlds where the actual brain-mind laws break down. After all, if we allow, as epiphenomenalists will, that there are metaphysically possible worlds in which I have physical duplicates with different conscious states, or with no conscious states at all, then we will not expect these mental differences, between me and my other-minded physical duplicates, to display themselves in any differential physical consequences.

So epiphenomenalist anti-physicalists will see no reason to concede that the manifestability premise holds across all metaphysically possible worlds, even if it holds in all naturally possible worlds. And correspondingly, they will not view the manifestability argument as providing any substantial reason to suppose that the mental supervenes on the physical across all metaphysically possible worlds. They can allow that mental differences will display themselves in differential physical consequences as long as we hold all laws of nature fixed and correspondingly concede the weak supervenience thesis that mental differences without physical differences are naturally impossible. However, this, as we have seen, falls short of physicalism proper. Genuine physicalism requires strong supervenience mental differences without physical differences are metaphysically impossible. But epiphenomenalists will see nothing in the manifestability argument to force them to this stronger claim, for they will have no inclination to accept that all metaphysically possible mental differences must display themselves in differential physical consequence.[13]

1.7 The Causal Argument for Physicalism

Let me now turn to a somewhat different argument for physicalism, which I shall call "the causal argument". This argument, like the manifestability argument, will hinge on the completeness of physics. But instead of appealing to requirements on the manifestation of mental states, it will appeal to the possession of causal powers by mental states. This shift of focus will yield a more effective line of reasoning against anti-physicalist view like epiphenomenalism.

Thus consider the following premise, which I shall call the "principle of mental efficacy":

Premise (3). Every mental occurrence causes some physical effect.

Note now that, on just about any account of causation[14], the following is an immediate corollary of the completeness of physics:

Premise (4). All physical effects have complete physical causes ("complete" in the sense that those causes on their own suffice by physical law to fix the chances of those effect).

Consider now some mental occurrence, and one of the physical effects which are required by (3). For example, suppose you decide to lift your arm, and as a result your arm rises[15]. By (4) this physical effect will also have a complete physical cause, which will presumably involve the neuronal and other physical antecedents of your arm rising. So it follows that your arm rising has two causes: a mental cause, your decision, and also a physical cause, your neurones firing.

Does this mean that such physical effects are always overdetermined, like the death of the man who is shot and simultaneously struck by a random bolt of lightning? This doesn't seem right. After all, when an effect is overdetermined by two causes, it follows that it would still have occurred if either one of the causes had been absent: the man would still have been killed by the lightning bolt even if he hadn't been shot, and vice versa. But we don't similarly want to say that your arm would still have gone up even if you hadn't wanted to lift it, or, alternatively, even if different neurones had fired in your brain.

The obvious conclusion is that your desire and your neurones are not two independent causes, like the shot and the lightning bolt, but are in some sense the same cause. We need somehow to identify the mental cause with the physical cause, so as to avoid the conclusion that the movement of your arm was overdetermined16.

Note how this argument differs from that in the last section. There the aim was to show that the physical always co-varies with the mental, and the argument was that physical variation is needed to produce the external evidence for mental variation; the trouble was that this argument only established co-variation across naturally possible worlds, which was too weak for physicalism. In this section the aim has been to show that the mental is ontologically inseparable from the physical, and the argument has been that such a separation would imply an absurd proliferation of causal overdetermination; if this ontological inseparability does follow, it will mean that there is no metaphysical room for mental properties to float free of physical ones, and so will establish genuine physicalism.

It might seem as if the causal argument begs the question against anti-physicalist epiphenomenalism just as much as the manifestability argument. Epiphenomenalists, after all, will deny the assumption of causal efficacy, just as they denied any strong manifestability premise. So they will escape the causal argument too. They don't need to explain why bodily movements aren't always overdetermined, since they don't admit they have mental causes in the first place.

But there is a significant dialectical difference between the two cases. There is nothing pre-theoretically objectionable about the denial of a strong manifestability premise: nothing obvious will go wrong with our overall view of the world just because we allow the mere metaphysical possibility of mental differences without physical manifestations. By contrast, it clearly flies in the face of any number of normal assumptions to deny that mental events have physical effects. If my conscious thirst isn't what causes me to go to the fridge for a beer, and my conscious map-reading isn't what causes me to choose one route rather than another in a strange city, and so on, then we are going to have to think again about most of our assumptions about the way the human world works.

Given this, we can well ask why the epiphenomenalist wants to adopt the curious view that conscious mental states are causally inefficacious, especially given the availability of physicalist alternatives which avoid it. The only plausible answer, I take it, is to do with consciousness: epiphenomenalists are persuaded that any physicalist account of the mental will leave out the essential conscious features of the mental, and so are persuaded to postulate a distinct, non-physical realm of mental events, even at the cost of denying that the mental affects the physical. I shall return to this issue in chapter 4 below, where I shall argue that there is nothing in consciousness that is left out by physicalism, and therefore that the epiphenomenalist denial of the causal efficacy of the mental is ill-motivated.

Before proceeding, let me quickly deal with one complication. This relates to broad mental states. We saw earlier how broad mental states complicated the manifestability argument. Similar complexities arise in connection with the causal argument.

Thus note how I illustrated the causal argument by focusing on the bodily effects of mental states, like arms rising, and then inferred that the mental causes of these bodily effects must be identical with their "neuronal causes". However, this specifically neuronal conclusion sits ill with the possibility of broad mental states. For broad mental states can't be identical with internal brain states, given that they depend on matters outside the skin. Carl and Lrac differ in their respective desires for H2O and XYZ, even though they are internally physically identical. In line with this, it seems wrong to say that their different desires cause their bodily movements: bodily movements are surely caused by matters inside the skin, not by features that stretch outside.

Still, the fact that broad mental causes can't be the same as brain states doesn't mean they can't be equated with any physical states, in particular with certain physical features of their possessors-and-relevantly-related-systems. And it is not hard to see how the causal argument might be made to deliver this weaker conclusion. All we need is a causal efficacy premise to the effect that broad mental states cause some "broad" physical consequences, even if they don't cause the bodily movements that result from neuronal causes alone. And there seems no difficulty about this version of the efficacy premise. For example, Carl's desire may cause a glass of H2O to move, where Lrac's desire will cause a glass of XYZ to move. And then, with the efficacy premise so restored, we can use the causal argument to argue that Carl and Lrac's desires must be equated with those physical features of themselves-and-their-surrounding-environments which are responsible for these broad effects.

1.8 Generous Causation and Alternatives to Type Identity

I argued in the last section that the mental causes of physical effects must be the same as the physical causes of those effects. Exactly how we construe this equivalence, however, depends on what view we take of the ontological status of causes in general. Some philosophers, most prominently Donald Davidson (1967), think that causation is a relation between events construed as "bare particulars" shorn of any general attributes. However, there are good arguments for being dissatisfied with this anaemic view of causation, and for preferring to view causal relata as facts rather than as Davidsonian bare particulars17. Accordingly, I shall assume the factual view of causal relata in what follows.

However, if you view causation as a relation between facts in this way, then it may seem as if the causal argument is in danger of proving too much. In particular, it may seem in danger of proving that mental properties must be type identical with physical properties,

notwithstanding the intrinsic implausibility of this type identity claim. For, if causes are facts, then the causal argument's conclusion, that mental causes must be identical with physical causes, will require that mental facts such as that I am in pain, say are identical to certain physical facts have a certain brain feature, say and these two facts cannot be identical unless the properties they involve being in pain, having that brain feature are themselves identical.

Well, this type identity would indeed follow from the causal argument if we take a very strict view of causation, and insist that the only thing that can cause a physical effect is another strictly physical fact. For then the principle of mental efficacy, according to which mental facts cause physical effects, can only be satisfied if mental facts are themselves instantiations of strictly physical properties. However, suppose we understand causation in a more generous sense, and allow that an instance of a strongly supervening property causes the effects of those facts on which it supervenes. Then the principle of mental efficacy will only require that mental properties are type identical to physical properties or that they strongly supervene on physical properties. For as long as the latter possibility is realized, then it will still be true, in the generous sense, that mental facts cause the physical effects of the physical facts on which they supervene.

As an illustration of this possibility, consider the functionalist view that mental states are causal intermediaries between perceptual inputs and behavioural outputs. The orthodox version of this view does not identify pain, say, with whichever first-order property mediates causally between damage detectors and avoidance behaviour in any given species. For this would have the "chauvinist" implication that species with different internal workings could not share the experience of pain. Rather the standard functionalist view is that pain is a second-order property, the property-of-having-some-property which mediates causally between damage detection and avoidance behaviour, which second-order property can therefore be present across beings with different innards.

Now, on this functionalist view, pains can't cause bodily movements in the strict sense which requires identity with strictly physical facts. For, if pains are instantiations of second-order properties, they cannot be identical with any first-order physical facts. Still, such functionally understood pains can still be "realized" by physical properties, in that they can be present purely because some first-order physical fact which mediates between damage and avoidance is present and in that case a pain will indeed cause bodily movements, in the generous sense in which superveners cause what their subveners cause. For if a pain is so realized by a physical fact, then it will supervene on this physical fact, even though not identical with it, in that any metaphysically possible being with this physical property will be in pain, since it will possess the property-of-having-some-property which mediates causally between damage and avoidance.

Let me clarify my direction of argument here. I am not at the moment concerned to uphold functionalism, nor, consequently, am I particularly concerned to argue that functionalist mental definitions are satisfied by physical states in humans18. Rather I have introduced functionalism merely as an illustration of how facts that are not themselves physical facts can nevertheless cause physical effects, at least in the generous sense of causation.

More generally, if we understand the causal argument in terms of the generous sense of causation, then the conclusion will be mental facts must in some way strongly supervene on physical facts (otherwise mental facts couldn't cause physical facts even in the generous sense, given the completeness of physics). Functionalism offers one illustration of how this

might be so, even when type identities are not available. But my conclusion is not that functionalism must be true, only that the mental must somehow strongly supervene on the physical.

For a further example of a theory of this form, consider Donald Davidson's view of the mental. (Davidson, 1980, passim. Though Davidson's view of the mental is standardly presented in harness with the Davidsonian view of causation mentioned above, it is helpful to separate out these two aspects of Davidson's thinking.) Davidson holds, in effect, that to be in a given mental state M is to be in some state which causes behaviour which would warrant the attribution of M to you. This is a different theory from functionalism, since it makes essential appeal to the non-scientific canons of interpretation which Davidson takes to govern our attributions of mental states to others. But, just like functionalism, it allows room for the idea that the mental may be realized by the physical, and consequently strongly supervene on it. For if it is physical state P which causes the behaviour which warrants the attribution of mental state M to person X, then X will be M purely in virtue of being in P, and correspondingly any possible creature with P will have M, since it will have some state which causes behaviour which would warrant the attribution of M.

So the Davidsonian view, like functionalism, will also satisfy the requirement that mental facts should cause physical facts, at least in the generous sense. Still, as with functionalism, I mention this, not as an argument for the Davidsonian view in particular, but simply as another illustration of how the requirement of supervenience on physical states allows the causation of physical effects, even in the absence of type identity.

I have no clear views about the full range of ways in which mental properties might supervene on physical ones. Functionalism and Davidsonianism are two such options, but there may well be others. However, there is no need to decide this issue here. It will be enough if I have shown that some such view of the mental is demanded by the causal argument.

Of course, there remains the option of embracing epiphenomenalism, and denying that the mental is efficacious, even in a generous sense. As I observed earlier, the normal motivation for this unpalatable view is to do with consciousness, and the conviction that conscious states at least must be ontologically quite distinct from any physical states. The question this raises, and to which I said I shall return in chapter 4, is whether this anti-physicalist conviction about consciousness rests on solid enough grounds to justify the radical step of denying that our thoughts and feelings affect our actions. But for the moment I am content merely to point out that the minimum price for rejecting physicalism is epiphenomenalism. Anti-physicalists need to deny some premise in the causal argument, and the line of least resistance is to deny the principle of mental efficacy.

1.9 The Completeness of Physics Defended

There is an alternative, if less obvious, way to resist the causal argument namely, by denying the completeness of physics. This assumption may seem initially plausible. But, as I allowed earlier, it is by no means entirely unproblematic.

The central difficulty facing defenders of this assumption is an obvious dilemma about what they mean by "physics". Either "physics" means the theory currently taught in university departments of physics and presupposed by articles in physics journals, or it means some ideal future theory that will succeed current theory.

The trouble with the first horn of this dilemma is that, if the past form of physical theorizing is anything to go by, current physics is no doubt inadequate in certain respects, and in particular in failing to identify all the antecedents for certain physical effects. So current physics is not complete.

The trouble with the other horn, by contrast, is that we don't yet know what physical categories will be assumed by the ideal future physics. So we scarcely seem to be in any position to maintain that those categories will suffice for complete explanations of all physical effects.

However, I think there is a version of the second horn of this dilemma which will serve the purposes of the arguments of this chapter. Suppose we simply define "physics" as the science of whatever categories are needed to give full explanations for all physical effects. I accept, as above, that this science will be different from current physical theory, and thus that we don't yet know what it is. But, even so, there is no difficulty about how we know that it is complete, for we have simply defined it so as to be complete.

The obvious worry about this definitional strategy is that it seems to remove any significant content from the thesis of completeness, and thereby to make it doubtful that the thesis could have any substantial conclusions. There are two dimensions to this worry. First, the definitional strategy characterizes physics as the science of whatever is needed to explain "physical" effects. But what are "physical effects", if we haven't yet specified what counts as "physics"? Second, even if we had some independent hold on "physical effects", the proposed strategy would still make the completeness thesis an empty analyticity, for it simply defines "physical" categories as all those needed to explain physical effects, from which completeness immediately follows.

Let me deal with these two worries in turn. To deal with the first worry, I propose that we simply postulate some pre-theoretically given class of paradigmatic physical effects, such as stones falling, the matter in our arms moving, and so on. If we take this class to be independently given, then we can effectively characterize the rest of physics as all the categories that need to be brought in to explain those paradigmatic physical effects.

But this still leaves us with the second worry, that even we help ourselves to a pre-theoretical class of paradigmatic physical effects, we are still defining physics in such a way as to make the completeness of physics a matter of definition. I still need to explain how substantial conclusions about the truth of physicalism could possibly follow from such a definition19.

My answer is that no substantial conclusions follow from the completeness of physics per se. But they do follow from the joint assumption that (a) physics is complete and (b) that it does not make any use of psychological categories.

Let me explain. In itself, the above definition of physics leaves it open that psychological categories may turn out to be needed as an essential part of physics. Maybe psychokinesis is true, and there are physical effects that can't be accounted for without making essential mention of distant volitions. Less exotically, maybe some bits of behaviour can't in fact fully be accounted for purely in terms of muscular activation, neuronal activity, and so, without bringing in extra mention of prior mental states. Now, if psychological categories do turn out to be needed to give full explanations for physical effects in this way, then the issue of

whether psychology supervenes on the physical, as I have defined it, becomes trivial. Psychology will indeed supervene on the physical, but only because it is included in the physical, not because psychological variation requires variation in something else.

On the other hand, if psychology is not part of the physical, as I have defined it, then the arguments of this chapter will go through as before. That is, if psychological categories are not in fact ever essential to explaining physical effects, then physics, in the sense of whatever is needed to explain physical effects, will be both complete and exclusive of psychology, and the arguments of this chapter will show that psychological states are non-trivially supervenient on physical states.

It seems to me highly unlikely that the psychological will turn out to be part of the physical. Current physics, I take it, aims to develop a complete theory of paradigm physical effects in terms of the categories of energy, field and spacetime structure. I am quite prepared to believe that this this aim cannot be achieved, and that the categories of current physics will need supplementation before we can get a genuinely complete theory. What I do not believe is that they will need supplementation by psychological categories.

I am here making an empirical claim. The history of science yields a great deal of empirical evidence about the kind of causes that are responsible for the motion of stones and other kinds of matter. This evidence does not, perhaps, allow us to formulate a definitive list of all the necessary categories. But it does, it seems to me, provide sufficient grounds for concluding that mental categories are not among them.[20]

To help see what is at issue here, it is illuminating to consider Descartes' views on the matter. Descartes did think that there were physical effects that could not be explained without bringing in mental antecedents. Descartes believed that the total amount of motion, in the sense of mass times speed, is conserved, according to regular laws, in all material interactions, and therefore that the speeds of all material bodies are determined by earlier such speeds. However, unlike us, Descartes did not believe in the conservation of momentum, considered as a directional quantity, and so did not think that the direction of motion of material bodies was necessarily determined by prior physical factors. And it was this gap that Descartes exploited to explain how the mental, although ontologically quite distinct from the physical, can nevertheless affect the physical: the mental interacts with the physical in the pineal gland, and influences the direction of motion of certain particles (though not their speed, since this is always fixed by prior physical states).

To hold that the psychological is part of the physical is to believe a version of what Descartes believed, namely, that there is a gap in the determination of certain physical effects, which can only be filled by mental occurrences. And this is what seems highly unlikely to me. It is one thing to hold that the current categories of energy, field and spacetime structure leave a gap in the determination of certain physical effects. It is another to hold that this gap cannot be filled without bringing in the mental. If that were true, after all, then the obvious moral would be that physicists needn't build expensive particle accelerators to generate theoretically anomalous physical phenomena; instead they could find plenty of currently inexplicable physical phenomena simply by looking inside people's heads.[21] I think we have good empirical reason to reject this possibility as absurd.

Extra References

Papineau, D. 1996: "Theory-Dependent Terms" Philosophy of Science 63.

[Because of this new reference, remove Stich (1991) from the Bibliography.]

Steward, H. 1996. "Comments on Philosophical Naturalism" Philosophy and Phenomenological Research 57.

Witmer, G. 1998. "What is Wrong with the Manifestability Argument for Supervenience" Australasian Journal of Philosophy 76.

1. For arguments in favour of physicalism, see Lewis (1966), Davidson (1970), Peacocke (1979, ch III.3), McGinn (1982, p 29), Smith and Jones (1986, pp 57-59), McFetridge (1990, p 86), Lycan (1987, pp 2-3). Reasoned opposition to physicalism is offered in Crane and Mellor (1990), Crane (1991). Most of these contributions will be referred to further in what follows.

2. It is sometimes suggested that this kind of shift, from a "local" to a more "global" supervenience, makes room for ad hoc defences of supervenience, and so dilutes physicalism beyond interest. To answer this charge, we should require that wider systems should be admitted as subvening bases only if there are independent grounds, apart from a desire to save physicalism, for regarding the putatively supervening properties as relational.

3. For brevity I shall often focus in this chapter on the relation between the psychological and the physical. But the analysis will be of general significance, as the structure of my arguments will indicate.

4. I think that there is good reason to think that theoretical concepts in general, and psychological concepts in particular, are vague, in that there will often be no fact of the matter about how to apply them to cases where the theory that defines them breaks down. For more on this see Papineau (1996).

5. See Putnam (1975), Burge (1979, 1982) and Evans (1982).

6. Some readers might baulk at my use of "explanation" here, on the grounds that a full physical specification of the antecedents of some large-scale physical outcome won't necessarily be illuminating for us humans, in the way that an explanation using chemical or biological or psychological terminology might be (cf Putnam, 1978, p 42). No matter. My argument only requires that the physical antecedents fix or cause the physical outcome, not that they illuminate it. David Owens (1992) is even more particular, and would baulk at this last use of "cause", on the grounds that causes aren't causes unless they illuminate. Again no matter. My arguments need only whatever is left in the notion of cause after we take away the anthropocentric factor of illumination.

7. Cf Crane (1991).

8. As it happens, when I do return to the completeness of physics in 1.9, I suggest that this thesis itself can most usefully be understood as an analytic truth, rather than as an empirical claim. However, when we do read it in this way, the burden of my argument for physicalism is then taken up by some closely related empirical assumptions.

9. This argument is found in McGinn (1982, p 29) and further discussed by McFetridge (1990, p 86). In Papineau (1990) I tried to run the argument with a weaker version of premise (2), requiring only that mental differences have some different consequences, not necessarily physical ones. But when I presented this version of the argument at the Analysis 50 Conference in Cambridge, Tomis Kapitan showed me that it begged some crucial questions.

10. Why doesn't the argument work in reverse, and also show that all physical differences must depend on mental differences? The essential reason is that the mental is not complete. Even if we accept, as is not entirely implausible, the "mental manifestability of the physical" ("if two systems are physically different, there must be contexts in which this will produce differential mental effects"), we cannot conclude that these differential mental effects must always depend on prior mental differences, for lack of the premise that mental effects are always fixed by mental antecedents.

11. This is the version of the argument articulated by McGinn. He does, however, observe that it may not apply to all mental states.

12. Some readers may be wondering whether this equation of physicalism with strong supervenience has not simply taken us a long way round back to the earlier equation of physicalism with type identity. For haven't I just argued that the virtue of strong supervenience is precisely that it ensures that special properties are ontologically inseparable from physical properties, by contrast with weak supervenience, which only requires that special properties are correlated with physical properties by the actual laws of nature, but need not be ontologically intertwined with them? Well, the virtue of strong supervenience is indeed that it ensures an ontological dependence of special properties on physical properties, and not just a correlation. But the point of formulating physicalism in terms of supervenience, rather than type identity, is precisely that it is possible to have such ontological dependence even when type identities are not available. I shall return to this in section 1.8 below.

13. For further discussion of the failings of the manifestability argument, and for other criticisms of the original English version of this chapter, see Steward (1996) and Witmer (1998).

14. David Owens (1992) is an exception. But, as I said in footnote 6, I could grant Owens his stronger notion of cause, and simply phrase my arguments in terms of a weaker one.

15. Are bodily movements, like arms raising, mouths moving, and so on, properly counted as physical effects? Strictly, no. "Arm" and "mouth" are biological terms, not physical ones, and it is doubtful that they can be reduced to physical notions. So for full accuracy we ought to take the physical effects of mental causes to be the motion of bits of matter, which happen to be in arms, mouths, and so on. However, it will smooth the exposition if I can be less than strict on this point.

16. This form of argument for token congruence is to be found in Peacocke (1979, ch III.3). It has obvious affinities with the discussion in Davidson (1970).

17. For a defence of this view. see Mellor (1987). Another alternative to Davidson's view of causation is to allow that causes are events, but insist that events are instantiations of

properties, rather than Davidsonian bare particulars (Kim, 1973). However, Mellor (op. cit.) argues that "events" of this kind are simply a subspecies of facts.

18. David Lewis (1966) does argue from a version of functionalism to mind-brain identity. Lewis's argument shares one premise with my causal argument, namely, the completeness of physics. But where my other premise is only that each particular mental cause has some physical effect, Lewis makes the stronger functionalist assumption that different mental types can distinguished by their characteristic causal role in mediating between physical causes and effects. (He then concludes, from the completeness of physics, that such roles are always filled by physical states.)

19. Crane (1991) argues on just these grounds that the version of my argument for physicalism in (Papineau, 1990) collapses into triviality.

20. Let me guard against one possible source of confusion here. When I say that a compete physics excludes psychology, and that psychological antecedents are therefore never needed to explain physical effects, the emphasis here is on "needed". I am quite happy to allow that psychological categories can be used to explain physical effects, as when I tell you that my arm rose because I wanted to lift it. My claim is only that in all such cases an alternative specification of a sufficient antecedent, which does not mention psychological categories, will also be available. I need the thesis that psychological terms are not included in the minimal set which provides sufficient conditions for all physical effects, not that they are not included in any such set.

21. Cf. Lycan (1987, pp 2-3).

CHAPTER 2 REDUCTION AND SELECTION

2.1 Introduction

In the last chapter I upheld the orthodox view that reducibility to physics, in the sense of type identity, is too strong a requirement for the categories of most special sciences, and for those of psychology in particular. In this chapter, however, I want to show that the case for reducibility to physics is rather stronger than is generally recognized. More specifically, I want to show that this case is compelling for those special sciences that do not have a teleological underpinning. As it happens, I think that psychology in particular does have such a teleological underpinning, and that its categories are therefore not reducible to physical categories. But my argument will imply that special sciences without a teleological underpinning are indeed reducible to physics.

2.2 An Unexplained Coincidence

It will be convenient to begin with the functionalist view of mental states. As I explained in the last chapter, functionalists view mental states as causal intermediaries between perceptual inputs and behavioural outputs. According to functionalism, you will be in a given mental state as long as you are in a physical[1] state which plays the relevant causal role between perception and behaviour.

It is widely regarded as a great merit in functionalism that it leaves room for irreducibility, and allows that mental states should have different physical realizations in different people, or even in the same person at different times. According to functionalism, what is common between John Major and Boris Yeltsin, when they each believe that there is an ice-cream in front of them, say, is that they are each in some physical state which is characteristically caused by the presence of an ice-cream, and which characteristically causes them to reach out if they want an ice-cream. But functionalism doesn't require that this be the same physical state in both cases -- which is just as well, given how unlikely it is that there should be some strictly physical feature common to all and only those people who believe that there is an ice-cream in front of them.

However, there is something rather puzzling about the picture that functionalism now invites us to accept. If states like believing there is an ice-cream in front of you, and wanting that ice-cream, are realized by different physical states in different people, then why do these states always have the same behavioural effect in all those different people, namely, reaching out for the ice-cream? In general, we expect physically similar states to have similar effects, and physically different states to have different effects. So why in this case should physically different states have the same effect?

Consider an analogy. Imagine that people forced to eat a certain restricted diet -- nothing but reheated brussels sprouts, say -- invariably develop certain characteristic symptoms -- inflamed ankles and knees, say. Nutritionists investigate this phenomenon. But they find no uniform explanation. In one case, the sprouts harbour a virus which flourishes in the ankles and knees and provokes the immune system. In another case, eating the sprouts leads to excess production of uric acid and hence to gouty attacks. In another, the diet leads to a nutritional deficiency which depletes the cartilage which protects the joints. And so on. For each person, there is some physiological explanation of why the diet leads to the inflammation, but the explanation is different in each case. I take it that this would be incredible. If the diet triggered either of just two different sequences, say, both of which then happened to cause inflammation in the ankles and knees, we could perhaps view this as a

curious coincidence, an amusing oddity with no further explanation. But that it should trigger an indefinite number of quite different sequences, yet all of them lead to the same inflammation, would surely be quite absurd, in the absence of further explanation.

Yet the thesis of variable realizability seems to commit us to something quite analogous, namely, that the same perceptual inputs give rise to quite different internal states in different people, and yet those different internal states will all end up generating the same behavioural outputs. This too is surely quite absurd, in the absence of further explanation.

Contrast the functionalist[2] picture with the kind of situation where physical reduction is possible, as when kinetic theory reduces the classical gas laws to the basic dynamics of molecular movement. At first sight it mightn't be clear how this kind of case differs from the functionalist picture. After all, aren't there lots of different ways in which the molecules can be moving around in a gas at a given temperature, thus giving us a heterogeneity of physical states for the single macro-state of having that temperature?

But, even so, there is still something physically in common between all those different physical states, namely, that the molecules have a given mean kinetic energy. It is this commonality that then enables us to explain such things as why an increase in temperature at constant volume always results in an increase in pressure.

Reducibility to physics does not involve the absurdly strong requirement that the instances of the reduced category should share all their physical properties. The requirement is only that there should be some physical property present in all and only those instances, which then allows a uniform physical explanation of why those instances always give rise to a certain sort of result.

But that is precisely what we don't have in the functionalist case. If there is nothing physically in common among the realizations of a given mental state, then there is no possibility of any uniform explanation of why they all give rise to a common physical result. And that's what I find puzzling.

Imagine that the temperatures and pressures of gases were always realized by internal molecular motions, and temperature increases always led to pressure increases, but yet it was impossible to explain this in terms of basic physics. I take it that this would be incredible. But that's what functionalism is asking us to believe about psychology.

It is worth emphasizing that I am not accusing the functionalist picture of inconsistency, but only of incredibility. The difficulty I am concerned with arises when some mental state S, which mediates between physical input R and physical output T, is realized by a range of different physical states Pi. The puzzle is: why do all the different Pis which result from R all nevertheless yield the common effect T? Now, it is possible that every such Pi should just happen to yield T, just as it is possible that all the different physical consequences of eating reheated brussels sprouts should just happen to cause inflamed ankles and knees. However, if this were so, it would be the kind of coincidence that cries out for explanation.

At bottom, the difficulty I am raising is an empirical one. Our experience of the world has shown us that if a certain physical result always appears after certain physically specified antecedents, then there is always some uniform explanation in terms of physical laws. But the functionalist picture violates this general principle. It commits us to the existence of a

physical generalization, namely, that R leads to T, but denies that it can be explained in physical terms. I think this ought to make us think again about functionalism. (For other versions of this argument, see Papineau, 1985; Searle, 1985, ch 5; MacDonald, 1986, sect II.2.)

2.3 Laws in the Special Sciences

Although the last section focused on functionalism in the philosophy of mind, the problem at issue is clearly generalizable to any category in any special science which (a) is related by empirical law to physical antecedents and physical consequents yet (b) is variably realized at the physical level. For in any such case we face the same puzzle of why all the different physical realizations of the special category should give rise to the same physical result.

One obvious way of resolving this puzzle would be to deny (a), that is, to deny that the special category in question is related by law to physical antecedents and consequents. For there obviously won't be any puzzle about how the different physical realizations of some special S all produce the same physical result in appropriate circumstances, if there isn't any such result that they all produce.

It do not propose to adjudicate how far this move is plausible for the different special sciences. The question of the existence of psychological, social and biological laws is a standard topic in the philosophy of these special sciences, and there is no question of engaging with the huge literature on these questions here. But what I shall show in this section is that, if you do want to resist reductionism by denying the existence of laws, then your denial needs to be whole-hearted. It is not enough merely to maintain that the laws of biology, say, are less strict than those in basic physics. For even lax biological laws will be puzzling, if there are no reductive relations between biological categories and physical ones.

Let me illustrate this point by considering the position adopted by Jerry Fodor in his influential article "Special Sciences" (1974). In that article, Fodor defends the general functionalist picture I have been concerned with so far: he takes it that any special S will be realized on different occasions by different physical Pis, but that nevertheless such special Ss can enter into laws linking them with subsequent results R, in virtue of the fact that processes operating on the physical level will generally lead from each Pi to R.

However, Fodor adds a twist, which might seem to avoid the difficulty I have raised for functionalism: Fodor insists that the law linking S with result R will have exceptions. On Fodor's picture, not all the Pis which realize S will give rise to result R, and in consequence the S-R law will not be invariable. So Fodor seems to have an immediate answer to the question of why all the different realizations of S yield the common result R -- they don't.

However, note that Fodor continues to hold that S usually leads to R, or tends to lead to R, or some such. And this in itself raises a puzzle, in the absence of any concessions to the reducibility of S. For if there is no common physical pattern at all to the realizations of S, then it will be puzzling that there is even a tendency for S to lead to R.

By denying that laws involving special categories are exact, Fodor can resist the argument for an exact reduction. But he still faces an argument for an approximate reduction, as long as his special science contains approximate laws. For even such approximate laws will be puzzling, unless there is some common physical category which usually realizes S, and thereby explains why S is usually followed by R.

There is, I think, a general pattern here. By weakening the extent to which a special science contains general truths, you can weaken the extent to which its categories have to be reducible to physics. But you can only avoid reductionist conclusions completely by denying that the special categories enter into general lawlike patterns at all.

By way of further illustration, consider a suggestion made by Davidson. In general Davidson is sceptical as to whether any serious laws can be framed using psychological terminology. However, in "Hempel on Explaining Action" (1976) he offers the suggestion that the generalizations involved in explaining and predicting actions are always person-specific. We can know that Jim, say, will buy an ice-cream when he wants one, will lose his temper if he thinks someone has been rude to him, and so on. But at the same time there are other people of whom these things aren't true. And so, even if we can know a law which applies to Jim, this doesn't mean that there are psychological laws ranging over people in general.

My immediate concern here is not to evaluate this ingenious suggestion, but just to point out that the reductionist argument still gets some grip on even this minimalist Davidsonian conception of psychology as a generalizing science. Davidson still holds that it is a general truth that if Jim want s an ice-cream, he buys one. And this itself would be mysterious unless there is at least a uniform physical realization of Jim wanting an ice-cream. (In fairness to Davidson, it should be noted that he takes the relevant generalizations to be dispositional as well as person-specific. This raises further issues which I discuss in the next section.)

It is true that a uniform physical realization for Jim wanting an ice-cream doesn't amount to a very strong form of reduction. But it's still something, given that even Jim wanting an ice-cream can in principle be realized by different physical states in different instances. Once more, the moral is that, insofar as generalizations of any kind are admitted in a special science, to that extent there will be an argument for a corresponding amount of reduction.[3]

2.4 Functionalism and Dispositions

There is one obvious way in which S could be variably realized and yet there be no mystery about a general law linking S and R. Namely, if S were a dispositional term, defined as the state of being disposed to give rise to R in appropriate circumstances. In that case there wouldn't be any further need to explain, via some physical reduction of S, why the different realizations of S all give rise to R: giving rise to R is precisely what makes those different states all count as realizations of S in the first place.

It is possible that this thought has tended to obscure the difficulty that I have been raising for functionalism. Functionalism makes it a matter of definition that any given mental state gives rise to specified behavioural effects. And so, if you focus on this aspect of functionalism, it may seem natural to conclude that there can't be any real difficulty about how such states can be differently realized physically, and yet have common physical effects. Isn't this just an upshot of their functionalist definitions?

However my puzzle can't be dissolved that easily. The basic difficulty is that functionalist concepts aren't so much dispositional concepts as theoretical ones. Functionalism defines special categories not just as states which produce certain effects, but rather as states which enter into a certain structure of causes and effects. According to the functionalist picture of psychology, for example, pain will be defined not just as the state that characteristically causes avoidance behaviour, but also a s the state that characteristically results from bodily

damage; again, the belief that there is an ice-cream in front of me will be defined not just as the state that characteristically causes me to reach out if I want an ice-cream, but is also the state that characteristically results from my looking at an ice-cream with my eyes open. In general, functionalist definitions of Ss allude to their resulting from Rs, as well as to their causing Ts.

This means that the argument against S being variably realized now goes through as before. We can put it like this: if the various realizations of the state which arise from R have nothing physically in common, then how come they all alike give rise to T? If the various realization s of the state which arises when people look at an ice-cream have nothing in common, how come they all alike lead to their reaching out for it?

The presence of R as an independent criterion for the presence of an S, independent of S's effects, means that we can't any longer simply account for S's realizations all yielding T by saying that's what makes them realizations of S. For now something else also makes them realizations of S, namely, that they arise from R.

It is worth emphasizing that the position which I am claiming faces a difficulty is not functionalism understood merely as the claim that psychological states can be defined in terms of structures of causes and effects. Rather, the difficulty arises specifically when functionalism in this sense is combined with the thesis that psychological states are variably realized. So I have no argument against the philosopher who insists that our concept of pain is a concept of a second-order state defined in terms of certain perceptual causes and behavioural effects, and not a concept of any specific physical state. My concern is only to point out that, unless there is a specific physical state which generally realizes pain, albeit an unknown one, it would be a puzzle why those perceptual causes are generally followed by those behavioural effects.

A connected point. I am exploring an argument for the conclusion that the categories of the special sciences must be reducible to physics. This is not an argument, however, for the conclusion that the practitioners of the special sciences have to know that reduction. For, even if such a reduction is in principle available, you don't necessarily have to know it to have an adequate conceptual grasp of the relevant special categories, and hence be in a position empirically to investigate them. After all, the classical gas laws were well known long before kinetic theory was developed. (This point would be too obvious to be worth making, were it not for my suspicion that many people are attracted to "functionalism", in the strong sense of variable realizability, because they think that without it the special sciences would be under an unfortunate immediate obligation to produce physical reductions of their categories. But of course there are plenty of other possible justifications for not producing immediate reductions, apart from functionalism in this strong sense.)

David Lewis (1980) combines a functionalist definition of psychological concepts with the view that those states are uniformly realized in any given species. Up to a point this position circumvents the difficulty I am raising. Within any given species, so to speak, there is no puzzle as to why the state that arises from R always gives rise to T, for by Lewis's own account that state will be a homogeneous physical state which will lead to T as a matter of physical law. However, there does remain a problem across species. If the central states that result from bodily damage in octopuses and frogs and humans are all so different, how come they all lead to movement away from the external cause of the pain? Of course, behavioural or neurophysiological observation of each such species could show us that the various central

states in question all give rise to such avoidance behaviour, that is, could show us that all these species had states that fitted the general functionalist definition of pain. But, without further explanation, there would still be a puzzle as to why, despite their physical differences, the different central states that arise from bodily damage should have the same physical effects. (By now a solution to this puzzle will no doubt be suggesting itself. Namely, that these states all have the same effect because they have all been naturally selected to produce that effect. But let us leave this solution until section 2.7. My current concern is only to establish that there is a puzzle here to be solved.)

2.5 The Irreducibility of Ordinary Dispositions

The argument of the last section showed that there is no reason to suppose that dispositional concepts in general will be physically reducible. Provided a given dispositional concept doesn't enter into any further laws, in addition to the definitional "law" that connects it to its display, then we have no reason to expect reduction. For example, redness is arguably definable as a dispositional characteristic of objects, namely, the characteristic of producing a certain kind of perceptual response in normal observers. But it remains perfectly possible that there is nothing physically in common between all the different objects that produce this response. If the notion of redness entered into certain kinds of further laws, then there would be reason to expect a reduction. But the mere fact that all red things make normal people see red doesn't itself give reason to expect reducibility.[4]

Again, the biological notion of fitness is arguably definable as a dispositional characteristic of biological traits, namely, the characteristic of enhancing survival better than alternative traits influenced by the same genetic locus. But it would be wrong to expect that just because of this there will be anything physically in common between different fit traits and the ways they enhance survival.

Of course, if we want to explain the display of a disposition by reference to that disposition itself, as when we explain someone's visually judging something to be red by reference to its redness, or when we explain some trait's selection by reference to its fitness, then we will be committed to the disposition as something more than just the property of producing that effect. But this commitment to "something more" could just be that there is some physical basis for the production of the effect. And this commitment can thus leave it open that the physical basis might be different in different instances of the dispositional property.

Let me emphasize the requirements for the irreducibility of a dispositional property. Such irreducibility requires that the property not enter into any substantial non-dispositional laws, that is, that there not be any uniform physical cause for the different physical bases of the disposition, nor any uniform physical effects which aren't themselves effects of the purely dispositional definition (for in this latter case we would the facet he non-definitional issue of why all the different realizations which are definitionally grouped together by their dispositional display also always have another common physical effect). These are fairly strong requirements, but I see no reason to suppose that they are not satisfied in plenty of familiar cases. For example, I take it that redness does not in fact have any such uniform causes or uniform effects. There is surely no uniform physical cause for all the different physical base s for redness, nor, arguably, any uniform effect of redness independently of its uniform effect on observers. And similarly with fitness: there is no single cause of all the different physical properties which make different traits fit, nor an y uniform effect of those properties apart from their influence on survival. Which is why, once more, there is no reason to expect redness or fitness to be reducible.

We might wonder why dispositional concepts are useful, if they cove r a heterogeneity of different physical bases, and have no uniform causes or effects apart from their defining display. However, there are obvious reasons why it might sometimes be useful to classify things together just in virtue of their producing a certain kind of effect. An interior designer may not care about the molecular constitution of a fabric, nor even about how it was made, but simply about the colour responses it will produce in humans. Again, suppose we are interested in predicting the spread or extinction of some biological trait. All we need to know is its fitness, not its physical nature or its developmental history. If two different traits have the same fitness relative to their competitors, then they will evolve in the same way, whatever the other differences between them.

2.6 The Meaning of Reduction

One question sometimes raised in the literature is whether there is really a principled difference between reduction and variable realization: for, in a case of variable realization, why shouldn't we simply disjoin the various Pis which realize S, and then say that S reduces to (P1 v P2 v . . . v Pn)?

In "Special Sciences" Fodor responds to this challenge by saying that the disjunctive property (P1 v . . . Pn) won't in general be a genuine natural physical kind, and the generalization (P1 v . . . Pn) -> Q won't therefore be a genuine law. However, Fodor's analysis stops here. As he admits, he has no explicit account of what makes some kinds natural and others not, and so at this point simply rests his case on intuitions.

Other anti-reductionists have adopted a rather more sophisticated approach. Instead of resting their case on intuitions as to whether the disjunction (P1 v . . . Pn) is a natural kind, they have pointed out that the relevant disjunction might be infinitely long, or indeed mightn't even be recursively specifiable (cf. Hellman and Thompson, 1975), and that therefore the question of any reductive explanation of high-level laws in terms of lower-level ones doesn't even arise

This more sophisticated line is certainly of technical interest. But I would like to point out that the argument of this chapter adds weight to Fodor's view that even a finite disjunction of physical states can fail to qualify as a reduction (even though it disagrees with Fodor on the extent to which such reductions are needed). For the argument of this chapter suggests that such a finite disjunction of different physical states ought not to count as a reduction, whenever the picture it leaves us with at the reducing level is physically incredible.

Suppose, to return to my earlier illustration, that the appearance of inflamed ankles and knees as a result of eating reheated brussels sprouts has a name -- say, brussitis. And suppose, as before, that different cases of brussitis are due to different physical processes. My point is that the list of such physical processes doesn't need to be recursively unspecifiable for us to feel that there is something unsatisfactory about the equation of brussitis with the disjunction of those physical processes. For, as soon as the number of the disjuncts gets above two or three, we will judge, quite rightly, that it is incredible that there should be no further explanation of why reheated brussels sprouts always lead to inflamed ankles and knees.

In effect I am suggesting that the notion of a reduction is precisely the notion of an account which shows that nothing incredible is happening at the physical level. Fodor says that a finite disjunction is not a reduction because the physical categorization involved isn't "natural". I am adding the thought that this lack of "naturalness" resides in the fact that such disjunctions are too heterogeneous for it to be plausible that there should be no further

explanation for the disjuncts all producing the same effect (a thought which is not available to Fodor himself, given that he sees no need for such explanations).

2.7 Teleology and Irreducibility

Despite the argument I have been developing in this chapter, I don't in fact think that psychological categories are reducible to physical ones. I think there is a different, non-reductive explanation for why variably realized psychological states often produce uniform physical effects. It is high time I explained how this might work.

By way of an analogy, consider this example. All domestic water heaters contain thermostatic devices which stop the heating when the water gets hot enough. If we denote the threshold temperature by R, the thermostat operating by S, and the heating stopping by T, then we have the generalization, applicable to all domestic water heaters, that R -> S -> T. However, there clearly isn't any physical reduction of S here: there are many different kinds of thermostats, with quite different designs and constitutions, and with nothing physically in common apart from their all turning the heater off when the water gets hot enough.

Even so, there is scarcely much of a puzzle here as to why all the physically different realizations of S produce the common result T. The obvious answer is that water heaters are designed by people not to burn out, and that's why they all contain thermostats that switch off the heating when the water gets hot enough. We can say the mechanisms in water heaters have been selected by the designers in order to switch the water heaters off. That's what the thermostats are there for.

Another example. All vertebrates who breed within a fixed location will act towards invaders of that territory in such a way as to frighten away those invaders. Here let R be the invasion of the territory, S the characteristic behaviour, and T the departure of invaders. Then, plausibly, for such animals, R -> S -> T. Yet there is no physical reduction of S: there is nothing physically in common between all the different forms of territorial behaviour displayed by vertebrates, apart from the fact that they all make intruders go away.

But, once more, there is scarcely anything puzzling here. The obvious explanation for the fact that these physically different kinds of behaviour all have the uniform effect of frightening away intruders is that natural selection has favoured those behaviours precisely because they frighten away intruders. As in the previous example, the different physical causes have all been selected in order to produce that effect.

I favour the "aetiological" theory of teleological notions like function, purpose and design. (See Wright, 1973; Millikan, 1989b; Neander, 1991a, 1991 b.) According to this theory, it is appropriate to say that item X has the function of doing Y just in case item X is now present as a result of causing Y.[5] The paradigm for the aetiological theory is the kind of case where X has been naturally selected by a mechanism which picks out things that cause Y, as in the case of biological evolution by genetic selection. But the aetiological theory can also be extended to cover artefacts like thermostats, and indeed human actions in genera l, since human decision-making can itself be thought of as a mechanism that selects artefacts and actions because they produce certain effects.

Not everybody agrees about the aetiological account of teleology. Some people will want to put scare quotes around words like "purpose" and "design" when they are used in connection with blind mechanisms like genetic natural selection. But we need not spend time on this

issue here. For the terminology of "purposes" is not essential to my central point -- namely, that there is nothing puzzling about physically quite different things all having the same effect, if those physical things are all products of some mechanism which selects items because they have that effect. Adherents of the aetiological theory will be able to express this point by saying that there is nothing puzzling about the non-reducibility of some special science's phenomena, if those phenomena are there for a purpose. But others can put scare quotes round "purpose " here if they like.

Let us now consider the specific question of the reducibility of the generalizations of psychology. Here another kind of selection mechanism, different from both biological genetic selection and from intelligent decision-making, becomes relevant. This is selection by individual psychodevelopmental learning. There are good general reasons for supposing that individual learning, at least in its early stages, must involve some innate tendency to enhance those neural pathways which lead to certain kinds of results, and to discourage neural pathways which lead to other results.[6] In this sense learning is itself a mechanism that selects items because they produce certain results.

As with all selection mechanisms, the items between which this mechanism chooses will be relatively random, depending on such things as idiosyncrasies of individual circumstance, linguistic training, knocks on the head, or even on genuinely chance occurrences in the brain. (Compare the "mutations" which are inputs to genetic selection.) From the point of view of learning, the precise physical nature of the relevant items doesn't matter, provided they produce the right kind of effect.

And this, finally , is why there is no reason to expect there to be anything physically in common between two people when they both believe, say, that there is an ice-cream in front of them, even though this state has similar effects on their behaviour. The physical realizations are likely to be different simply because the inputs to each individual's learning mechanism (the "mutations") will be relatively random. But we needn't be puzzled as to how there can be similarity of effects without the physical commonality, for the one thing that the learning mechanism will have ensured that the different states which arise when different people look at an ice-cream will at least share the feature that they will produce appropriate effects in appropriate circumstances (such as reaching out for it when you are hungry).[7]

I earlier mentioned David Lewis's view that mental states are variably realized across species, but uniformly realized within species. The argument of this section corroborates Lewi s's commitment to variable realization across species. But it also suggests that he is wrong to stop there, and that we should expect to find variable realization within species too. Across species we find variable realization of innate mental states because genetic natural selection preserves any genetic mutation with beneficial effects, and such mutations are likely to be different in different species. Entirely analogously, if each individual contains a learning mechanism which preserves any "physiological mutation" with certain beneficial effects, and if these physiological mutations are different in different individuals, then the upshot will be that even among conspecifics we will find variable physical realizations of acquired mental states.

Actually, in the case of the mental state that Lewis himself concentrates on, namely, pain, there is a reason to expect uniform realizations within species. For pain is best thought of as part of our learning mechanism, rather than a s the kind of mental ability that this mechanism produces, given that learning selects precisely those mental items that don't cause pain. Since

our basic ability to learn isn't itself learnt, but a consequence of our genetic endowment, this is then a reason for thinking that pain, and similar basic mental states like hunger, temperature perception, and so on, are uniformly realized within species.[8]

So far in this section I have indicated a way of understanding how various biological and psychological non-dispositional categories might have uniform effects even though variably realized: such variable causes can have uniform effects in virtue of mechanisms which select items because they have that effect. The corollary, however, is that we shouldn't expect non-reducibility in those special sciences where no such selection mechanisms are to hand.[9] This seems to me to include nearly all special sciences apart from biology and psychology. Perhaps there are some rudimentary selection mechanisms in some of the social sciences, in the form of economic or social competition. But there are certainly none in such special physical sciences as meteorology, or chemistry. In any case, the general moral should be clear: special categories that aren't products of selection will be reducible.

Of course, reducibility to physics won't be at issue for sciences whose entities are partly psychologically constituted, like sociology or economics, or partly biologically constituted, like demography or epidemiology. For, even if such sciences are reducible to psychology or biology, the selection-based irreducibility of the latter sciences to physics will block the overall reduction. However, the point remains that the absence of selection mechanisms within such sciences as sociology or demography will imply that such sciences will at least be reducible to their psychological or biological constituents.

This last point is illustrated by Fodor' s (1974) discussion of Gresham's law, the economic principle that bad money drives out good. Fodor stresses the extreme implausibility of any uniform physical realization of such categories as good and bad money. And of course he is right about this. But at the same time it is worth noting that there are obvious psychological reductions of these categories, and correspondingly an obvious psychological explanation of Gresham's law. (Money consists of items which people exchange for goods and services because they expect others to do the same; such money is good or bad to the extent people believe it will continue to be so exchangeable; which is why people will circulate their bad money, and hold onto their good.) Of course these psychological facts won't in turn reduce to physical facts, given the teleological underpinnings of psychology. But the reduction of the economic facts to psychological facts is just what my overall theory predicts, given that there are no economic selection mechanisms to provide an alternative explanation of Gresham's law.

2.8 Selectional Explanations

An obvious question raised by the argument of this chapter is the status of selection al explanations themselves. I have argued that in disciplines with a teleological underpinning, such as psychology, we can explain why the disparate physical realizations Pi, of some given special category S, all have the same effect T, by invoking a selection mechanism which picks out Pis precisely because they cause T. Implicit here is the general claim that, in systems of the relevant kind, if any Pi causes T, then that Pi will tend to be preserved. But what now about this general claim? Is it itself reducible to physics? And, if not, doesn't it raise just the same puzzle about variably realized generalizations having common effects with which I started the paper?

I think that some such selectionist generalizations (if Pi causes T, then Pi is preserved) are physically explainable. Others, however, will be variably realized. But then there will be

some more general selection mechanism which in turn explains the existence of the specific selection mechanisms which pick things that cause T. What about the generalization implicit in this last explanation (if a selection mechanism picks Pis that cause T, then this selection mechanism will be preserved)? Well, either this generalization will be physic ally explainable, or it will be due to a yet more general selection mechanism which . . . is physically explainable.

So I would say that we can explain patterns that aren't themselves physically explainable in terms of selection mechanisms which are, or at least in terms of selection mechanisms whose selection is physically explainable, or . . . and so on. We can have hierarchies of selection mechanisms, with variable realizations at each level until the last. But the last level should always offer a uniform physical story, for until we have such a uniform physical story the explanatory buck with which I been concerned in this paper won't have stopped.

Let me give an illustration of the simplest case, where some variably realized pattern S -> T is explained by a physically uniform selection mechanism. Consider some simple biological organism which is capable of learning how to get rid of some given painful stimulus. Different individuals in this species will learn physically different ways getting rid of the pain. Then, if we think of the avoidance behaviour as S, its different realizations as the Pis, and the disappearance of pain as T, we will have the generalization S -> T, and this will be variably realized by the different Pis. This is the kind of variable generalization that this paper has argued to be prima facie puzzling. In the example at hand we can remove this puzzlement by invoking the learning process which ensures that, if any Pi causes T, that Pi will be preserved. The current issue, however, is what kind of explanation we can give of this general selectionist fact. But now recall a point made in the last section, that pain and associated learning mechanisms are likely to be innate and so uniformly realized within any given species. If this is right, then a uniform physical explanation for the selectionist generalization will be available. The story will no doubt be complicated. Nevertheless, to postulate that the learning mechanism is uniformly realized throughout the species is precisely to postulate some physically uniform feedback mechanism which is triggered by the disappearance of pain, and which then operates to preserve whatever physical behaviour caused that disappearance.

Does this mean that in this kind of case the original behavioural generalization S -> T will be reducible to physics after all? Only in an extended sense. There is still no uniform physical explanation of why the behaviour S generally gets rid of pain, for S is still variably realized at the physical level. Rather, what we can explain physically is why each individual, on receipt of the painful stimulus, performs some bit of behaviour which gets rid of the pain. In effect, what the selectionist story allows us to explain is not so much why the behaviour has the effect it does, but rather why each individual is disposed to some bit of behaviour with that effect. This explanation does, it is true, imply that all the behaviours in question have the effect they do; but it doesn't do this by identifying a uniform physical reduction for those behaviours; rather it switches to a broader context and instead gives a uniform physical account f or each individual having some behaviour with that effect to start with.

Now for a more complicated case. Suppose again that some group of animals have learned some common but variably realized behaviour, but not now because of their innate tendency to avoid pain, but rather because they have all acquired a common desire in virtue of their common experience. Suppose, for example, they have all learned to like bananas. And suppose that, as a result of having this desire, they have all learned ways (though not

necessarily the same ways) of doing such things as getting bananas down from trees. Now in this case there won't be any uniform physical explanation for their all acquiring such behaviour. For, if the desires for bananas are themselves acquired by learning, it is unlikely that those desires will themselves have a uniform physical realization, and so unlikely that the feedback mechanism responsible for learning how to get bananas down from trees will be physically uniform across the different individuals.

Here we need to shift to a yet wider context, and focus on how the desires for bananas were acquired in the first place. At which point we will presumably want to tell a story about a mechanism which selects states (namely, desires) which will cause, and help develop, behaviour which will yield results, such as bananas, which in the organism's experience have been associated with pleasure or the avoidance of pain -- where this selection mechanism is uniformly realized across the species. And this will then, as before, offer a uniform physical explanation, not of how the movements each animal performs lead to bananas (for these movements are physically non-uniform), nor even of how each animal has learned some bit of behaviour which gets bananas (these learning processes are physically non-uniform too), but rather of why each animal has acquired a state which disposes it to learn some bit of behaviour which will get it bananas.

We could go on. Generalizations about suitably experienced individuals seeking out bananas will hold good across species, as well as within them. But this then opens up the possibility that the associated innate mechanism for acquiring desires will be differently physically realized in different individuals who alike seek out bananas, thus undermining the physical uniformity of the story told in the last paragraph. But then we can widen the context even further, and appeal to intergenerational genetic selection, which will then explain these variably realized innate learning mechanisms as themselves selected by the physically uniform process which preserves things which cause survival and the replication of genetic DNA.

I am not of course suggesting that special scientists need to go into all this every time they appeal to some variably realized special generalization in explaining something. The idea that you must explain everything you use in an explanation is obviously self-defeating. Nevertheless, on the metaphysical level, as opposed to the methodological level, it is worth knowing that if we widen the context enough we can in principle always show that any variably realized special generalization is the upshot of some uniform physical process. For if such physical explanations weren't in principle available, it would be incredible that such variably realized generalizations should be true.

1. Functionalism per se can be defined in a way that does not require second-order causal roles to be realized by physical states. However, I have already argued for physicalism, in the last chapter. So I shall henceforth understand "functionalism" to stand for the narrower doctrine which does specify physical realizations.

2. It will ease the exposition if I can assume for the moment that functionalism includes the thesis of variable realizability. So for now functionalism is not just a thesis about the meanings of special terms, but that plus the denial of any type reducibility of the special to the physical. I shall return to this point in section 2.4 below.

3. Searle (1985) agrees with this moral, but takes it to provide a reductio of the possibility of special scientific laws: that is, he argues that, if there were special laws, then the categories of

the special sciences would have to be reducible to physics; and so, since the categories of the special sciences clearly aren't reducible to physics, there can't be any special laws.

4. I am here thinking primarily of reducibility to intrinsic physical characteristics of red objects, such as the molecular constitution of their surfaces. (Cf. Smith, 1987). But the point also applies (pace Smith) to the reducibility of redness to such relational physical characteristics as transmitting certain wavelengths of light in certain sorts of illumination; maybe there isn't even anything in common about the relational physical properties of red things, apart from the fact they make normal people see red. Indeed, there arguably isn't even any immediate reason for supposing there must be something physically in common even between the way different people respond perceptually to red objects: couldn't each person learn some physical way of reliably responding to what everybody else called "red", but with different people doing this in physically different ways? But perhaps this last extreme anti-reductionist conjecture is in tension with the ability of different people to agree in identifying previously unobserved objects as red, even though redness is a secondary quality whose instances have nothing in common except their ability to make people experience red.

5. It is unfortunate that "function" and its cognates are used both for the teleological notion of causes that are explained by their effects, and for the definitional notion of concepts that can be defined as elements in a structure of causes and effects. The two ideas are quite distinct.

6. Se e Chapter 3 of Daniel Dennett's Content and Consciousness (1969) for general reasons why learning must work like this, and recent "connectionist" models of pattern recognition for specific illustrations. In particular, see Andy Clark (1989, pp 12-13) for an instance of how similar teaching can train neural nets with random initial conditions into the same (high-level) structures.

7. I am in effect arguing that the categories of non-reducible special sciences must have purposive functions, to exp lain their non-reducibility. Lycan (19xx) and Sober (19yy) have argued similarly that functionalist theories of mind need to be supplemented by teleology, lest too many systems count as minds: their idea is that not anything with a certain causal structure should count as a mind, just those systems which are designed to have that structure. However, while the conclusions are similar, it is worth noting that the argument I have given has both wider scope and greater force. It argues that teleology is needed not just in functionalist accounts of mind, but in any law-governed non-reducible realm. And the rationale is not just that the functionalist theory of mental concepts would founder without teleology, but that there wouldn't be an y non-reducible special laws without teleology.

8. This point might seem open to dispute. If, within a species, two physically different genes produce exactly the same phenotypic effects in different ways, then both genes will be equally favoured by natural selection. Indeed, short of knowledge of their detailed DNA structure, two such genotypes are likely be counted as one by biologists. This suggests the possibility of two physically different types of pain within one species; However, it is empirically unlikely that variant physical components in such a complex mechanism as pain would ever have exactly the same selection-relevant effects. This case is different from the case of different species, for in different species natural selection will favour different physical bases for roughly the same job as long as they have roughly similar effects. But within an interbreeding species there will be direct competition between such physical alternatives, and so such alternatives will only both be preserved if they have an exact selective equivalence.

9. Why should appeal to selection mechanisms be the only way of explaining away non-reducibility? But what else could account for the fact that physically disparate items have the same effect, except some mechanism that picks them out because they have that effect?

CHAPTER 3 THE TELEOLOGICAL THEORY OF REPRESENTATION

3.1 Introduction

In this chapter and the next I shall be considering two topics which are widely regarded as raising difficulties for physicalism. This chapter will be concerned with mental representation. The next chapter will deal with consciousness. It is not difficult to see why mental representation is often thought to present a problem for physicalism. Mental states like beliefs, desires, hopes, fears, and the other propositional attitudes have representational contents: they represent the world as being a certain way. But how can this be, if such mental states involve nothing more than physical states of the brain? If my belief that Lima is the capital of Peru is realized by an arrangement of neurones, then how does this belief manage to reach out across the world and latch on to a city I have never seen? How can a bank of neurones be about something outside my head?[1]

Different physicalist theories of mind, such as functionalism, or Davidsonian anomalous monism, or any of the many other physicalist accounts of mind currently on offer, will make this problem precise in rather different ways. However, since my aim in this chap ter is to defend a positive solution -- the teleological theory of representation -- which will be available to physicalists of all kinds, it will not matter greatly exactly which version of physicalism we start with. So I shall follow the pattern of much recent literature, and start once more with functionalism.

The overall plan of this chapter will be as follows. In the next section (3.2) I shall show how representation arises as a problem for functionalism, and offer the teleological theory of representation as an initial solution. Then, after some brief comments about broad propositional attitudes (3.3), I shall elaborate some of the details of the teleological theory, in the course of answering the standard objection that some beliefs serve biological purposes even when they are false (3.4). This will prompt some discussion of the status of belief-desire psychology (3.5), and also show how the teleological theory incorporates, rather than competes with, the ide a that truth guarantees the satisfaction of desires (3.6). Sections 3.7-10 will then defend this satisfaction-guaranteeing component in the teleological theory against a number of objections, and will also consider some alternative theories which share this satisfaction-guaranteeing assumption, but do not incorporate it within a teleological context.

After this I shall return to the issue of broad beliefs, showing how it is unsurprising, given the teleological theory, that beliefs and desires should fail to supervene on brain states (3.11-12). The final two sections of the chapter will then discuss the availability of empirical evidence for the teleological theory (3.13), and point out the radically anti-verificationist implications of the theory (3.14).

3.2 Functionalism and Representation

Functionalism views beliefs and desires and other mental states as internal causal intermediaries between perception and behaviour. For functionalism, we might say, beliefs and desires are part of a system of internal pushes and pulls which explains why people behave as they do. This functionalist picture of mental states raises immediate questions about representation. After all, why should components in an internal causal structure be credited with representational powers? Surely an internal causal role is one thing, and a representational relationship to an (almost invariably) extra-cranial state of affairs another. Functionalism seems to describe only the first, causal aspect of mental states, and to omit the

second, representational aspect. As it is sometimes put, functionalism seems to give us only the "syntax" of mental states, and to leave out their "semantics".

It is true that most versions of functionalism follow everyday practice and identify beliefs and desires in terms of "content clauses", as the belief that p, the desire that q, and so on. However, from the perspective of the rest of the functionalist package, this need only be viewed as the most convenient among many possible ways of indicating the causal structure of beliefs and desires, as one way of "labelling" causal roles, and not as an essential use of representational notions. After all, how could representational relationships to often distant states of affairs be intrinsic to the internal causal roles of mental states?

It is perhaps worth pausing on this point. Despite what I have just said, doesn't the functionalist approach to the mind need to invoke assumptions about what desires are for and beliefs are about, in order to infer what agents will do? Well, functionalism does indeed attend to the causal roles of mental states; and, a s I have just said, it does take these causal roles to be indexed by content clauses. But, to repeat, it is not essential to this that the content clauses specify what beliefs are about or desires are for. A nice way to bring this out is to think of contents, as some philosophers do, in terms of sets of possible worlds. On this account, the content of an instrumental belief that F will cause G is the set of worlds in which F does cause G, and the content of a desire for G is the set of worlds in which G obtains. Given this, and given that agents tend to perform those actions that they believe are necessary for what they want, functionalism could then invoke, as a first approximation, the generalization that an agent will do F just in case the set of worlds which comprises the content of the agent's desires is contained in the set of worlds which comprises the content of the agent's instrumental beliefs about F. Note, however, that it does not matter to this generalization that these beliefs and desires represent the world as being a certain way -- that they are true (in the case of beliefs) or satisfied (desires) just in case the actual world is a member of the set of worlds which constitutes the content. All the generalization needs are the overall sets of worlds which comprise the contents, since these alone suffice to specify the interdependent causal roles of beliefs and desires; it is irrelevant that these contents also determine, together with the actual world, whether beliefs are true or desires satisfied. Which is why, from the functionalist point of view, any other similar structures could in principle serve to specify causal roles instead, even if they didn't involve the entities we normally think of beliefs and desires as about -- provided, that is, that they at least succeed in tying mental states to the bits of behaviour, the Fs, which are the end points of the causal roles the functionalist is interested in.

So the complaint is that functionalism gives only internal causal roles, and not representation. It might seem to some readers, however, that the difficulty is easily remedied. Isn't the trouble just that functionalism thinks of the "inputs" and "outputs" of causal roles too narrowly, with inputs starting with the sense organs, and outputs finishing with bodily movements? So why not simply extend our causal net to allow more distal causes of perception, on the input side, and more distal effects of behaviour, on the output side? This would allow us to analyse the truth conditions of beliefs as those distal circumstances which cause them, and the satisfaction conditions of desires as those distal states of affairs they give rise to, and would thereby seem to reintroduce aboutness without further ado.

This move, however, is fatally afflicted by the disease known as "disjunctivitis".[2] The belief that there is an ice-cream in front of you can be caused, not only by a real ice-cream, but also by a plastic ice-cream, or a hologram of an ice-cream, or so on. So, on the current suggestion,

the belief in question ought to represent either-a-real-ice-cream-or-a-plastic-one-or-any-of-the-other-things-that-might -fool-you. Which of course it doesn't.
Similarly with desires. The results which follow any given desire include not only the real object of the desire, but also various unintended consequences. So the current suggestion would imply that the object of any desire is the disjunction of its real object with all those unintended consequences. Which of course it isn't.

So, even if we widen functionalism's causal roles to include distal causes and effect s, we still need somehow to winnow out, from the various causes that give rise to beliefs, and the various results that eventuate from desires, those which the beliefs are about, and which the desires are for.

This is where an appeal to teleological considerations seems to yield a natural and satisfying answer. We can pick out a desire's real satisfaction condition as that effect which it is the desire's biological purpose to produce. And, similarly, we can pick out the real truth condition of a belief as that condition which it is the biological purpose of the belief to be co-present with.[3]

This teleological theory of representation will be elaborated and defended in detail in what follows. But at this stage let me make two immediate points. First, my use of "purpose" and similar phrases should be understood, as in chapter 2, in terms of the aetiological account of teleological notions. That is, I take it that the purpose of A is to do B just in case A is now present because in the past some selection process selected items that do B. So, in the specific context at hand, when I speak of that condition which it a desire's biological purpose to produce, I take it that some past selection mechanism has favoured that desire -- or, more precisely, the ability to form that type of desire -- in virtue of that desire producing that effect. And when I speak of the condition which it is the bio logical purpose of a belief to be co-present with, I take it that some past selection mechanism has selected that belief -- or, more precisely, the ability to form that belief type -- in virtue of its occurring in conjunction with that condition. (As in chapter 2, those readers who dislike the aetiological analysis of purposive talk can simply replace all my references to purposes by references to selection mechanisms. What matters to my story is that mental states should be the products of selection processes, not what terminology we use to specify this.)

The second immediate point I wish to make is that this selectionist-teleological approach to mental representation does not imply that all representational abilities must be genetically innate products of inter-generational selection. For selection-based teleology can also be a product of individual learning (cf. "The pigeon is pressing the bar in order to get food"). And so, if some non-innate belief or desire is selected in the course of individual learning in virtue of the condition it is co-present with, or the result it gives rise to, then that belief or desire will have a genuine selection-based representational purpose, despite its non-innateness.

3.3 Broad Contents

It will be helpful, before proceeding to further details of the teleological theory, to comment briefly on the relation between my argument so far and the recent debate about "broad" versus "narrow contents". My reference, in the middle of the last section, to the possibility of "widening" the functionalist net, may have made some readers think of recent philosophical discussions of "broad contents". However, our current concerns are rather more general than the debate about broad contents. Our present topic is to understand content as such: why do mental states, of whatever kind, have contents? The debate about broad contents, by contrast,

takes the existence of contents as such for granted, and is concerned with more detailed questions about which specific mental states have which specific contents.
The debate about broad contents arises from the observation, to which I drew attention in chapter 1, that the content of many beliefs seems to depend, not just on the believer's physical make-up, but also on features of the context. Thus Hilary Putnam has argued that the identity of beliefs about natural kinds depends on what kind s are actually present in the believer's world (1975); Tyler Burge has argued that the contents of theoretical beliefs can depend on features of the social context (1979, 1982); and Gareth Evans has maintained that the possession of singular beliefs demands the existence of the objects those beliefs are about (1982).

These philosophers and their followers form one side of the debate about broad contents. On the other side is a sizeable minority who are suspicious of broad beliefs, on the grounds, roughly, that it is hard to see how differences which lie outside the head can matter to the explanatory significance of mental states (cf Fodor, 1987).

This is why I said the debate about broad contents is less general than our current concerns. The participants in the debate take it as given that our beliefs and other attitudes have representational contents. The point of dispute is only whether or not these contents are fixed by internal physical make-up.

John McDowell (1986, sect 5) has suggested that it is only possible to get worried about the general possibility of representation if you make the mistake of thinking that all beliefs are narrow. McDowell's thought is that a problem about representation only arises as long as we think of beliefs as things inside people's head. Once we recognize that the very possession of a belief can involve extra-cranial facts, we ought no longer to be puzzled about how things inside the head can stand for things outside.

This seems to me to get things exactly the wrong way round. Merely accepting that the possession of beliefs involves entities outside believers' heads does little to explain how representation as such is possible. After all, plenty of the other states that people possess involve entities outside their heads -- for example, financial solvency, or popularity, or being married -- without thereby becoming representational.

Far from appealing to broad contents to dissolve the general problem of representation, I think we will do better to solve the general problem of representation first, and then apply the solution to the issue of broad contents. In the absence of any general understanding of representation as such, much of the current debate between the friends and enemies of broad contents has collapsed into an indecisive trading of intuitions. However, once we have arrived at a satisfactory general theory of content, then we shall understand why it is quite unsurprising that some contents should be broad. I shall return to this issue in section 3.10 below.

3.4 Functional Falsity
A good way to develop the details of the teleological theory is to consider a familiar objection. This is the objection that certain beliefs have biological purposes which require them to be present when they are false, and so constitute prima facie counter -examples to the teleological thesis that truth conditions can be analysed as those circumstances in which beliefs are biologically supposed to be present. (Ned Block has urged this objection on me. See also Stich, 1982, p 53.)

For example, consider the belief that you are not going to be injured in some unavoidable and imminent trial of violence. It is arguable that natural selection has bequeathed us an innate disposition to form this belief, even in cases where it highly likely that we will in fact be injured, in order to ensure that we will not flinch in battle. But it then seems to follow that, according to the teleological theory, the truth condition of this belief will include many cases where we will be injured -- since such cases will be among those where we are biologically supposed to have the belief. So we seem to have a reductio of the teleological theory. For by hypothesis the truth condition of the belief is that we won't be injured.

Examples like this are interesting, but I don't think they suffice to discredit the teleological theory. In order to see why not, we need to consider the way that beliefs and desires combine to generate actions in the overall human decision-making system. It will emerge that the purpose of beliefs in this system is to guide actions in such a way that desires will be satisfied. And then, by understanding the teleological theory as focusing on this specific purpose of beliefs, we will be able to accommodate examples of the above kind. The point will be that stopping you flinching is a special kind of biological purpose, which cuts across the purpose of satisfying desires,[4] and which therefore does not require the truth of beliefs in the way that the satisfaction of desires does.

The overall biological function of the human decision-making system is to generate actions that cause biologically suitable results. Beliefs and desires both contribute to this purpose. However, they contribute in different ways. The role of desires is to do with the fact that different results are suitable at different times: our desires vary in order that our actions will produce different results at different times. The role of beliefs is to do with the fact that, given any result, different means are appropriate to that result at different times: our beliefs vary in order that we can choose the most effective means at any time to the results that we desire at that time.

In the end, all selection-based purposes depend on results: to have a purpose is to have been selected by a mechanism which favours certain results. However, the above remarks show that this is true of beliefs only in an indirect sense. For beliefs don't have any results of their own. Rather, their standard purpose is to produce whichever results will satisfy the desires they are acting in concert with. In effect, beliefs get selected at one remove, in virtue of being good at causing actions which cause desired results.

Note that this means that, according to the teleological theory, there is a sense in which the representational powers of desires are prior to those of beliefs. Any given desire will be present in order to produce a certain result r, which result is therefore its satisfaction condition. Given this explanation of satisfaction for desires, we can then explain the purposes of beliefs. Any given belief will be present in order to produce actions which will produce desired results if a certain condition p obtains, which condition is therefore that belief's truth condition.[5]

Let us now return to the example with which I began this section. I have just argued that the biological purpose of any belief is to be present in those circumstances in which the actions it prompts will satisfy desires -- which circumstances therefore count as its truth condition. However, the example about not flinching in battle involves a different kind of purpose. For in this case the belief at issue, the belief that you won't be injured, has an extra biological

purpose, apart from its role in aiding the satisfaction of desires, namely, to ensure that you do not flinch in battle.

In order to deal with such examples, we need to distinguish the "normal" purpose of beliefs, namely, to ensure the satisfaction of desires, from such "special" purposes as stopping you flinching in battle. This distinction then allows us to frame the teleological theory in a way which is consistent with the existence of such special purposes. That is, we should understand the teleological theory as relating specifically to the normal purpose of beliefs. For, as long as we stick to those normal purposes, then truth is still the requirement for achieving them, in line with the teleological theory of representation.[6]

If you are unconvinced that the belief about invulnerability needs to be true in order to serve its "normal" function, consider the case, say, of Cuthbert Coward. Cuthbert would far rather remain unscratched than win the battle. Still, if Cuthbert were somehow to be persuaded that he won't be injured (though in fact he will), then even he might be induced to enter the fray. But then he won't get what he desires, which is above all to remain unscathed. It is only the special purpose of getting him to fight, even though he's doesn't really want to, that gets satisfied when the belief is false. By contrast, the normal purpose, of satisfying his desire to remain unscathed, still requires his belie f to be true -- just as the teleological theory, as now proposed, requires.

Of course, Cuthbert has somewhat unsatisfactory desires, from a biological point of view, in the sense that the satisfaction of his desires is unlikely to further his overall chances of survival and reproduction. This is why beliefs sometimes have special purposes. The point of these special purposes is in effect to by-pass the normal role of beliefs in satisfying desires, and to ensure instead that agents with biologically inappropriate desires don't end up performing biologically inappropriate actions. Cowards are a case in point. Their unfortunate desires mean that they are likely to end up running from battle, and thus losing the any chance of biologically important spoils, just in order to avoid a scratch. And so, in order to protect them against the biological dangers of such consequences, natural selection predisposes them to believe that they are invulnerable, even when the evidence doesn't warrant this belief, so as to stop them performing those actions which would in fact satisfy their desires.

It might seem puzzling that natural selection should give some beliefs two different purposes. After all, natural selection presumably designs biological systems for one ultimate end, namely, the bequest of genes. So why don't beliefs simply have the single purpose of ensuring such gene bequests?

The answer relates once more to the nature of the human decision-making system. Note that this system doesn't work by always choosing that action which is most likely to ensure gene bequests. Rather it chooses that action which is most likely to satisfy existing desires. It is not impossible to imagine biological systems of the former kind, which always aimed directly for gene bequests. But it seems likely that the limitations of our cognitive capacities have prevented us from doing things in this way. Instead we aim for such relatively short-term goals as warmth, sex and chocolate ice-cream.

By and large such short-term goals correlate reasonably well with ultimate biological success, which is no doubt why our innate desires, and our ways of acquiring non-innate desires, have evolved as they have. But the satisfaction of our desires won't always coincide with

biological success (not all sex leads, or even can lead, to reproduction). And this then means that there are certain biological risks consequent on our way of doing things. Now, it may be that some of these risks are inevitable by-products of our desire-based decision-making system: for example, it may be inevitable that humans will have extremely strong desi res to avoid injuries, and so inevitable that in certain circumstances this will lead them to act against their biological interests. And this will then lead to natural selection interfering with the normal operation of decision-making system, by giving us beliefs which lead us to act in ways that frustrate our desires, but satisfy our biological needs.

Let me sum up the argument of this section. Certain beliefs do indeed have some biological purposes that require them to be false. However, this doesn't invalidate the teleological theory of representation. For we can understand the teleological theory as focusing specifically on the normal purposes of beliefs, namely, to guarantee the satisfaction of desires. And these normal purposes don't ever require beliefs to be false.

3.5 The Reality of Beliefs and Desires

In the last section I made a number of definite assumptions about the role of beliefs and desires in our overall decision-making system. Some readers may want to ask how this tallies with the agnostic attitude to everyday psychology I expressed in chapter 1, when I said that my references to the entities of everyday psychology should be understood merely as place holders for the true theoretical explanation of human cognition, whatever that may be.

One possible response to this query would be to maintain that the last section's comments about the roles of beliefs and desires need not be read realistically, as committing me to substantial claims about the causal structure of our cognitive system. Daniel Dennett, for instance, argues (1971, 1978, 1987) that everyday psychology commits us only to the "intentional stance", to the view that an individual's behaviour is somehow appropriate to his or her environment and needs, and not to any "design" or "physical" assumptions about the mechanisms that might be responsible for generating that behaviour. Dennett holds that this intentional stance is underpinned by general evolutionary considerations, which tell us that our cognitive systems must have some design that will enable us to choose actions that will further our welfare, while leaving open the internal details of that design. On Dennett's conception, then, references to such everyday concepts as belief and desire need not be taken as realistic hypotheses about internal structures, but simply as a way of pointing to the appropriateness of actions.

However, I shall not take this Dennettian line. For one thing, it sits ill with the teleological theory of representation. According to the teleological theory, the representational contents of beliefs and desires depend on how (the abilities to form) these states have been shaped by natural selection. But if beliefs and desires aren't real states, but only constructs by which we indicate the appropriateness of actions to circumstances, then it is hard to see how natural selection can operate on them. Natural selection favours things which produce certain effects. But it can't favour things which don't exist.[7]

In any case, there is good reason to doubt Dennett's view that everyday psychology is restricted to the "intentional stance". This relates to a point made in the last section. As we saw, natural selection hasn't arranged our brains so that we always choose actions that are likely to maximize gene bequests. Instead it has fixed on certain relatively short-term goals, like warmth and sex, and on certain ways of acquiring further short-term goals, and arranged for our brains to choose actions which are likely at least to satisfy these goals.

As I observed at the end of the last section, this makes sense from the point of view of natural selection, given that these short-term goals correlate reasonably well with long-term reproductive success, whereas aiming directly for such long-term reproductive success would no doubt overtax our cognitive capacities (not to mention the cognitive capacities of our evolutionary ancestors). But the fact that the installation of short-term desires constitutes a sensible strategy from the point of view of natural selection shouldn't obscure the fact that it is a definite design option, a choice of one among a number of different possible internal structures which could ensure that behaviour is more or less appropriate to needs and environment. After all, we can easily enough imagine hyper-intelligent non-human beings whom natural selection had made "super-rational", by giving them no short-term desires as such, but simply the sole aim of maximizing gene bequests by always choosing that action which available information indicated as most likely to achieve that end. And, at the other extreme, we already have terrestrial examples of simple organisms, like insects, with plenty of hard-wired routines driven by short-term needs, but scarcely any ability to modify their behaviour in response to information about the environment.

So everyday psychology, with its distinction between beliefs and desires, takes us beyond the thought that evolution has somehow arranged that we will choose actions appropriate to our needs and environment, to a specific theory of how evolution has arranged this: evolution has arranged for us to have information about our circumstances, in the form of our current beliefs, and then to choose actions which those beliefs indicate will satisfy the goals signalled by our current desires. In these respects we are different from the "super-rationalists", since they are not interested in any intermediate goals except gene bequests; and we are different from the insects, in that their behaviour is almost entirely insensitive to information about their circumstances.

In the light of these points, I accept that my appeal to beliefs and desires in the last section does indeed take me beyond the stance of chap ter 1, and commit me to the truth of certain basic everyday psychological assumptions as realistic hypotheses. However, now that we have seen why this commitment is inescapable, we can also see why it is unburdensome. For there is plenty of uncontentious empirical evidence that everyday psychology is true at just those points where it takes us beyond Dennett's intentional stance. The significant point is not just that everyday psychology says that we are different from the super-rationalists and the insects in having an internal structure of beliefs and desires; in addition, our actual behaviour shows that we are different in this respect. If we didn't differ from the super-rationalists in having desires, then we wouldn't continue to act in pursuit of short-term aims, like eating chocolate, even after we know that doing so only makes us fat and so is no help to our reproductive success; and if we didn't differ from the insects in having beliefs, we wouldn't be able to figure out that one way to acquire some chocolate would be to go to the new confectionery shop around the corner.[8]

This kind of general evidence does not of course confirm every detail of the complex set of assumptions and attitudes which constitute our everyday psychological thinking. But it does seem to me to be enough to justify the kind of core assumptions about the existence of beliefs and desires that I made in the last section. Empirical psychology still has much to discover, both about the more detailed claims made by everyday psychology, and about the "sub-personal" structures by which such everyday psychological claims are implemented. But I don't think it need do anything further to establish that human actions are generated by

internal causal processes involving beliefs and desires. If our actions were not generated in this way, we would behave quite differently from the way we know we do behave.[9]

3.6 Truth as the Guarantee of Success

In section 3.4 I stressed that the teleological theory of representation needs to be understood as focusing specifically on the role that beliefs play in facilitating the satisfaction of desires, rather than on any further role they may have in fulfilling further biological purposes. However, once we focus on desire satisfaction in this way, then do we still need teleology to explain truth-conditional content? Why not simply explain content directly, by saying it is that property of beliefs which will ensure the satisfaction of desires?

At the beginning of this chapter I argued that functionalism leaves out representation, and that the teleological theory is needed to bring it back in. But perhaps the moral of my remarks in 3.4 about the relation between belief content and desire satisfaction is that we shouldn't start with functionalism in the first place. For what those remarks in effect show is that functionalism presents only a limited picture of the role that mental states play in psychological explanation, a picture which leaves out the role of truth in ensuring the satisfaction of desires. Perhaps once we fill in the missing components of the picture, we won't any longer need teleology to explain representation.

In due course I shall show that this is not so: even after we have paid due accord to the role of truth in ensuring desire satisfaction, we will still need teleology for a full explanation of representation. But it will be worth proceeding slowly.

The limitations of functionalism can be brought out by contrasting two different pictures of the structure of action explanation. The first picture, the picture embodied in functionalism, focuses on the internal roles that beliefs and desires play in causing behaviour, and so takes psychological explanation to conform to this pattern:

(A) X desires G
X believes that F will produce G

_____

X does F.

However, there is also a second picture of the structure of action explanation, a picture embodied in my remarks about the human decision-making system. According to this picture, psychological explanation is not solely an internal matter, but also has an "external" structure, which explains, not behaviour, but the achievement of results:

(B) X desires G
X believes, of some behaviour, that it will produce G
This belief is true

_____

X achieves G.

If we restrict our attention to "internal" explanations of form (A), as functionalism does, then it scarcely surprising that we become puzzled about the significance of representational notions, since the only role that beliefs an d desires play in (A) is that of causal pushes from

the inside, as it were, and not as representers of the external world. But in "external" explanations of form (B), the representational features of beliefs and desires become crucial: the satisfaction condition of the desire specifies what external result is at issue, the truth condition of the belief specifies how things must be to ensure this result, and the actual truth of the belief specifies that things are indeed so. Far from being limited to the internal causes of behaviour, explanations like (B) specify that external circumstances are such as to lead from the agent's behaviour to result G.

This is why we now need to ask whether we really need to appeal to teleology in our theory of representation. The original puzzle that led us to this theory was, in effect, that internal explanations like (A) make no use of representational notions. But now we see that external explanations like (B) do use representational notions. And this suggests that we might be able to analyse representational notions purely in terms of the way they enter into such external explanations -- explaining truth, say, as that property which ensures desire satisfaction -- without needing to appeal to teleological considerations after all.

In a moment I shall explain why this doesn't quite work. But let me deal with a minor point first. There is an extensive literature on the question of whether representational notions are essential to (B). (See Loar 1981; Devitt, 1984; Field, 1986.) Can't an explanation like (B) always be replaced by a two-stage explanation which first explains behaviour F, as in (A), and then explains G by reference to the fact that F causes G, and thereby omits any explicit mention of truth? Well, maybe so. But the obvious question is why we should want to dispense with truth in this way. The answer, for most of the contributions to the relevant literature, is to do with "deflationary" or "minimalist" theories of truth: defenders of such theories are committed to the replaceability of (B)s by (A)s, since they think that mention of truth is always simply a "quotational" variant of what can be said in disquoted terms; while opponents of such theories want to show that (B)s involve ineliminable appeal to truth as a real property of beliefs. My present concerns, however, are orthogonal to this debate. I am not concerned to decide how far talk of truth might be eliminable in favour of something else, but simply to take it at face value, and understand what work it does in our thinking about the world. The question at hand is not whether we can do without truth, but what we do with it . (As it happens, I think that replacing (B)s by (A)s loses sight of a general explanatory pattern, the pattern displayed in schema (B). On the other hand, I don't think that this is the most effective way to argue against the deflationary theory, given that arguments based on the importance of explanatory patterns are notoriously inconclusive. A better strategy is to press the deflationalist for a theory of translational content. I shall return to this issue in section 3.9 below.)

The question currently at issue is whether we can analyse representational notions simply on the basis of the way that they enter into external explanations like (B), and without appeal to teleological considerations. Let us consider in more detail how this might work. The idea, in outline, is that truth conditional content might be analysed in terms of the role of truth in ensuring desire satisfaction. We can formulate this suggestion explicitly as follows:

(C) The truth condition, for any belief, is that
condition which guarantees that actions based on
that belief will satisfy the desires it is acting
in concert with.

Something like this success-guaranteeing analysis of truth has been proposed by a number of other writers (Ramsey, 1927, p 29; Putnam, 1978, part 3; Appiah 1986; Mellor 1988; Whyte,

1990.) However, there is an obvious reason why it is not, as it stands, an adequate substitute for the teleological theory. Namely, that it explains truth, for beliefs, only by assuming the notion of satisfaction, for desires. Yet satisfaction is as much a representational notion as truth, and so ought itself to be explained by an adequate philosophical theory of representation.

It is no good simply offering an account of desire satisfaction parallel to (C), such as:

(D) The satisfaction condition of a desire is that
condition which is guaranteed to result from
actions based on that desire, if the beliefs behind
the action are true.

For simply adding (D) to (C), without offering any further hold on representational notions, is like trying to solve a single equation with two unknowns. Both (C) and (D) are expressions of the principle:

(E) Actions based on true beliefs will satisfy the
desires they are aimed at.

(E) places a mutual constraint on the representational values that a person's beliefs and desires can have. But on its own it does not suffice to pin down those values uniquely. If a given attribution of truth and satisfaction conditions satisfies (E), then so will any attribution that simply permutes referents for names and predicates, provided it does so in the same way in both truth and satisfaction conditions. (Cf Stalnaker, 1984, pp 17-18; Papineau, 1984, p 555.)

This means is that any theory of representation that explains truth by (C) needs to add something further -- not just (D) -- to explain satisfaction. I add teleology. I explain desire satisfaction in terms of the results that desires are biologically supposed to produce, and then plug this into (C), thus giving truth the biological purpose of satisfying desires.

There are perhaps other possible options at this point. You might agree with the success-guaranteeing account of truth, as in (C), and agree that something extra is needed, yet disagree that the requisite addition is teleology. However, let us postpone the question of whether (C) can be appropriately supplemented in non-teleological ways until the section after next. For the prior question is whether (C) is even defensible as part of a full account of representation. There a number of standard objections to the idea that truth is what guarantees desire satisfaction, which both the teleological theorist and those who want to supplement (C) in other ways need to answer. It will be convenient at this point to deal with these objections.

3.7 Objections to a Success-Guaranteeing Account of Truth

(i) Non-Instrumental Beliefs

Doesn't (C) apply only to beliefs of the form: s will bring about t? For these are the only beliefs which are directly relevant to the satisfaction of desires, as schema (B) makes clear. Surely, however, an analysis of truth conditions ought to deal with beliefs of all forms, and not just with beliefs about means to ends.

It is not difficult, however, to see why (C) should be considered to hold for beliefs of all forms, as well as for means-ends beliefs. It is true that the relevance of beliefs to actions always depends in the last instance on what they imply about appropriate means. And in this sense it is only means-ends beliefs that are directly relevant to actions. But, still, such means-end beliefs, that s will bring about t, will as a rule be inferred by the agent from various other beliefs. And this then institutes the requisite general connection between truth and satisfaction. For if those other beliefs are true, and the inferences from them valid, then the belief that s will bring about t will be true too, and the resulting action will succeed. So it is a general principle that actions based on true beliefs will succeed, and not just a principle about means-ends beliefs as such.[10] Consequently, when we invert this principle into an analysis of truth conditions -- analysis (C) -- the analysis promises to apply to beliefs in general, and not just to beliefs of the means-end form.

(ii) Actions Based on More Than One Belief

In general a number of beliefs will lie behind any given action. But this means that the truth of any one belief will be insufficient to guarantee the success of ensuing actions. For desire satisfaction will only be guaranteed if the other beliefs behind the action are also true. So strictly analysis (C) ought to be formulated:

The truth condition, for any belief, is that
condition which guarantees that actions bas ed on
that belief will satisfy the desires it is acting
in concert with, assuming that any other beliefs it
is also acting in concert with are true as well.

But this then disqualifies (C) as analysis of truth-conditional representation, for it assumes the notion of truth in explaining it.

It might seem that we could deal with this difficulty by thinking of analysis (C) as applying specifically to cases where single beliefs generate actions on their own, without the assistance of other beliefs (cf Mellor, 1988, p 86). Truth conditions could then be identified as what guarantees satisfaction in such single-belief cases. But the trouble with this is that we then run into objection (1) again, since the only kind of beliefs that can generate actions on their own are means-ends beliefs. If we want an analysis of truth that works for beliefs in general, and not just for means-ends beliefs, then we need a way of extending (C) beyond single-belief choices of action.

A better way to deal with the problem is to think of analysis (C) as being applied simultaneously to all the belief types in an agent's repertoire. That is, we should think of (C) as fixing the truth conditions for all those beliefs collectively by, as it were, solving a set of simultaneous equations. The "equations" are the assumptions that the truth condition of each belief will guarantee desire satisfaction, if other relevant beliefs are true; the overall "solution" is then a collective assignment of truth conditions which satisfies all those equations.

(iii) Can't False Beliefs Satisfy Desi res?

Another initial worry about (C) might be that it makes truth too easy. Surely we don't want to count beliefs as true whenever the actions they prompt have satisfactory results. Can't an action achieve a desired result by accident, even though some of the beliefs behind it are false (as when they involve some self-correcting mistake)?

But (C) doesn't in fact rule out this possibility. The suggestion isn't that it's enough, for the truth of a set of token beliefs, that a particular action, prompted by those particular tokens, should satisfy desires. Rather (C) specifies a condition which guarantees, for all tokens of the relevant types, that ensuing actions will satisfy desires.

(iv) Decisions Made Under Uncertainty

In many case an agent will act, not on full beliefs, but on partial beliefs. In such cases the agent's thinking won't pick out any action as certain to satisfy desires, but rather select the action that is subjectively most likely to satisfy desires. But then, if the action does succeed, that won't have been guaranteed by the truth of the agent's beliefs about the world.

It is an interesting question as to how far the well-foundedness of decisions made under uncertainty depends on objective features of the world, such as the existence of objective chances. But we can by-pass this issue here. For, once more, there is nothing in (C) which rules out the possibility of actions whose success isn't guaranteed by the truth of the beliefs behind them. The idea behind (C) is rather that we should focus on the kind of case where success is so guaranteed, and then analyse truth as what guarantees desire satisfaction in just those cases. So uncertain decisions issuing from partial beliefs are beside the point. To apply (C) to a given belief, we should stick to cases where that belief is held fully, and figures in decisions which aren't uncertain: truth is what guarantees satisfaction in those cases.[11]

(v) Is Truth Just Pragmatic?

Analysis (C) seems to imply that the virtue of truth is essentially pragmatic, that the reason for wanting truth is always so as to satisfy desires. But surely truth can be pursued as an end in itself, and not just because of its pragmatic value. Indeed there are certain questions, about the farther reaches of the universe, say, or the distant past, where our interest in having true beliefs can't possibly be practical, since such beliefs can make no difference to our actions.

But this complaint misses its target. (C) isn't a theory about why we should want truth. It's a theory of what truth is: namely, for a belief, the obtaining of a condition which guarantees that, if an agent were to act on that belief, the ensuing action would satisfy desires. This doesn't presuppose that anybody will actually act on the belief. Nor does it presuppose that the only reason for wanting the truth in respect of that belief is to be able to act so as to satisfy desires. To be sure, if you do want to satisfy desires, then (C) does immediately imply that you have a motive for wanting the beliefs behind it to be true. But that leaves room for other motives for wanting truth, both in the case of practically significant beliefs and practically insignificant ones. In particular, it leaves room for truth to be valued as an end in itself. (Can't we now ask: why should truth be valued as an end in itself? But I take it to be a virtue of (C) that it allows this as a significant question.)

It might still seem that there are some beliefs that couldn't, even counterfactually, be relevant to an action satisfying a desire. What about the belief that there are no agents, or the belief that all my actions are doomed to failure?[12] At this point we need to appeal to the compositionality of beliefs. As I shall explain in section 3.9, we need to recognize that beliefs are made up of components ("concepts"), the representational significance of which derives from their systematic contribution to the truth conditions of the beliefs they enter into, that is, from their systematic contribution to conditions which guarantee that actions based on those beliefs will satisfy desires. Once we recognize this, then we can hope to pin down the representational significance of concepts like agent, doomed to failure, and so on, in terms of their contribution to beliefs which can be relevant to action, and then use those

representational values to build up truth conditions for such special beliefs as can't be relevant to action.

(vi) Non-Natural Beliefs

Analysis (C) applies only to beliefs whose truth is of potential causal relevance to the success of actions. Perhaps this will enable it to accommodate beliefs about the natural world. But what about moral, or modal, or mathematical judgements? In what sense, if any, can the truth of such non-natural judgements matter to the success of action?

I don't propose to pursue this complex topic at this point. Whether or not analysis (C) might apply to a given category of judgement depends on the details of the workings of such judgements, and such details are matters of active controversy for moral, modal, and mathematical judgements. I shall offer some further comments on these issues in chapter 6 below.

(viii) Doesn't (C) Presuppose Validity and Hence Truth?

In my answer to objection (i) I appealed to the notion of validity: I argued that analysis (C) could be extended from means-ends beliefs to other beliefs because valid inferences from true beliefs of any kind will lead to true conclusions about appropriate means. However, it might be argued that this appeal to validity is illegitimate, on the grounds that the notion of validity presupposes the notion of truth.

Analysis (C) certainly needs the notion of validity. Often agents will draw invalid inferences about means (imagine that they have to decide what to do quickly, or that their situation is very complicated) and then the truth of the beliefs on which those inferences are based won't guarantee the success of their actions. So if analysis (C) is to apply generally, and not just to means-end beliefs, it should strictly be formulated as:

The truth condition of any belief is that
condition which guarantees that actions validly
based on that belief will satisfy desires.

But this now makes the problem clear: (C) can scarcely be held to constitute an analysis of truth, if it presupposes validity and validity presupposes truth.

One possible move here might be to deny y that validity does depend on truth. Thus we might seek some purely syntactic notion of validity, defined in terms of some specified structure of rules of inference, rather than the semantic notion of any truth-preserving form of inference. However, this syntactic strategy seems unpromising. For a start, there are technical difficulties about the completeness of syntactic characterizations of non-first-order validity. And, in any case, given that syntactic characterizations are always answerable to the semantic conception of validity (cf Dummett, 1974), even for first-order validity, it is doubtful that the syntactic strategy will really dispose of the circularity, rather than just brushing it under the carpet.

To deal with this difficulty, I think it is necessary to broaden the focus away from analysis (C) itself, and reintroduce teleological considerations. We need to think of validity as playing a part, alongside truth and desire satisfaction, in fulfilling the biological purposes of the overall human decision-making system.

It is fairly obvious, on reflection, that this decision-making system needs some mechanism for generating beliefs about means, beliefs that are directly relevant to actions, from the total set of background beliefs that may bear indirectly on the achievability of desires. And it will clearly be part of the biological purpose of this mechanism to produce true beliefs about such means, given that the background beliefs are true. Of course this inferential mechanism won't always succeed in fulfilling this purpose: as I just observed, humans often draw invalid conclusions about which means to adopt. But that doesn't show that validity isn't the inferential mechanism's purpose, any more than heart failures show that blood circulation isn't the heart's purpose.

As I mentioned earlier (see footnote 5), the biological purposes of beliefs and desires are interdependent, in the sense that desires will only fulfil their biological purposes if beliefs fulfil theirs, and vice versa. We now see that there is a further interdependency, in that both beliefs and desires will only fulfil their biological purposes if the inferential mechanism fulfils its purpose too, and vice versa. There is of course nothing surprising about such interdependencies. They are a common feature of biological systems. For example, the lungs will only fulfil their biological purpose, of oxygenating the blood, if the heart fulfils its purpose, of circulating the blood, and vice versa.

It might not be immediately clear how these observations about biological purposes are supposed to solve the original problem. Don't they just amplify the point that truth, in beliefs, and validity, in inferences, presuppose each other, thereby blocking any possibility of explaining one in terms of the other? But the point of reintroducing biological considerations is not to deny this interdependence, but rather to show how we can analyse truth and validity simultaneously.

Suppose we start off not presupposing any representational terms like "truth" or "validity". We proceed to describe the workings of the human decision-making system. It has various interdependent components: some states (desires) have the biological purpose of prompting actions which will produce specific results; others (beliefs) have the biological purpose of prompting actions which are appropriate to specific circumstances, and hence the biological purpose of co-varying with those circumstances; and then there is an (inferential) mechanism whose purpose is to generate new beliefs out of old ones, under the constraint that the circumstances which the latter beliefs are supposed to co-vary with should be guaranteed by the circumstances the former beliefs are supposed to co-vary with. And then having done all that, without using the notions of "truth" and "validity", we can now account for these notions, by saying that beliefs are true when they fulfil their purpose of co-varying with the relevant circumstances, and that inferences are valid when they fulfil their purpose of preserving such truth.

3.8 Alternative Accounts of Desire Satisfaction

This completes my catalogue of standard objections to the success-guaranteeing account of truth-conditional content given by (C). My answer to the last objection returns us to the point at which we left the overall argument. For this answer dealt with the difficulty about validity by locating (C) within the biological analysis of the overall human decision-making system. But we have already noted, at the end of the section before last, a rather more straightforward reason for making this move. Namely, that (C) on its own simply explains truth, for beliefs, in terms of satisfaction, for desires, and therefore needs supplementation by an independent account of des ire satisfaction. My earlier suggestion was that we should fill this gap too by placing (C) within the biological context of the overall human decision-making system. For

this move then allows us to view desires as having a biological purpose, namely to prompt actions which produce specific results, and so enables us to analyse desire satisfaction in terms of this purpose.

A question raised when I made this suggestion was whether this is the only way to remedy the philosophical incompleteness of (C). Couldn't opponents agree with the rest of my argument, but disagree about the teleology? That is, couldn't they agree that (C) is only part of the truth about truth, which therefore needs to be supplemented with some further account of desire satisfaction, but then diverge by offering some different explanation of satisfaction for desires, which does not appeal to considerations of biological purpose?

For example, they might try to identify the results which satisfy desires as those which extinguish those desires (cf Russell, 1921, ch 3; Whyte, 1991). In general, when some desired result is achieved, then that desire disappears. So perhaps we can identify which results are the objects of which desires by reference to which results make those desires go away.

Another alternative would be to appeal to the reinforcement of behaviour (cf Dretske, 1988). Often, when a desire prompts some behaviour which produces a given result, that behaviour is reinforced, in the sense that it is more likely to be repeated when that desire next arises. So perhaps we can identify the results which satisfy desires as those results whose achievement leads to the reinforcement of behaviour.

One problem facing theories of this kind is that they will still face the problem about validity raised at the end of the last section. I dealt with this problem by viewing inferential abilities as part of the overall biologic al system, and accounting for validity in terms of the biological purpose of this ability. Accounts which seek to dispense with considerations of biological purpose obviously cannot offer this solution. Yet they will still face the problem, for merely adding an independent account of desire satisfaction to (C) will still leave us with the problem that (C), if it is to work at all, needs implicitly to presuppose an idea of validity, and hence of truth.[13]

There are other problems facing the alternative suggestions about desire satisfaction. Take the "extinction theory" of satisfaction first. On the face of it, some desires are only fuelled their own satisfaction (salted peanuts), while others are quenched by their non-satisfaction (sour grapes). Perhaps an extinction theory can somehow be elaborated so as to deal with these prima facie counter-examples. But until this is done, the teleological theory seems to offer a far more powerful and promising approach to desire satisfaction.

As to the "reinforcement theory", it seems odd to view this as a more fundamental account of desire satisfaction than that provided by the teleological theory. The phenomenon at issue is that a given action X prompted by a given desire will tend to be repeated just in case that action gives rise to a given result G -- which result the reinforcement theory therefore counts as the desire's satisfaction condition. Now, such reinforcement is certainly a genuine phenomenon. But consider it from a biological point of view. From the biological perspective, reinforcement of some means X amounts to an alternative route to achieving G, alongside the cognitive procedure of noticing that in general X leads to G and acting on this belief. That is, natural selection in effect sometimes arranges for us to acquire a derived desire for X in itself, instead of leaving it to our cognitive system to choose X on the basis of our prior desire for G and the explicit belief that X is an effective means to G.

This suggests, however, that reinforcement is, in evolutionary terms, a relatively primitive method of generating actions. In section 3.5 above I ha d occasion to observe that we human beings fall short of the kind of biological "super-rationality" which would always choose actions on the basis of explicit beliefs about the most effective way to maximize gene bequests. But at the same time I pointed out that we have moved some way in this direction, in that we are capable of doing things which we do not desire in themselves, simply because we believe them to be means to things we do desire. To this extent, then, we are more sophisticated than organisms who rely entirely on reinforcement, and whose only way of benefitting from evidence that X is normally followed by G would be to acquire a derived desire for X.

In view of our greater sophistication in this respect, it would be surprising if our successes in achieving desires were always followed by the reinforcement of the means adopted. Given that we humans can select actions as a result of deliberation as well as conditioning, such automatic reinforcement would be both unnecessary and potentially disadvantageous. And in fact it doesn't always happen. Even after much experience of satisfying my desire for chocolate by going to the corner shop, I do not find that I have any desire to visit the corner shop as an end in itself.

This implies that the reinforcement theory cannot suffice as an account of desire satisfaction. To the extent that some desires can be satisfied without the means adopted being reinforced, as in this last example , we will be unable to equate the satisfaction conditions of those desires with results which lead to the reinforcement of means. If we want a theory of satisfaction that works across the board, we will do better concentrate on those results which desires are supposed to produce when they combine with beliefs in the deliberate choice of actions.

3.9 Do We Need Reified Truth Conditions?

In this section I want to focus on the ontological commitments of the account of representation I have developed so far. (It will be convenient to concentrate on beliefs, but most of the points which follow could be applied to desires too.) We can summarize the account of truth conditional content we have now arrived at a s follows:

(F) The truth condition, for any belief, is that
condition which guarantees that actions generated
by that belief will fulfil its biological purpose
of satisfying desires.

Note, however, that this analysis (F) (like (C) before it) refers to "truth conditions", and implicitly views truth itself as a matter of such conditions "obtaining". This creates a prima facie problem. For such talk, if taken at face value, commits us to dubious entities like propositions, or possible states of affairs, or sets of possible worlds.

Some philosophers would be untroubled by commitments to abstract objects like propositions and sets. They can skip ahead to the next section. But I am unhappy with such commitments, for reasons to be given in chapter 6 below, and so in this section I want to try to show that the reification of truth conditions is not essential to (F).

My argument so far implies that, for any belief-type in an individual's repertoire, an instance of the following schema will hold:

(G) actions generated by that belief will fulfil the belief's purpose of satisfying desires if and only if p

Given this, then one way of understanding analysis (F) is as asserting that claims of the form (G) specify the truth-conditional contents of beliefs. That is, analysis (F) can be understood as asserting that (G) is an equivalent substitute for:

(H) the belief in question is true if and only if p.

Note now that neither (G) nor (H) refer to truth conditions as such. So if the import of analysis (F) is simply that (G) is equivalent to (H), then analysis (F) will be free of any substantial commitment to truth conditions too.

What we want from analysis (F) is a theory of content for beliefs. That is, we want an analysis which explains what it is for a belief to have a truth-conditional content, and which therefore gives us a recipe for determining the specific content of any given belief. But we can achieve all this without reifying truth-conditions as objects which attach to belie fs. For we can simply understand (F) as saying that claims like (H), about truth-conditional content, can always be replaced by claims like (G), about biological purposes. I shall understand (F) in this way from now on.

There is a well-known difficulty facing this kind of approach. If we take (G) at face value, and in particular don't read "if and only if" in an inadmissibly intentional way, then we ought to accept such instances as:

the belief that snow is white will fulfil its
biological purpose if and only if grass is green.

But this is surely unacceptable, if instances of (G) are supposed to amount to specifications of truth-conditional contents. For the truth condition of the belief that snow is white is certainly not that grass is green.[14]

The trouble here, as students of Donald Davidson's theory of meaning will know, is that any "standing belief", such as snow is white, is "always" true, if true at all. So we can get a true instance of (G) simply by mentioning a true standing belief on the left hand side and placing any true statement whatsoever on the right.

This is where we need to recognize the compositionality of beliefs. Instead of starting with whole beliefs, and taking analysis (F) to explain truth conditions by equating them directly with instances of (G), we need to start with the components of beliefs, such as singular concepts, predicate concepts, ways of combining concepts, and so on, and to focus on the referential values of such components, in the sense of the contributions that such components make to the biological purposes of the beliefs they enter into. Analysis (F) can then be viewed as equating truth conditions with conditions built up from such referential contributions. So now we will not construct instances of (G) directly, but only by inference from a set of assumptions about belief components, assumptions which will specify what is required for the whole beliefs those components enter into to generate successful actions. And then, since we will now be building up the (G)-claim for the belief that snow is white,

say, from assumptions about the systematic contribution that the concepts snow and - is white make to success-guaranteeing conditions across the board, we can expect to derive:

the belief that snow is white will fulfil its
biological purpose if and only if snow is white

as desired, rather than:

the belief that snow is white will fulfil its biological purpose if and only if grass is green.[15]

What now of truth itself? Those who reify truth conditions as possible states of affairs, or sets of possible worlds, or some such, can simply say that a belief is true just in case its truth condition obtains (the possible state of affairs is actual, the actual world is one of the set of possible worlds, . . .) But those of us who want to avoid reified truth conditions need to proceed more circumspectly. My current thinking on this knotty issue is that we don't need anything more to understand truth itself ap art from an ability to generate the appropriate instance of the schema (H) for any given belief. For, if we are able to do this, then we will have a recipe which tells us what is required for the belief that snow is white to be true, namely, that this belief is true if and only if snow is white; and what is required for the belief that grass is green to be true, namely, that this belief is true if and only if grass is green; and so on, for beliefs in general. And what more do we ne ed to understand truth, if we have a recipe which tells us what is required for the truth of any given belief?

This is to argue for a version of the redundancy theory of truth, according to which nothing more is needed to understand claim s about the truth of beliefs than to understand that such claims stand or fall with the claims made by the beliefs themselves. It is important, however, to distinguish sharply between the redundancy theory of truth, in this sense, and recent "deflationary" theories of truth.[16] The difference is that the redundancy theory leaves room for a substantial theory of content, a substantial theory of what determines the truth conditions of beliefs, whereas advocates of the deflationary theory argue that such substantial accounts of content are both unnecessary and misguided.

This difference between the redundancy and deflationary theories is best brought out by focusing on the question of how someone might master the ability "to generate appropriate instances of the schema (H)" -- which is how I phrased, at the end of the paragraph before last, the requirement which, according to the redundancy theory of truth, is supposed to render any further understanding of truth redundant. Deflationalists argue that it is sufficient to know that the sentence +p+ used to identify the belief that p on the left hand side of any instance of (H) should be the same as the sentence +p+ used on the right to specify the requirement for that belief's truth; or, alternatively, for versions of (H) which specify truth conditions for sentences, that it is sufficient to know that the sentence mentioned on the left of any instance of (H) should be used on the right to specify that sentence's truth condition. The redundancy theory, by contrast, is committed to no such "minimalist" account of how to generate instances of (H); it may for instance be combined, as I would combine it, with the view that the appropriate way to generate instances of (H) is to accept (H)'s equivalence with (G), and therefore to derive (H)'s instances by determining the biological purposes of the relevant beliefs.

This shows that the redundancy theory should be viewed, not as a competitor to the deflationary theory, but as something on which both deflationalists and their opponents can

agree. That is, both sides can agree that nothing more is needed to understand truth itself than a recipe which will tell you for any be lief (or sentence) what is required for its truth. Disagreement arises only on the further issue of what such a recipe need involve. Deflationists think that we need only require that the same phrase appear on the left and right hand sides of (H)-claims. Their opponents will contend that we do not have an adequate recipe for generating (H)-claims until we have a substantial theory of what determines the truth conditions of beliefs (or sentences).

On this issue there seem s to me little doubt that deflationalists are wrong. The point is clearest for the analogue of (H) for sentences. The deflationalist says that you will know how to generate the instances of (H) if you know that the sentence mentioned on the le ft hand side of any instance should be used on the right hand side. But of course this only works if the sentence mentioned is in the language you speak, so that you can use it on the right hand side. To get a notion of truth that applies to sentences in general, and not just sentences of your own language, the deflationalist needs to add that you will get an appropriate instance of (H) if the sentence used on the right hand side translates the sentence mentioned on the left hand side. However, this appeal to translation destroys the deflationalist position. For what is it for one sentence to translate another, in the relevant sense, except for them to have the same truth-conditional content? So in order to have an adequate recipe for generating (H)-claims, you will need to grasp what it is for two sentences to have the same truth condition. And it is hard to see how you can do this without a substantial account of what determines the truth conditions of sentences.

A similar point applies to the version of (H) for beliefs. The deflationary strategy works fine for beliefs already identified in terms of their truth-conditional contents, as beliefs that p. But for beliefs otherwise identified, in terms of causal relations, say, then we won't know what to put on the right hand side of the relevant instance of (H), unless we have a substantial theory of what determines truth conditions for beliefs in general.

So, while I think that the redundancy theory gives the right account of truth, I also think that this account needs to be located within a substantial theory of content.[17] The substantial theory of content I favour is in terms of success conditions and biological purposes. However, I don't necessarily want to argue that you need to embrace this specific theory of content to understand the notion of truth. For I certainly want to leave room for lay people who do not share this philosophical theory of content to understand truth. My view is that such lay people have an "everyday" or "folk" theory of content which is substantial enough to allow a satisfactory recipe for generating instances of (H), but which is philosophically inferior in various respects to the teleological theory of content. I shall not pursue this issue here, however, though I shall return to it in section 3.12 below. For the moment we can simply note that the teleological theory itself is certainly a substantia l theory of content, and so a suitable philosophical setting for the redundancy theory of truth.

One last point about the redundancy theory of truth. As I have explained it, this theory has the disadvantage that it does not provide an explicit analysis of the notion of truth. It tells us that the belief that snow is white is true if and only if snow is white, and the belief that grass is green is true if and only if grass is green, . . . But it does not analyse truth as a property that is common to these and other true beliefs.

If we build up the truth conditions for a given repertoire of beliefs recursively from semantic clauses for the components of those beliefs, then Tarski showed us how to construct a

predicate which applies to all and only the truths among those beliefs. This construction, however, makes essential use of set theory. Moreover, it only gives us a predicate equivalent to truth-in-R (where R is the relevant repertoire of beliefs), not a predicate equivalent to truth for beliefs in general. The latter problem can perhaps be solved by equating truth, not with truth-in-any-particular-R, but rather with the second-order property of satisfying-the-correct-Tarski-definition-o f-truth-in-R-for-the-R-you-belong-to.[18] There remains the commitment to set theory. Perhaps there is some way of finessing this problem too. But, rather than digressing further down this by-way, let me simply observe that those, like myself, who want to avoid commitment to sets, have the option of abandoning the quest for an analysis of a property common to all true beliefs, and simply settling for what the redundancy theory does undoubtedly give us, namely, knowledge of what is required for any given belief to be true.[19]

3.10 Broad Contents Revisited

Let me now return to the issue of broad beliefs, beliefs that physical identicals can differ in. As I said at the end of section 3.3, the theory of content developed in this chapter will enable us to understand why some beliefs are broad in this sense.

The best way to appreciate the issue of broad beliefs is to return to the contrast I drew between two pictures of action explanation in section 3.5 above. On the one hand were "internal" explanations, as in:

(A) 1. X desires that G
2. X believes that F will bring about G
Therefore,
3. X does F

If we focus on explanations of this kind, then it is easy to become puzzled about the existence of broad beliefs. For explanations of form (A) don't require beliefs to do anything except give a causal push to actions from the inside, as it were. And on this conception of beliefs it would indeed be puzzling that differences outside believers' heads can make any difference to what they believe.

However, as we saw, this isn't the only kind of action explanation. There are also "external" action explanations, which explain, not just means, but results:

(B) X desires G
X believes, of some behaviour, that it will
produce G
This belief is true

———————

X achieves G

Once we focus on this kind of explanation, the kind of explanation to which truth-conditional content matters, then the existence of broad beliefs and desires becomes unsurprising. Explanations of form (B) show that truth-conditions are nothing to do with internal push es. Rather, they specify the conditions required for beliefs to satisfy desires. Given this explanatory role, it is easy to understand why some beliefs should have world-dependent contents. For such broad contents will be found whenever two physically identical people are in different contexts in which different conditions are needed to ensure that some piece of behaviour satisfies their desires.

The point is clearest for explicitly indexical beliefs. Suppose Bill and Ben are physically identical, and that they both have the desire and belief that they express by "I want to be warm" and by "Running around will make me warm". Then they are both likely to start performing the same bodily movements, namely, running around. And to this extent their beliefs are the same: both beliefs "push from the inside" in the same way. But note now that the conditions that will satisfy their respective desires are different: Bill's desire will be satisfied by Bill getting warm, whereas Ben's desire will be satisfied by Ben getting warm. And because of this the condition required for Bill's and Ben's actions to succeed will be different: Bill's action will succeed just in case Bill 's running around will make Bill warm, whereas the success of Ben's action requires the quite different condition than Ben's running around will make Ben warm. And that is why the truth conditions of Bill's and Ben's beliefs are different, despite their physical identity. It is simply due to the fact that Bill's and Ben's actions have different success conditions.

This kind of explanation of broadness is not restricted to explicitly indexical beliefs. It will apply whenever the satisfaction conditions of the desires of two physical identicals are different; for then, as above, the truth conditions of beliefs germane to the satisfaction of those desires will be different too. Given the teleological theory of desire satisfaction, we can expect this phenomenon to be widespread, even in the absence of explicit indexicality: for the processes which select desires, in genetic evolution and in individual learning, will often select desires because of certain environment-dependent effects of those desires, effects which will not necessarily be present in the different environments of physically identical doppelgangers. So, for example, our desire for water has arguably been selected by a process that favours actions that lead us to H2O; by contrast, a being on a planet with XYZ instead of H2O could not have desires which have been selected in this way. Again, it is arguable that my desire for the company of certain people, say, is the result of learning processes in which those people played an essential role; again, a being who had never met those people could not have developed these desires in this way.

An important special case of broad mental states will be those ac quired in the course of learning a public language. Here we will find mental states whose biological purpose is in essential part to enable us to conform to community usage. (Think of a child being encouraged when it speaks correctly, and discouraged when it makes mistakes.) So, for example, I may acquire a concept of arthritis, whose biological purpose is to enable me to apply the word "arthritis" as the rest of my community does. This yields another kind of reason why the mental states of physically identicals may have different contents: for somebody may be physically identical to me, and yet live in a community in which "arthritis" is used differently.

3.11 Accidental Replicas

We have just seen how the teleological theory of representation can help us to understand why supervenience is violated by broad beliefs and desires. However, the teleological theory of representation also implies that the supervenience of the mental on brain physics is violated in a far more radical way, a way which is widely regarded as constituting a reductio ad absurdum of the teleological theory.

This more radical violation of supervenience arises because the teleological theory makes representational content depend on selectional history. The content of your beliefs and desires depends, according to the teleological theory, on what purposes they were selected to fulfil.

So it follows that another being could be physically identical to you, and yet not share your representational states, because it did not share a similarly structured selectional history.

Imagine, to make the issue graphic, that you have a physically identical doppelganger who does not have any selectional history at all, but who simply coagulated out of passing molecules a few moments ago, in some massive cosmic coincidence. Then, according to the teleological theory, this doppelganger will not share any of your con tentful beliefs and desires, despite sharing your physical make-up, since none of its brain states have been produced by any selection processes. And this seems absurd to many philosophers.[20]

An initial point that might be made on behalf of the teleological theory is that a failure of mind-brain supervenience as such can scarcely refute the teleological theory. After all, the example of broad beliefs and desires already shows that the possession of contentful states will often require certain kinds of context and history, as well as certain kinds of brain states. So why is it at all surprising that your accidental replica should lack contentful states? Of course, if we still upheld the philosophical view, which was widespread before the recognition of broad beliefs, that differences outside the head cannot matter to mental make-up, then the accidental replica would be a knock-down refutation of the teleological theory. But, as it is, why not s imply accept that the accidental replica is another being whose idiosyncratic background gives it states with different contents to ours?

However, this reply is less than entirely persuasive. The existence of broad beliefs can be de fended on independent grounds, by appeal to pre-theoretical intuitions which owe nothing to the teleological theory of representation. Because of this, the teleological theory is confirmed by its ability to explain of the existence of broad beliefs. However, there are no such pre-theoretical intuitions which show that an accidental replica does not have any contentful states at all;[21] indeed, as I said, most philosophers view this implication as intuitively absurd. So this implication, unlike the existence of broad beliefs, counts against, rather than in favour of, the teleological theory.

Perhaps defenders of the teleological theory can contest the awkward intuitions about the accidental replica. Intuitions a bout complicated counterfactual situations are notoriously insecure. Can we be sure, when we imagine your accidental replica, that we are really imagining a purely accidental being, rather than one that has somehow been designed, if not by natural selection, then by some supernatural power (such as an omnipotent philosopher who is able to create beings as required to illustrate philosophical points)? If we were imagining such a designed being, then the intuition that it has contentful states would be no problem for the teleological theory, for designed states have purposes and so teleological contents. Conversely, if we really are imagining an accidental being, then perhaps we ought therewith to relax the intuition that it has contentful states, which would again let the teleological theory off the hook.

I shall not develop this line of argument any further, however. For, even if we allow that intuition can somehow simultaneously guarantee both that an imagined being is genuinely accidental and that it has contentful beliefs, there is still a natural way to defend the teleological theory. A defender of this theory can simply point out that the theory is intended as a theoretical reduction of the everyday notion of representational content, not as a piece of conceptual analysis. And as such it can be expected to overturn some of the intuitive judgements we are inclined to make on the basis of the everyday notion. Consider, for example, the theoretic al reduction of the everyday notion of a liquid, to the notion of the

state of matter in which the molecules cohere but form no long-range order. This is clearly not a conceptual analysis of the everyday concept, since the everyday concept presuppose s nothing about molecular structure. In consequence, this reduction corrects some of the judgements which flow from the everyday concept, such as the judgement that glass is not a liquid.

This appeal to the idea of a theoretic al reduction might strike some readers as an ad hoc response to the problem of the accidental replica. But this reaction would be unreasonable. For it should have been clear from the start that, if the teleological theory of representation is acceptable at all, it must be as a reduction, not a piece of conceptual analysis. After all, there is clearly nothing about the natural selection of brain states in the everyday notions of beliefs and desires.

Perhaps the teleologic al theory of representation will one day become part of our everyday concept of representation. By way of analogy, consider the aetiological theory of teleology itself. When, in the nineteenth-century, biologists first started to understand biological functions in terms of their Darwinian aetiology, this was inevitably a matter of theoretical reduction, rather than conceptual analysis, since the requisite Darwinian notions were simply not available to pre-Darwinian biological thought. Bu t it is arguable that in the intervening years Darwinian ideas have come to penetrate the concept of function itself, with the result that, to biologists, function now just means: effect for which some trait has been naturally selected. (Cf Neander, 1991a.)

This process, of new theoretical ideas being absorbed into old concepts, is a common enough upshot of the general acceptance of a theoretical reduction. So, as I said, perhaps one day we will all intuitively think of representation in teleological terms. At which point the accidental replica will cease to be a problem, for our intuitions will then come to tell us that its internal states do indeed lack representational contents (provided, that is, that we succeed in imagining a being who is genuinely accidental). However, all these conceptual changes will happen, if at all, only after the teleological theory of representation has won general acceptance. So for the time being advocates of this theory will do better to rest their case on the arguments for theoretical reduction.

3.12 Empirical Evidence for the Teleological Theory

At this point it might occur to some readers to ask: what exactly is the case for the theoretical reduction of representation to teleology? Normally theoretical reductions are supported by empirical evidence. When chemists established that water was H20, for example, they adduced a body of empirical evidence which showed that the extensions of "water", as used by most people, and "H20", as used by the chemists, were in close agreement. So, by analogy, the teleological theory of representation ought also to be supported by empirical evidence, in particular evidence which shows that the teleological theory's ascriptions of content coincide with those made by everyday psychology. But where is this evidence? What grounds have I offered for believing that the everyday desire for r will in fact turn out to have been selected to produce r, rather than s, or nothing at all, or the everyday belief that p will turn out to have been selected to be co-present with p, rather than q, or whatever?[22]

In this respect the teleological theory of representation is worse off than those other theories, discussed in 3.8 above, which agree that truth is the guarantee of desire satisfaction, but then explain desire satisfaction in terms of extinction of desires or reinforcement of behaviour.

For, whatever other difficulties these theories may face, they can at least make a plausible case that they are part of everyday thinking about representation.

The teleological theory of representation, by contrast, needs to be defended as a theoretical reduction, not as a piece of conceptual analysis. So its defenders need to produce empirical evidence that its ascriptions of content coincide with those made by everyday psychology.

I think that they can meet this challenge. But first, before explaining the solution, let me say a bit more about the problem.

Defenders of the teleological theory obviously need to recognize that everyday thought embodies a working notion of representational content, which is available prior to any analysis of representation which the teleological theory may offer. After all, everyday thinkers who are quite ignorant of the teleological theory are able to ascribe beliefs, desires, and other contentful states to people, and by and large they are able to agree with each other in such ascriptions.

I take it that such ascriptions are informed by a body of folk psychological assumptions. These will include such general principles as that people act in ways which their beliefs indicate will satisfy their desires; that the truth conditions of belief are conditions which actually produce the satisfaction of desires; that the satisfaction of desires will often, if not always, lead to their extinction, and to the reinforcement of the behaviour by which they were achieved; and so on. These general principles will be supplemented by some more piecemeal truisms, such as that people can normally see what is in front of them, that they normally mean what they say , that they can remember what happened yesterday, that they will desire what they previously desired in similar circumstances, that they will be thirsty if they have had nothing to drink for days, and so on.

Together this body of everyday knowledge constitutes an implicit grasp of representational notions, a grasp that enables everyday thinkers to ascribe beliefs and desires with specific contents to people. The teleological theory should be understood as offering a deepening and refinement of this everyday understanding. It deepens everyday understanding, as do all theoretical reductions, by giving us fuller information about the nature of the reduced phenomenon, information which takes us beyond the surface features by which the phenomenon is normally identified, to the underlying features which explain those manifest appearances. And it refines everyday thinking by adding precision to our assumptions about representation and the propositional ascriptions they inform.

Let me say a bit more about the way the teleological theory refines everyday thinking. It is an implication of the arguments earlier in this chapter that the general assumptions of everyday psychology do not by themselves yield complete determinacy in ascriptions of propositional attitudes. I pointed out in section 3.6 that the assumption that truth guarantees satisfaction places a joint constraint on ascriptions of truth and satisfaction conditions, but that, without some further account of desire satisfaction, this constraint can be satisfied by deviant permutations of normal ascriptions of truth and satisfaction conditions. And I argued that the everyday idea that satisfaction extinguishes desires, or the idea that it reinforces the means which achieved them, are not adequate to fill this gap, since they fail to apply to desires in general. In practice everyday thought no doubt fills much of this gap by appeal to such piecemeal rules as that people will desire what they previously desired, that they will be thirsty if they have had nothing to drink for days, and so on. But we can expect that, even so,

there will be certain cases where everyday thinking is unable to decide about the content of certain desires, nor, therefore, of the beliefs which inform their pursuit. And in these cases the teleological theory of representation will be able to make determinate what everyday thinking does not. Imagine, for instance, a woman who has a recurring desire which leads her to visit a certain spot in a park. She is not sure why she does this; it could be for the flowers, or the restful atmosphere, or various other reasons. There might be nothing in everyday psychology to determine the content of her desire. But there will still be a fact of the matter as to which previous effect of this desire has led its being preserved, and the teleological theory will fix on this on the content of the desire.

This would be a case w here the teleological theory fills a gap left by everyday thinking. There is also the more extreme possibility that the teleological theory may actively overturn ascriptions of content made by everyday psychology. The accidental replica discussed in the last section is one example of this. And we can imagine other, more mundane, cases in which everyday psychology's ascriptions of content do not tally with the selectional history of the relevant states, and so are deemed wrong by the teleological theory. However, to return to the main issue to be addressed in this section, cases like this had better be the exception rather than the rule. For, before the teleological theory can start overturning everyday judgements, we need some evidence that it is an acceptable theoretical reduction in the first place, and this requires, as pointed out earlier, reason to suppose that the teleological theory agrees, if not in every case, at least in most of the prior ascriptions of propositional content made by everday psychology.

The complaint made at the beginning of this section was that as yet we seem to have no evidence for such agreement. Let me now face up to this challenge. My strategy here will be to appeal to the argument of chapter 2 to provide the requisite evidence. In that chapter I argued that it would be incredible that special-scientific properties should be variably realized, unless their instances are the product of some selection mechanism. I think that this line of argument will serve to answer our present difficulty. For it implies that it would be incredible that human beings should conform to the assumptions made by everyday psychology, unless their beliefs and desires had been selected by processes which give them purposes corresponding to their contents.

Before going into details, it is probably worth clarifying the sense in which this argument provides empirical evidence for the coextensionality of the teleological theory's and everyday psychology's ascriptions of content. This relates to the point, originally made in section 2.2, that the "incredibility" of variably realized special-scientific laws without a teleological underpinning is an empirical matter: the objection to such laws is not just that they offend brute intuition, but, more importantly, that they run counter to the wealth of experience which testifies to the general principle that uniform physical patterns have uniform physical explanations.

To see how these considerations help with the particular problem at hand, recall generalization (E):

(E) Actions based on true beliefs will satisfy the
desires they are aimed at.

Now consider an instance involving a desire for some specific physical result, r, like getting hold of an ice-cream:

(I) Agents who act on true beliefs and the
desire for ice-cream will get some ice-cream.

Note that (I) specifies a uniform physical state in the consequent. Yet the antecedent conditions -- desiring ice-cream, and being a true-believer -- are presumably not themselves uniformly physically realize d. So we might well ask, "Why do these all physically different antecedents have a uniform physical effect?"

This was just the kind of question we asked in chapter 2. And the answer we gave there was that in such cases there will always be a selection mechanism which selected the physically disparate instances of the antecedent because they produce the common effect. In fact chapter 2 has already applied this analysis to the specific issue of variably realized desires for ice cream, and argued that the reason the different physical realizations of the desire for ice-cream all lead to the ingestion of ice cream is that this is why they were selected in the first place.

This observation now provides an answer to the question of why we should expect the teleological theory to agree with everyday psychology in ascriptions of content, at least in respect of desires. The answer is simply that it would be a mystery that the desire for some physical result r should do what everyday psychology says it does, as in (I), unless it has been selected to produce r.[23]

The corresponding point about true belief is more interesting. Since generalization (I) generalizes across belief types, not requiring that the agent have any specific beliefs, but just that the agent's beliefs, whatever they are, be true, the "true belief" requirement in the antecedent will be variably realized by the truth of different belief types. Thus, being a true-believer can be realized by: believing that the shop is open and the shop being open; or believing that there is ice-cream in the shop and ice-cream being in the shop; or believing that an ice-cream is within reach and an ice-cream being within reach; and so on. In different cases, different external conditions are required for an agent to be a true-believer, and so for the agent's behaviour to lead to the desired result. And so now we have this version of the variable realizability puzzle: why do all the quite different conditions required for different beliefs to be true all lead, when conjoined with the possession of those beliefs, to the desired result?

And the solution, once more, is that a mechanism has selected those conjunctions of condition and belief precisely because they produce such results. To be more accurate, we should think of the relevant mechanisms as selecting dispositions to form-certain-beliefs-when[24]-certain-circumstances- obtain: for instance, the disposition to form-the-belief-that-an-ice-cream-is-within-reach-when-an-ice-cream-is-within-reach. And the reason why different exercises of these disparate dispositions on different occasions will nevertheless all produce the same result, as required by (I), is that these dispositions will have been selected precisely because of the kind of effect the relevant beliefs have when their associated circumstances obtain. It's the conjunction of beliefs and their truth conditions that ensures success, and so it's dispositions to form beliefs in conjunction with their truth conditions that is selected.

And this now show us how to answer the challenge of this section in connection with belief contents , analogously to the way we answered it for desires. The teleological theory must match everyday psychology on ascriptions of belief contents because, as before, it would be a

mystery that beliefs that p should do what everyday psychology says they do, as in (I), unless they had been selected to be present when condition p obtains.

Perhaps we could have reached this conclusion by a shorter, if less illuminating, route. In the course of this chapter we have had occasion to note that our actions are directed by two kinds of mental states, beliefs and desires: desires have ends attached, and vary over time in ways attuned to our needs, while beliefs tend to "track" specific external conditions; and these beliefs and desires then combine to cause behaviour which causes those ends if those conditions do obtain. Now, this carefully orchestrated arrangement could scarcely have arisen by chance. If this is really how our psychology works, then surely it must have been designed for that purpose -- not by a conscious design er, of course, but by the blind selection mechanisms of learning and evolution. (Cf Millikan, 1989a, pp 292-4.) So once more we have prior reason to think that beliefs and desires must have been selected for purposes corresponding to their contents, as the teleological theory of representation claims.

3.13 Verificationism Refuted

There is an important general moral to be drawn from the argument of this chapter, a moral which will be central to the epistemological arguments in the third part of this book. Namely, that the teleological theory is radically at variance with verificationist analyses of meaning which imply a conceptual tie between the truth conditions of judgements and the conditions under which those judgements are asserted. For there is nothing in the teleological theory of representation, when properly understood, to imply that there should be any definite correlation between the circumstances in which we are inclined to form beliefs, and those in which those beliefs are true. The reason is that truth-conditional content, for the teleological theory, hinges on the results of beliefs, not their causes. In particular, the teleological theory identifies truth conditions as those circumstances in which the actions prompted by a belief cause the satisfaction of desires. These are not the same circumstances as those which lead us to adopt the belief. And there is nothing in the teleological theory to imply any special link between these two sets of circumstances.

It is true, of course, that there will generally have been some biological pressure in favour of belief-forming processes which tend to yield true beliefs, since true beliefs ensure the satisfaction of desires, and in general the satisfaction of desires is biologically advantageous. But this link is easily disrupted. Most obviously, there is the point that our natural inclinations to form beliefs will have been fostered by a limited range of environments, with the result that, if we move to new environments, those inclinations may tend systematically to give us false beliefs. To take a simple example, humans are notoriously inefficient at judging sizes underwater.

Rather more interesting are cases where our systematic tendencies to false belief are themselves the upshot of biological design, rather than simply the result of changed environments . One illustration of this possibility is the belief about immunity to injury discussed in section 3.4. In cases of this kind the normal biological pressure in favour of true beliefs is counterbalanced by a contrary biological pressure, which encourages us to form the belief about immunity even when it is false, so as to get us to fight and win. And there are many other similar[25] cases in which biological pressures produce systematic inclinations towards false beliefs. These false beliefs then lead us to act in ways that frustrate our desires, but tend to further our biological needs. And in consequence the circumstances in which we form such beliefs will be systematically different from those which make them true, for truth conditions are tied to the satisfaction of desires, rather than biological needs.

Verificationists might feel inclined to respond that these observations are beside the point, on the grounds that verificationism only asserts a tie bet ween truth and normative assertion conditions, not actual ones. What matters is when people ought to assert claims, not when they do. So biological demonstrations that people often do assert false claims are beside the point. (After all, verificationists can point out, a distinction between "canonical" assertion conditions and actual practice has always been implicit in verificationist thinking, for without some such distinction verificationism will fail to leave any room for false judgements.)

However, I don't think that an appeal to this kind of distinction can save verificationism from the biological facts. For once we allow the kind of radical gap between truth and assertion that is implied by biology, then verificationists face the problem of providing some independent grounding for assertoric norms. It is one thing to allow, say, that individual assertoric practice sometimes falls out of step with the majority line. For then the majority will provide the norm for individual practice. But if verificationism accepts that there can be judgements which nearly everybody gets wrong nearly all the time, then what basis is left for the thought that nevertheless there are agreed standards of correct judgement which are conceptually tied to the truth?[26]

Of course, there is one way of construing "assertoric norms" which will create a conceptual link between truth and normative conformity -- namely, we can equate such norms with whichever judgemental procedures will lead us to the truth, and then giving some independent analysis of truth. This is how I myself think of assertoric norms, with the independent analysis of truth being provided by the teleological theory of representation.

But this is not verificationism. Verificationism aims to proceed in the opposite direction, by given some self-standing account of assertoric norms, and then defining truth in terms of the satisfaction of such norms. The normal basis for such a verificationist account of norms is the actual assertoric practice of the community. My point is that this route ceases to be available once verificationism concedes to the teleological theory that the whole community can usually be wrong.

1. I am interested in representation as a problem for physicalism. It is worth observing, however, that the problem is scarcely peculiar to physicalism. Even dualists, for example, have an obligation to explain how their special mind-stuff can stand for other things. Not that they have always recognized this problem, no doubt because their mind-stuff had so many special powers anyway -- such as the ability to exist outside space but in time, to be transparent to itself, and so on -- that one more special power scarcely seemed worth worrying about.

2. Cf Fodor (1990, pp 63 ff).

3. Versions of this teleological approach to mental representation are found in Dennett (1969, ch 9; 1987, ch 8), Fodor (1984), Millikan (1984, 1986, 1989a), Papineau (1984, 1986b, 1987), McGinn (1989, ch 2). Fodor has since recanted. He now holds (1990, Ch 3) that the teleological approach fails to solve the disjunction problem. He says that there is nothing in teleology to tell us that a frog's fly detector, say, represents flies rather than flies-or-any-other-small-black-moving-objects, since a properly working detector will respond to any small black moving thing . But Fodor is here assuming that the purpose of the fly detector is fixed by what causes it, rather than by what it is supposed to cause. However, as I shall stress

in what follows, biological purposes are always a matter of results. In particular, the purposes of beliefs are to get the organism to behave in a way appropriate to certain circumstances. This is why the frog's detector registers flies: the frog's states cause the frog to behave in a way appropriate to flies, and not just to any small black inedible dots. (Why flies, rather than food, or survival, or gene perpetuation? This is a different question, about a "vertical" indeterminacy which is orthogonal to the "horizontal" indeterminacy of the disjunction problem. I shall answer it in footnote 8.)

4. This arguably oversimplifies the example somewhat (cf footnote 25 below). But for the time being it will be helpful to sacrifice biological realism to explanatory convenience.

5. While this gives us one sense in which desire satisfaction is prior, there are other senses in which the representational powers of beliefs and desires are mutually dependent. The sense in which desire satisfaction comes first is this: the biological aim of desires is not (except in special cases) to produce true beliefs, but the biological aim of beliefs is standardly to satisfy desires. However, this is consistent with the point that desires always act in concert with beliefs when prompting actions, just as much as vice versa, and therefore that any desire fulfilling its biological purpose will depend on beliefs fulfilling their biological purposes too. Moreover, because of this, we can expect desires and beliefs also to be psychodeveloped mentally interdependent, each category becoming differentiated as a distinct psychological state only when the other is.

6. Ruth Millikan (1984, 1989b) uses the phrase "proper function" for those effects of biological traits which they have been selected to produce (that is, for the aetiological notion of "function" or "purpose"). It is perhaps worth observing that in this sense both the "normal" and "special" purposes of beliefs are "proper functions".

7. Dennett himself favours a selectionist account of representational content (1969, 1987). However, he seems not to have noticed the tension between this and his non-realism about beliefs and desires.

8. These points about our human structure of beliefs and desires now answers the question raised at the end of footnote 2, and explains why our different desires have different satisfaction conditions, rather than all being aimed alike at the ultimate evolutionary end of gene perpetuation. For while it is true that the biological purpose of all desires is in the end to foster gene bequests, different desires have been designed to foster this end in determinately different ways. This shows up in the fact that the desire for chocolate, say, doesn't disappear when you accept that eating more chocolate won't help you pass on your genes. The appropriate way to think of the purpose peculiar to a given desire is as that result the desire will lead us to pursue whether or not we believe that result is a means to further ends. For further discussion of this point, see Papineau (1987, sect 4.3).

9. Dennett's non-realism about belief-desire psychology makes him think that it is absurd to suppose that empirical discoveries might show that we don't have beliefs and desires (1987, p 233-235). I agree that this is absurd. But this is not because I agree with Dennett that belief-desire psychology is non-theoretical and so somehow insulated from empirical evidence. Rather, I think it is theoretical, but already established by a wealth of evidence.
10. I owe the argument for this principle to Horwich (1990).

11. This response to the objection about uncertainty was suggested to me by Hugh Mellor.

12. These examples were put to me by David Owens and David Sanford respectively.

13. Whyte (1990) aims to deal with this problem by arguing that the causal roles by which we ordinarily identify beliefs, and which then fix their success conditions, happen specifically to involve valid rather than invalid inferential moves. I agree that common sense psychology regards the valid implications of beliefs as constitutive of those beliefs, by contrast with any characteristic tendencies to generate invalid conclusions. But I think this is because ordinary thought identifies beliefs by their truth conditional contents, and then helps itself to the idea of those consequences which validly follow. This means that any attempt to reduce truth conditional content cannot appeal to common sense psychology's view that certain inferential consequences are constitutive of the identity of beliefs. For these sets of constitutive consequences cannot be characterized without the notion of semantic validity.

14. Why shouldn't we read the "if and only if" in an intentional way? Well, if we read (G) as saying that ". . . the belief . . . will fulfil its biological purpose in all possible worlds where p", this will solve the snow is white/grass is green difficulty, but only at the cost of introducing possible worlds. It is true that an explicit reference to possible worlds is only one possible way of analysing ". . . the belief . . . will necessarily fulfil its biological purpose if and only if p". In chapter 6, however, I shall argue that, whether or not we adopt the possible worlds analysis of modality, modal judgements cannot be viewed as legitimate expressions of belief, and so are ineligible for essential roles in our best theories. So I prefer to solve the snow is white/grass is green difficulty without using modal notions.

15. Donald Davidson's approach to meaning (1984) can also be viewed as offering a kind of analysis of truth-conditional claims like (H) (rephrased to apply to sentences rather than beliefs), through not a reductive analysis, as above, but rather an implicit analysis, via an explanation of how to test an empirical "meaning-theory" which specifies (H)-claims for all a community's sentences. (For an exposition of this interpretation of Davidson, see Papineau, 1987, sections 2.4-8). This Davidsonian approach to truth-conditional content has extra difficulties with the "snow is white/grass is green" problem, however. For, while the problem can still be solved, given strong enough requirements about the need for "meaning-theories" to derive their (H)-claims from separate assumptions for sub-sentential components, it is unclear how to motivate these requirements within the Davidsonian programme. From my perspective, this is not a difficulty. I take a realistic view of belief components and their referential contributions to biological purposes. So I don't need any independent justification of the compositionality requirement, of the kind essayed by Davidsonian theorist s, such as that native speakers, or perhaps meaning theorists, need to derive their knowledge of an indefinite number of (H)-claims from a finite amount of sub-sentential information. From my point of view such doubtful appeals to the preconditions for knowledge of meaning-theories are irrelevant, since I take the compositionality requirement to be a direct upshot of the semantic facts, irrespective of whether or not any native speakers, or meaning theorists, know a theory of those facts.

16. Fo r the redundancy theory, see Ramsey (1927). Deflationary theories are defended in Quine (1970), Leeds (1978), Horwich (1982, 1990); for a general discussion, see Field (1986).

17. From this perspective, the redundancy theory can also be viewed as consonant with the idea that truth involves correspondence with the facts. Those who adopt the redundancy

approach to truth will not, of course, want to explain truth in terms of possible facts "obtaining". But it seems natural to say, given the redundancy theory, that when the belief that snow is white is true, for instance, this is in virtue of the fact that snow is white. This doesn't explain truth in terms of facts, but rather introduces facts as what make true beliefs true; still, when a belief is true, there will be a corresponding fact. There remain questions about the "thickness" of any such fact; are there any other reasons, apart from the truth of the corresponding judgement, for recognizing the fact, such as, say, its causal significance?

18. Since this definition generalizes over Rs, the "adequacy condition" which provides an "external" test for the correctness of Tarski-style definitions of truth-in-particular-Rs will become part of the definition of truth-in-general. This is what we should expect: it reflects the point that you cannot have a general recipe for generating (H)-claims without a substantial theory of what determines the contents of beliefs.

19. At one time I thought that we could equate truth for beliefs in general with the property of "generating actions which are guaranteed to succeed" (cf 1990, p 30). But I now think that the "guaranteed" here conceals a reference to a reified truth condition, since what we need, for the truth of any token of a belief type, is not just some actual fact that will cause the action based on the belief to satisfy desires, but, more specifically, the obtaining of that possible fact which guarantees success for all actions generated by tokens of the belief type. An analogous point would apply if we tried to equate truth for beliefs with the property of "fulfilling their biological purposes of satisfying desires"; for such fulfilment needs to be understood in terms of the general condition required for the relevant belief type to fulfil its purpose, not just in terms of any accidental route to desire satisfaction.

20. Cf Cummins (1989, ch 7); Whyte (1993). The problem of the accidental replica is also discussed by Millikan (1984, p 94). Note that the accidental replica wouldn't present a problem if we divorced the teleological theory of representation from the aetiological theory of teleology. But since I see no virtue in non-aetiological accounts of teleology, I shall not pursue this option further.

21.This is perhaps a bit strong, given that some of the arguments for broad mental states do arguably have the corollary that your accidental replica will lack some of your mental states. Thus, if the broadness of your concept of water depends upon which liquid was around when you learnt this concept, then a being that never learnt anything couldn't share your concept. Still, many other broad attitudes don't depend on learning in this way, and so there will be no pre-theoretical reason to deny them to your accidental replica. And, apart from that, plenty of your beliefs and desires aren't broad at all, and so intuitions about broadness will do nothing to explain why your replica lacks these.

22. I owe this objection to the teleological theory to a conversation with Andrew Woodfield.

23. So far this only deal with desires for physical things. But the story can be elaborated to accommodate desires for non-physical things, provided those non-physical things in turn have physical effects, by reference to which the desires in question can then be selected. A similar point applies to beliefs. Beliefs can be selected to be co-present with non-physical conditions, provided those conditions have physical effects by reference to which such co-presence can be selected. The discussion of hierarchies of selection mechanisms in section 2.8 is relevant here.

24. "When" only makes immediate sense for indexical beliefs. For standing beliefs, we need the compositionality of beliefs to give it substance. That is, we need to remember that beliefs are made of components, whose representational significance depends on their systematic contribution to the truth conditions of those beliefs, and that what gets selected, in the first instance, are therefore dispositions to deploy such components in just those cases when their contribution to the truth condition of the resulting belief will be satisfied.

25. In fact the earlier description of the immunity-from-injury case was something of an oversimplification. The real biological problem in such cases is not that there is no psychological desire corresponding to the relevant biological need (to fight and triumph), but rather that this desire is insufficiently strong in comparison with other conflicting desires (like wanting to avoid injury). Our biology then compensates by favouring beliefs that will get us to pursue such insufficiently strong desires, even on scanty evidence. (Perhaps the best-known of the many other examples of this structure is the human readiness to conclude that given foods are poisonous, thereby compensating for our biologically inappropriate tendency to let our hunger outweigh our fear of poisoning.)

26. Followers of Michael Dummett might feel inclined to argue that there must be such standards, in order for people to be able to acquire or manifest their grasp of judgements. (Cf Dummett, 1976, p 101.) But the thesis that acquisition and manifestation depend on agreed standards of correct judgement is itself undermined by the observation that there are judgements which everybody tends to get wrong.

CHAPTER 4 CONSCIOUSNESS AND THE ANTIPATHETIC FALLACY

4.1 Introduction

In chapter 1 I had occasion to mention dualism, which I characterized as the view that conscious mental states have features which cannot possibly be possessed by physical states. At that stage I argued that the price of dualism was epiphenomenalism, on the grounds, roughly, that dualism requires conscious mental states to be distinct from the physical causes of behavioural and other physical effects. But I left it as an open question whether the arguments for dualism, and in particular for its conception of consciousness, made epiphenomenalism a price worth paying.

In this chapter I want to argue that there is in fact no good motivation for the dualist view of consciousness, and that we should therefore uphold the simple physicalist position that all mental states, including conscious states, are identical with or realized by physical states. The advantage of this physicalist position is that, unlike dualism, it allows us to view conscious mental states as genuine causes of behavioural effects.

It will be convenient to use the term "physical state" in a liberal sense throughout this chapter, to include not only strictly physical states in the sense of chapter 1, but also any second-order or higher-order states which are realized by physical states.[1] The differences between these different kinds of "physical states" will not matter for most of this chapter, since most of the arguments between dualism and physicalism arise in exactly the same way whichever kind of "physical state" the physicalist identifies conscious mental states with.

At the end of the chapter, however, the difference between "first-order" and "higher-order" physical states will become relevant. So far in this book I have tended to assume that mental states are, if anything, identical with higher-order physical states, rather than with first-order ones: as I observed in chapter 1, it seems unreasonable to hold that extraterrestrials, or people with brain prostheses, cannot share thoughts with us, just on the grounds that their brains contain different kinds of molecules. With the specifically conscious features of mental states, however, the situation is somewhat different. For, as we shall see, there are persuasive arguments, based on various inverted spectrum thought experiments, for holding that such conscious features in particular depend on the physics of the brain, rather than on its higher-order organization. This is fine-tuning, however. The prior issue is whether conscious features can be identical with any kind of physical property, first-order or higher-order. After that we can worry about which kind of physical property they might be identical with.[2]

I shall proceed as follows. In the next three sections I shall consider some recent arguments for dualism. I shall argue that they are quite ineffective. Accordingly, in section 4.5, I shall ask why the intuitive pull of dualism is nevertheless so strong. My diagnosis will be that we are seduced by a fallacy, which I shall call the "antipathetic fallacy", into thinking of consciousness as something distinct from the physics of the brain. The final five sections of the chapter will then explore some of the consequences of this diagnosis.

4.2 What is it Like to be a Bat?

Much of the contemporary literature on consciousness begins with Thomas Nagel's article "What is it Like to be a Bat?" (1974). Nagel argues that conscious mental life involves certain essentially subjective facts, facts that can only be appreciated from the "first-person" point of view, from the point of view of the subject of those conscious experiences. Such subjective facts contrast with objective facts, like physical facts, which are accessible from the "third-

person" perspective, independent of any particular subjective po int of view. Nagel concludes on this basis that any physicalist account of mind must fail to account for the subjective aspect of mental life.

Nagel illustrates this thesis by inviting us to reflect on the echolocatory experience of bats. Nagel takes it that bats, like other mammals, have conscious experiences. In particular, he takes it that bats have conscious sensory experiences when they echolocate. But, he points out, we human beings are unable to adopt the bats' point of view, and so have no idea what those bat experiences are like. You might think that echolocation would be like flying about in the dark and hearing lots of high-pitched noises. But that would be what it is like for beings like us, with human perceptual apparatuses, to echolocate, and not, presumably, what it is like for bats to echolocate. And, indeed, the more we think about it, the more it becomes clear that we have no grip on the subjective nature of the bat's echolocator y experience.

Nagel focuses on bats, not because he has any doubts about bats being conscious, but rather because our complete inability to adopt the bat point of view highlights the existence of the subjective side of bat experience. In the case of other human beings, and perhaps even of chimpanzees and dogs and squirrels, we can put ourselves in their places, and imagine having experiences like theirs. And so it is easier, in these cases, not to notice that grasping the subjective aspect of experience requires us to abandon the objective, third-person perspective. But since we can't put ourselves in the place of bats, this particular example forces us to recognize that a purely objective perspective does not in fact give us any access to the subjective reality of experience.

Despite the plausibility of Nagel's line of argument, I think that physicalism can meet the challenge he poses. Let us proceed in stages. For a start, we should immediately concede that there is one sense in which we human beings are indeed cut off from the facts of bat experience. We do not have echolocatory experiences, whereas bats do. In this sense it is undoubtedly true that we "lack access to", "cannot appreciate", or whatever phrase you prefer, the "subjective reality" of bat experience. But this observation in itself clearly yields no argument against physicalism. For physicalists are just as well placed as anybody else to explain this difference between bats and humans. Physicalists think that conscious experiences are identical with certain physical events in the brain. So physicalists can say that the difference between bats, who have echolocatory experiences, and humans, who do not, is simply that certain physical events, namely, those which constitute echolocatory experiences, occur in bats, but not in humans. In this sense the physicalist can happily agree that bats have access to experiences which humans cannot appreciate.

This point is central to the physicalist view of conscious experience. Physicalism does not deny that there are conscious experiences, nor, if you wish, that "that it is like something to have them". The claim is only that this is nothing different from what it is to be a physical system of the relevant kind. Of course there is something it is like to experience pain, or to see red, or to taste cheese. And such things are highly important, especially for the subjects of those experiences. But, insists the physicalist, they are not non-physical things. What makes it like that for you is that you are you, that is, that you are a physical system of a certain sort. If you were physically different in the relevant respects, things would be different for you.

This is the initial physicalist response to Nagel's challenge. There is, however, a more persuasive line of argument suggested by Nagel's position. Suppose that you did somehow come to have echolocatory experiences. Wouldn't you then differ from other human beings,

not just in having had those experiences, but in then knowing something that other humans didn't know, namely, what echolocatory experiences were like? And wouldn't this be knowledge of an essentially subjective fact? For note that you could have known everything there is to know about bat experience from an objective point of view -- you could have been an expert on bat echolocation, who knew all about the physics and physiology and computational workings of the bat's brain -- and yet, prior to having had the echolocatory experiences, you would not have known what they were like. So it seems to follow that after you have had an experience you acquire knowledge of certain facts -- the subjective, phenomenal features of the experience -- which are necessarily omitted by any objective, physicalist story.

4.3 What Mary Didn't Know

This "knowledge argument" is developed explicitly by Frank Jackson in "Epiphenomenal Qualia" (1982) and "What Mary Didn't Know" (1986). Jackson simplifies the issue by focusing on a case where it is mere happenstance, rather than the wrong cognitive apparatus, that prevents somebody having certain experiences. He tells the story of Mary, who is an expert on the psychology and physiology of human colour vision. Mary knows everything there is to know about the goings-on in our brains when we when we see red, say. However, Mary has always lived in a restricted black-and-white environment. All the objects she has ever seen are black or white or grey. She has never herself seen anything red. Then one day she is presented with a red object. She then has the experience of seeing something red. And as a result she learns something she didn't know before. She now knows about the phenomenal nature of red colour experiences, when before she was ignorant of this. Remember, however, that Mary had always possessed complete objective information about colour experiences. So, once more, it seems to follow that there are items of information about experience that must be omitted by any physicalist account.

This "knowledge argument" adds an extra dimension to Nagel's original defence of subjective facts. But there is still plenty of room for physicalism to resist it. The natural physicalist response to this argument is to admit that there are indeed before-and-after differences in Mary, consequent on her having had her first experience of red, but to deny that these involve her becoming acquainted with some subjective feature of colour experience. There are other ways of construing the changes in Mary, which do not require the postulation of such subjective facts, and which do not therefore imply that a physicalist account of experience must be incomplete.

In this section and the next I shall outline a physicalist construal of these changes. In this section I shall consider changes in Mary's recreative powers of imagination and recall, and in her ability to reidentify her experiences. In the next section I shall consider changes in her concepts of experience. This will involve retreading some relatively familiar philosophical ground. But my aim is not just to block Jackson's argument -- other philosophers, referred to below, have already shown how to do that -- but rather to point t o a striking common feature of the experientially produced changes in Mary, namely, that they all yield ways of thinking about experiences that deploy versions of those same experiences. This point will be central to my subsequent diagnosis in section 4.5 of the "antipathetic fallacy" which I take to be responsible for the intuitive pull of dualism.

The first before-and-after change to be considered concerns Mary's new powers of recreation. Once she has seen red, Mary can recreate the experience of seeing red, in imagination and memory, whereas before she couldn't. Mary could of course always imagine, in the third-

person, so to speak, that somebody else was seeing red, in the sense that she could imagine such-and-such physiological or behavioural occurrences in that person. And, similarly, she was always able to remember, in the third-person again, that somebody had seen red. But now she has a new ability, the ability to imagine or recall having the experience itself, from the inside, as it were. She can now relive the experience, as opposed to just thinking about it.

An anti-physicalist like Jackson can account for this change in terms of Mary's new knowledge of a non-physical fact. When Mary experiences red, on this anti-physicalist account, she discovers that the experience has a characteristic phenomenal feature P. And then, because she has this new knowledge, she can imagine the experience by entertaining the thought that someone has an experience with property P. Similarly, she is now able to recall the experience by remembering that she herself had an experience with property P.

Physicalists will offer an alternative account. Suppose that the kind of imagination and memory at issue depends on the brain literally recreating a version of the experience being imagined or remembered. That is, suppose that first-person imagination or memory requires that brain be in a state which is similar to the state constituting the original experience. It won't be exactly the same state, since imagining or recalling a pain is different from having a pain. But it could well be a similar state, a kind of faint replica, which would fit with the fact that an imagined or remembered pain shares to some slight extent the unpleasantness of a real pain.

This alternative suggestion yields as good an explanation of the fact that you can only imagine or recall experiences you have previously undergone as the theory which postulates new knowledge of phenomenal property P. For it seems highly plausible that the brain's ability to recreate an experience depends, as a matter of empirical fact, on its having at some time had an original version of that experience, to give it, so to speak, the mould from which to make the replicas.[3]

What is more, this alternative account of Mary's new ability is quite consistent with an objective, physicalist account of conscious experiences. For on this alternative account the difference produced in Mary by her original experience of seeing red is not that she acquires some new item of knowledge, but simply that she can now do something she could not do before, namely, recreate that experience in imagination and memory.[4] The earlier account, which attributed new knowledge of phenomenal property P to Mary, implied that her previous third-person information about the experience left something out. However, since the new account does not credit Mary with any such new knowledge, there is now no implication that a physicalist account of conscious experience is incomplete.

Some readers may feel that this physicalist account of first-person imagination and memory is an ad hoc theory whose only attraction is that it saves physicalism. But this would be unjust. For the account also has the positive virtue, noted in passing above, of offering some explanation of why an imagined or remembered experience resembles the original experience itself -- namely, that such imaginings and rememberings literally involve a copy of the original experience.[5]

D.H. Mellor uses the term "secondary" to refer to this kind of copied experience, the kind of experience which occurs when we recreate in imagination or memory those primary experiences we have previously undergone.[6] The existence of such secondary experiences

which resemble their primary versions will be central to my eventual explanation of the antipathetic fallacy.

Of course this talk of "resemblance" between secondary and primary experiences needs further elaboration, both to specify what kind of replication i s involved, and to explain how the resulting replicas mimic the original experiences in our cognitive workings. But I take it to be uncontentious that there is some phenomenon of resemblance here, and that the model of "secondary" replicas of primary experiences offers a promising route to an explanation.

So far I have considered the new recreative powers of imagination and recall produced by Mary's first experience of seeing red. Another such before-and-after change is that Mary acquires a new introspective power to reidentify that experience when she has it again. Mary of course always had the ability to recognize "from the outside" when somebody was seeing red, from environmental or behavioural or physiological evidence. But now she has a new ability, to recognize, by direct introspection, that she herself is seeing red.

Again, one possible explanation of this new first-person ability would be that Mary discovers that the experience of seeing red has phenomenal property P, and that as a result she can now pick out experiences with property P as instances of seeing red. But, as before, this is not the only possible explanation of the new ability. For we can suppose instead that Mary simply acquires a non-conceptual "template", in David Lewis's phrase[7], which can then be compared directly with further experiences, and cause Mary to believe that she is experiencing red again. She doesn't arrive at this belief by noting that the experience has property P, and concluding that it is an experience of seeing red. There is simply a mechanism in her brain which compares the experience with the template and yields this belief directly.
As with the "secondary experience" account of imagination and memory, the "template" account of introspective recognition both yields a plausible account of why we should need the original experience in order to acquire the recognitional ability (namely, because the brain ne eds the original to have the materials from which to form the template)[8], and remains consistent with physicalism (since it doesn't explain Mary's new recognitional ability by attributing knowledge of phenomenal property P to her, but simply by postulating a new mechanism in her brain).

It is an interesting further hypothesis that the same cognitive operations may be involved both in recreative and in recognitional abilities. Perhaps the brain uses the processes constituting our "secondary experiences" themselves as the "templates" by which it classifies new experiences: that is, perhaps its mechanism for recognizing such new experiences is simply to compare them with the replicas which are activated in imagination and recall. It does not seem inevitable that things should work like this: there is no contradiction in the idea of beings who could classify new experiences by some template process, and yet lacked the ability to recreate those experiences in imagination or memory; and perhaps it is even possible for there to be beings who could recreate experiences, but who lacked the second-order mental ability to classify them. But it seems clear that in human beings the two abilities always go together, and the natural explanation is that they do so because the same mechanism subserves both.

4.4 Concepts of Experience

The overall argument of the last section can be put as follows. In so far as Mary's first experience of red leads to her knowing something she didn't know before -- leads to her "knowing what the experience is like", if you want to put it that way -- this new knowledge

can be construed as her knowing how to do something new, rather than as her knowing that anything new. There are indeed genuine changes produced by Mary's new experience. But these changes are all a matter of her acquiring new abilities -- to recreate or recognize the experience -- not of her forming any new kinds of judgements about the world.

But can this be the full story? Surely, many readers will feel, new experiences doesn't just give us the abilities described in the last section. They also enable us to think new thoughts. Once you've seen red, then can't you think of that colour, and judge it to be vibrant, or threatening, or something everybody should experience at least once, in a way you couldn't before?

I agree. But I think that this too can be accommodated by physicalism. The important question for physicalism is whether new experiences lead to our knowing about any new features of the world. Physicalists need to deny this. But they can consistently allow that new experiences lead to our acquiring new concepts for thinking about those features. In Fregean terms, the change would be at the level of sense, not reference. Mary's thinking about the experience of seeing red would change, but what she was thinking about would be exactly the same thing as she used to think about when she was a scientist who had never herself seen red.[9]

In order to bring out this point, it will be helpful to switch examples slightly for a moment, and consider, not Mary, but Jane, let us call her. Jane has always shared Mary's black-and-white environment. But Jane is no expert on colour vision. Indeed she has never heard of such things as colours and of people experiencing them.

Then one day Jane sees something red. Unlike Mary, she does not have available any public concept of the visual events that take place in people when they are presented with red objects. Indeed she may not even realize that the sensation she is currently experiencing is caused by some observable feature of her environment. Yet we would surely expect Jane to be able thereafter to form beliefs about that sensation, such as that it was vibrant, threatening, something everybody should experience it at least once, and so on.[10]

Mary, on the other hand, does have a public concept of the visual events that take place in people when they are presented with red objects. But, despite this, there seems no question but that Mary might acquire just the same kind of new thoughts as Jane does after experiencing red for the first time. For imagine that, the first time Mary experiences red, she does not know what it is -- she simply is not aware that the curious experience she is now having for the first time is the experience that is characteristically caused by red objects. In this case Mary will surely respond just like Jane, and start forming beliefs such as that this new experience is vibrant, threatening, and so on.

At first sight this might seem to substantiate Jackson's knowledge argument. Doesn't the fact that Mary follows Jane in forming new sorts of beliefs after her experience show that Mary's original set of physicalist beliefs must have left something out, namely, information about the subjective side of the experience? But this conclusion does not follow if, in line with my earlier suggestion, the novelty in Mary's beliefs lies at the level of sense rather than reference. And this of course is how the physicalist will diagnose the situation. Before Mary sees red, she has a "third-person" concept of this experience. Afterwards she also has a "first-person" concept. But they are concepts of the same thing. Mary is in the position of somebody who has thoughts about both Cicero and Tully, without realizing they are the same person.

A natural hypothesis about the structure of the new first- person concept acquired in common by Jane and Mary is that it involves a kind of exemplificatory reference by secondary experience. I earlier expressed the new belief formed by Jane and Mary as "that experience is vibrant". I now suggest that we take this construction at face value. Jane and Mary think: THAT experience is vibrant, accompanied by a secondary version of seeing red; they thereby secure reference to the experience of seeing red.[11]

This account of first-person concepts of experience shows, as before, why you can't refer to experiences first-personally until you have had them (you need the primary mould to form the secondary replicas), and it does so consistently with physicalism (since you don't become acquainted with any new facts, but just acquire new concepts).

It is interesting to consider what will happen if and when Mary figures out that her new experience is the kind of experience which is characteristically occasioned by red objects. The natural upshot, assuming that Mary herself is a physicalist,[12] it for her to conclude that she has two concepts with the same referent.[13] And then, as with anybody who realizes this, the t wo concepts will tend to "merge", each becoming merely an aspect of the unified concept with which she refers to the experience of seeing red.

The net result would be the same even if Mary were aware from the start that the new experience she is having is the one characteristically occasioned by red objects. For even in this case we could expect Mary to come to share the first-person mode of thinking about this experience displayed by Jane, albeit that this new mode of thinking would now, supposing still that Mary is a physicalist, be "merged" ab initio with the concept Mary had when she was just a colour vision scientist.

Perhaps there is room to dispute whether such a "merged" concept is really a new concept, compared with the concept Mary had before seeing red. The merged concept will incorporate both the old third-person physical information Mary always had, plus the new first-person mode of thinking she shares with Jane. There are, however, familiar difficulties about whether such amplifications of existing concepts count as genuinely distinct concepts, rather than alterations of old concepts. But rather than getting bogged down in the knotty issue of concept identity, let us simply agree that Mary's concept of the experience of seeing red has been modified, in a way that would not have been possible if Mary had not seen red herself.

So, to sum up the argument of this section, once we have new experiences, we are led to form new sorts of beliefs about those experiences. But this does not show that we thereby come to refer to any distinctively subjectively phenomena. For the distinctive element in these beliefs need be nothing more that the deployment of first -person concepts, and, for all that has been said so far, there is no reason to suppose that such first-person concepts are not co-referential with third-person concepts of experience.

4.5 The Antipathetic Fallacy

I expect that, despite everything I have said so far, many readers will feel strongly that it is a mistake to conclude that "first-person" and "third-person" concepts of experience refer to the same things. For my arguments in the last three sections will have done nothing to shake the widespread intuition that conscious experiences and brain states are as different as anything can be.

Let me summarize the state of play. So far in this chapter I have considered the strength of arguments against the physicalist identification of conscious experiences with brain states. And I take myself to have shown that these arguments are ineffective. There is no valid argument from "what it is like", or from "knowing what it is like", to discredit the physicalist view that having a given conscious experience is nothing more nor less than being a certain kind of physical system.

What is more, I take myself already to have shown, in chapter 1, that the cos t of viewing conscious mental states as something distinct from brain states is the denial of the efficacy of the mental: if you think that consciousness is non-physical, then you are forced to such undesirable conclusions as that your pain is never the cause of the motion of your arm.

I think that together these findings give us good reason to accept the physicalist view that conscious experiences are not distinct from brain states, and therefore to reject any intuitions to the contrary. However, it would be foolish to deny that such intuitions exist. Such non-physicalist intuitions exert a strong pull on all of us, even on us physicalist philosophers who are committed to rejecting them. So in this section I want to offer a diagnosis of these intuitions, with the intention of explaining why they arise even though they are mistaken.

In the previous two sections I have discussed a variety of ways in which we can focus mentally on conscious experiences, a variety of mental acts which refer to types of experience. These acts can be divided into two main categories: those "third-person" acts which are possible prior to your actually having had the experience in question, and those "first-person" acts which are only possible after you have had the experience. In the former category are all the mental acts Mary could perform before she saw red: her "third-person" imaginings and memories of other people experiencing red; her non-introspective identification on behavioural or physiological grounds of certain events as experiences of seeing red; her "third-person" beliefs, conjectures, and other propositional attitudes about the experience of seeing red. In the latter category are the "subjective" analogues of all these mental acts: the "first-person" imaginings and rememberings that involve internal recreation of an original experience; the introspective identifications of new experiences by direct comparison with a "template"; the beliefs, conjectures, and other attitudes that can be formed by people like Jane whose concept of seeing red involves an element of ostension by internal exemplification.

The common feature of the se latter "first-person" acts, and what distinguishes them from the corresponding "third-person" acts, is that they all deploy a secondary version of the experience being referred to. This is the reason, I have suggested, why the first-person acts are only possible after you have had the experience in question yourself. For it is only after you have had the experience that your brain will have the materials necessary to form secondary versions of that experience.

I think that this broad division between first-person ways of thinking about experience, which employ secondary versions which resemble those experiences, and third-person ways, which do not, is the source the strong intuition that conscious experiences involve something more than the physics of the brain. For it is all too easy to conclude, when we reflect on the difference between these two categories of thought, that only the first-person thoughts really refer to experiences, while the third-person thoughts refer to nothing except physical states.

The route to this conclusion begins with the perfectly accurate observation that first-person thoughts include an experiential element which is absent from the third-person cases. First-person thoughts portray the relevant experience directly, so to speak, by giving the thinker a simulacrum, by recreating in the thinker a version of the experience being thought about. Third-person thoughts, on the other hand, do not do this, since they do not involve secondary experiences.[14]

So there is a sense in which third-person thoughts do indeed "leave something out": they do not give us (versions of) the experience being referred to. And this observation can then easily lead to the further conclusion that third-person thoughts are about something different from first-person thoughts: where first-person thoughts refer to the experience itself, in all its conscious immediacy, third-person thoughts merely refer to the external trappings of the conscious event, the physical goings-on which accompany it.

But of course this last step is a fallacy. The fact that we do not have certain experiences when we think third-person thoughts does not mean that we are not referring to them. To make this move is to succumb to a species of the use-mention confusion: we slide from (a) third-person thoughts, unlike first-person thoughts, do not use (secondary versions) of conscious experiences to portray conscious experiences to (b) third-person thoughts, unlike first-person thoughts, do not mention conscious experiences. There is no reason, however, why third-person thought about experiences, like nearly all other thoughts about anything, should not succeed in referring to items they do not use.

I propose to call the above fallacy the "antipathetic fallacy". Ruskin coined the phrase "pathetic fallacy" for the poetic figure of speech which attributes human feelings to nature ("the deep and gloomy wood", "the shady sadness of a vale"). I am currently discussing a converse fallacy, where we refuse to recognize that conscious feelings inhere in certain parts of nature, namely, the brains of conscious beings.

Let me be specific about the target of this charge of fallacy. My target is not the explicit argument against physicalist views of consciousness offered by Jackson. I take the points made in the last two sections already to have shown what is wrong with Jackson's argument. Rather my target is a covert line of thought, whose fallaciousness is obvious once it is spelt out, but which I think has nevertheless seduced a great many thinkers into dualism: namely, the argument which moves from the true premise that third-person ways of thinking about conscious experiences do not use versions of those conscious experiences, to the false conclusion that those ways of thinking do not mention those conscious experience, but only physical states.

Let me also be specific about what I take the identification of this fallacy to explain. It is supposed to explain why many people believe that some mental states are distinctively non-physical. It is not supposed to explain why some physical states are distinctively conscious. This latter kind of question will be addressed in the next section, and I shall there agree that our ability to think about certain states in first-person ways does nothing to account for their possessing the distinctive inner light of consciousness -- though I shall also argue there that the desire to account for such inner lights rests on a confusion. My present concern, however, is not to explain why the states we can think about in first person ways are distinctively conscious, but rather to explain why these states are widely taken to be non-physical.

Both Thomas Nagel, in a well-known footnote in "What is it Like to be a Bat?" (1974, pp 446-7), and Willi am Lycan, in his book Consciousness (1987, pp 76-7), briefly allude to versions of the fallacy I am concerned with. My treatment here enlarges on their remarks in two respects. First, both Nagel and Lycan focus specifically on the contrast bet ween first-person imagination of conscious experiences and the third-person perceptual imagination of the associated brain states: the contrast, for example, between imagining having a pain and imagining the visual appearance of the relevant parts of the sufferer's brain. This is certainly one example of the kind of contrast I am interested in, but this exclusive focus underemphasizes the extent of this contrast. For, as I have observed, the contrast between first-person and third-person modes of thought is not restricted to imagination, but also includes memory, identification, and believing, desiring and other propositional attitudinizing. And even within the category of imagination, perceptual imagination is not the only kind of third -person imagination: if we can form non-perceptual beliefs and other propositional attitudes about brain states, as we surely can, then presumably we can imagine them non-perceptually too. (Nagel does mention "symbolic imagination", but only to exclude it from his analysis.)

Second, neither Nagel nor Lycan emphasize the way that first-person modes of thinking about experiences deploy secondary versions of those experiences. Nagel does, it is true, say that first-person imaginings "resemble" the experiences being imagined. But when he goes on to explain how the fallacy arises, his explanation, like Lycan's, is simply that first-person and third-person imaginings are independent mental acts, each of which can happen without the other, and that therefore we are inclined to conclude that they are about different things.[15] But this diagnosis fails to distinguish the antipathetic fallacy from all the other cases where different modes of thought about the same entity can create the impression that two different entities are being thought about. What is distinctive about the antipathetic fallacy, and what makes it so very seductive, is the fact that one set of ways of thinking about experiences -- the first-person ways -- involve versions of the experience itself, and so create the impression that the other ways of thinking about experiences -- the third-person ways -- leave something out. In general, when two different mode of thought create the impression that two things are being thought about (for example, Cicero and Tully), the illusion is easily enough dispelled on receipt of evidence that there is in fact only one referent. But in the mind-body case the impression of difference continues even in the face of any amount of such evidence, precisely because of the extra feature -- the first-person use of secondary versions -- that makes it seem as if the third-person modes of thought omit mention of the experience altogether.[16]

4.6 Theories of Consciousness

So far I have argued that there are no effective arguments against the physicalist identification of conscious states with physical states, and that the admittedly strong in tuitions which run counter to this view can be explained away. It may still seem to some readers, however, that a further obligation faces defenders of a physicalist view of consciousness: namely, to answer the question raised briefly in the middle of the last section, and explain why some states are conscious and others not.

The obligation I am thinking of here is not just to provide physicalistically acceptable accounts of such specific conscious states as being in pain, seeing red, having an itch in your left finger, or so on. We can suppose for the moment that physicalists can somehow specify which physical occurrences constitute each of these specific mental states. The current challenge is rather to give an explanation of the generic difference between conscious and non-conscious states as such. Why is consciousness present when a person is in pain, or

happy, or itching, but not when a stone is falling, or a tree is growing, or, for that matter, when an anaesthetized human is breathing?

Some philosophers of physicalist inclinations have proposed "theories of consciousness" in answer to this kind of question. I have in mind the kind of theory which aims to identify a physicalistically acceptable characteristic common to all and only conscious states. Some such theories are based on assumptions drawn from everyday thought (for example, Armstrong, 1968, pp 92-99, holds that the states of any self-representing system are conscious); others appeal to the resources of cognitive science (for example, Dennett, 1978, ch 9, suggests that cognitive systems with short-term buffer memories are conscious); and no such theory, I think, commands universal assent.

However, we can leave the details of such theories to one side. For a natural reaction to all such theories is that they simply fail to address the philosophical question at issue. At best such a theory will specify some structural or other physically acceptable characteristic (A, say) which is coextensive with the class of states we are pretheoretically inclined to count as conscious. But then we still seem to face the question: why does consciousness emerge in just those cases? And to this question physicalist "theories of consciousness" seem to provide no answer.

I suspect that many philosophers regard the inability to answer this question as the fatal flaw in the physicalist approach to consciousness. Surely, they feel, any satisfactory philosophical view of consciousness ought to tell us why consciousness emerges in some physical systems but not others.

I think that physicalists should simply reject this question. For the question presupposes that there are two different features at issue, the physically acceptable characteristic A, and being conscious. The physicalist is then challenged to explain the relation between these properties, and in particular to explain why they are always found together. But the physicalist should simply deny that there are two properties here. Being conscious isn't something over and above having A, it just is having A. (In the section after next I shall ask some questions about the sharpness and determinacy to be expected from any A which might provide such a physicalist reduction of consciousness. But it will helpful to shelve such worries for the moment, and assume that some suitable property A is available.)

The idea that being conscious just is having some physical state might seem intuitively implausible: surely the difference between conscious and non-conscious systems is something more than the difference between having and lac king some physical feature. But the defender of a physicalist theory of consciousness, while not denying that these intuitions exist, can account for them as a further manifestation of the antipathetic fallacy. The earlier sections of this chapter were concerned with the thesis that specific conscious states, like seeing red, are identical with specific physical states; and I argued there that our strong contrary intuitions can be explained away as due to the antipathetic fallacy. I would now like to suggest that a generalized version of this fallacy is responsible for the intuition that any physicalist theory of consciousness will necessarily be incomplete.

We can think of the general property of being conscious as standing to experiences like seeing red as determinable to determinate. Seeing red, being jealous, feeling cold, and so on, are the determinate states which have in common the determinable state of being conscious. And so, just as the antipathetic fallacy makes us think that such determinate states as seeing

red are distinct from any specific physical states, so it makes us think that the determinable state of being conscious is similarly distinct from any more general physical state. We are inclined to think of the determinable feature as a kind of generalized non-physical light, which stands to the non-physical features of particular experiences, as, say, the property of being illuminated as such stands to being illuminated with red light. But we shouldn't. Just as it is a mistake to think of experiencing red as something additional to the relevant physical property, so it is a mistake to think of being conscious as an extra inner light, over and above the physical feature A.

Once we fully free ourselves from the seductive "inner light" picture of consciousness, and take seriously the idea that being conscious may literally be identical with some physical A, then we should stop hankering for any further explanation of why physical state A yields consciousness. Consider this parable. Suppose that there are two groups of historians, one of which studies the famous American writer Mark Twain, while the other studies his less well-known contemporary , Samuel Clemens. The two groups have heard of each other, but their paths have tended not to cross. Then one year they both hold symposia at the American Historical Association, and late one night in the bar of the Chicago Sheraton the penny drops. They realize that they have both been studying the same person. At this stage there are plenty of questions they might ask. Why did this person go under two names? Why did it take so long to find out Mark Twain and Samuel Clemens were the same person? But it doesn't make sense for them to ask: why were Mark Twain and Samuel Clemens the same person? If they were, they were, and there's an end on it.[17]

Similarly, the defenders of a physicalist theory of consciousness can say, with consciousness and the physical property A. The defenders of such a theory will take themselves to have discovered that consciousness and A are the same property. So they will allow that we can sensibly ask why there should be different concepts of this property, and why it took us so long to realize that they stand for the same thing; and indeed they can answer these questions, by explaining that there are ways of referring to conscious phenomena that use secondary versions of those phenomena, and ways that don't, and that this in itself makes it easy to succumb to the antipathetic fallacy of supposing that different things are being referred to. But, they will insist, there is no further question of why consciousness is always present when physical property A is. If they really are the same thing, then we can't explain why they are the same thing. Somebody who feels there is still a question here has simply failed full y to grasp the thesis that consciousness is identical with a physical property.

4.7 Life and Consciousness

It may seem to some readers that a physicalist theory of consciousness will come close to denying the existence of consciousness. But that would be a mistake. It doesn't deny consciousness, just a certain conception of consciousness.

It denies that consciousness is some kind of extra inner light, some further non-physical property which exists over and above any physicalistically specifiable property. But this is quite consistent with holding that consciousness is a real property which distinguishes some kinds of systems from others. This combination of views requires only that we accept that consciousness is identical with some property which is specifiable in a physicalistically acceptable way.

An analogy may be helpful here. In the nineteenth century there was a heated theoretical debate about the essence of life. The participants had a satisfactory enough working notion of

life: they agreed about which kinds of behaviour and physical organization are characteristic of life, and in consequence were clear enough about where in practice the line should be drawn. Everything from humans to microbes are alive, while planets and pebbles are dead. (Perhaps there were some borderline cases; but the penumbra of vagueness was not wide.)

Still, despite this wide degree of agreement on the nature of life, nineteenth-century thinkers took there to be a further question. Why are these systems alive? What mysterious power animates them? And why is this power present in certain cases, such as trees and oysters, and not in others, like volcanos and clouds?

These questions have disappeared from active debate. Biology textbooks sometimes begin with a few perfunctory paragraphs about the distinguishing characteristics of their subject matter. But the nature of life is no longer a topic of serious theoretical controversy. Everybody now agrees that the difference between living and non-living systems is simply having a certain kind of physical organization (roughly, we would now say, the kind of physical organization which fosters survival and reproduction).

The explanation for this nineteenth-century debate, and of its subsequent disappearance, was that it was premised on the notion that living systems w ere animated by the presence of a special substance, a vital spirit, or elan vital, which was postulated to account for those features of living systems, such as generation and development, which were thought to be beyond physical explanation. And of course, if you do believe in such a vital spirit, then you will want to know about its nature, and why it arises in certain circumstances and not others.

However, nobody nowadays believes in vital spirits anymore, not least because it i s now generally accepted that the characteristic features of living systems can in principle all be accounted for in physical terms. In consequence, it no longer makes sense to puzzle about why living systems are alive. To be alive is just to be a physical system of a certain general kind. There isn't any extra property present in living systems, over and above their physical features, which distinguishes them from non-living systems. So we have stopped asking questions which presuppose such an extra property.

I recommend that we do the same with consciousness. The apparently nagging question, "Why does consciousness arise in certain physical systems?", is premised, I claim, on the assumption that consciousness is some extra feature, over and above any physical characteristic. But if we accept, as I have argued, that there is no reason to view consciousness in this way, then we ought therewith to stop asking why consciousness is present in the relevant kind of physical system.

Of course the parallel is not complete. In the case of life, the motivation for postulating an elan vital is purely explanatory, a desire to find a cause for phenomena which do not appear to be physically explainable. In the case of consciousness, by contrast, there is also the extra pressure of the antipathetic fallacy. Still, this doesn't affect the point. There may be extra reasons for thinking of consciousness as non-physical, which don 't apply to life. But once we recognize that it is physical, we should do what we did with life, namely, stop asking why it arises in the right physical circumstances.

One last point about the analogy with life. Note that the rejection of an elan vital does not mean that there is no life. There may be nothing special about living systems except a certain kind of physical organization. But this does not mean that the difference between being alive

and not being alive is not real. The postulation of an elan vital was simply one theory about the nature of life. We can reject this theory, and yet still uphold, as we do, the distinction between living and inanimate systems.

A similar point applies to consciousness. We should reject the theory that consciousness involves an extra inner light in addition to facts of physical organization. But we can reject this theory without rejecting consciousness. Even if consciousness is just a kind of abstract physical organization, the difference between being conscious and not being conscious can still be perfectly real.

4.8 Consciousness is Vague

So far I have been assuming that there is some well-defined and precise physical characteristic A which picks out just those states we are pre-theoretically inclined to count as conscious. However, I doubt that this assumption is justified. In this section I shall argue that any physicalist account of consciousness is likely to make consciousness a vague property. In the next section I shall argue that questions of consciousness may not only be vague, but quite indeterminate, in application to beings unlike ourselves. I do not intend these points as criticisms of physicalism. Rather my aim is to show that if we take physicalism seriously, some assumptions that we take for granted about consciousness may have to go.

The point about vagueness is suggested by the analogy with life. If life is simply a matter of a certain kind of physical complexity -- the kind of complexity that fosters survival and reproduction, as I put it above -- then it would seem to follow that there is no sharp line between life and non-life. For there is nothing in the idea of such physical complexity to give us a definite cut-off point beyond which you have enough complexity to qualify as alive. Rather as with baldness, or being a pile of sand, we should expect there to be some clear cases of life, and some clear cases of non-life, but a grey area in between where there is no fact of the matter. And of course this is just what we do find. While there is no doubt that trees are alive and stones are not, there are borderline cases in between, like viruses, or certain kinds of simpler self-replicating molecules, where our physicalist account of life simply leaves it indeterminate whether these are living beings or not.

But now, if consciousness is like life, we should expect a similar point to apply to consciousness. For any physicalist account of consciousness is likely to make consciousness depend similarly on the possession of some kind of structural complexity -- the kind of complexity which qualifies you as having self-representing states, say, or short-term memories. Yet any kind of such complexity is likely to come in degrees, with no clear cut-off point beyond which you definitely qualify as conscious, and before which you don't. So we should expect there to be borderline cases -- such as the states of certain kinds of insects, say, or fishes, or cybernetic devices -- where our physicalist account simply leaves it indeterminate whether these are conscious states or not.

Some philosophers regard this as a reductio ad absurdum of the physicalist view of consciousness. They take it to be intuitively obvious that there is a sharp line between conscious and non-conscious states.[18] So they conclude that there must be something more to consciousness than a certain kind of physical complexity.

I go the other way. I think that the physicalist approach to consciousness is correct. So I reject the intuition that there is a sharp line between conscious and non-conscious states.[19]

I accept, of course, that such intuitions exist. But I regard them as a further consequence of the "inner light" picture of consciousness, the picture into which it is so easy to be seduced by the antipathetic fallacy. For if you do think of consciousness as such an extra inner light, then you will no doubt think it is a sharp matter which states are conscious -- states which possess the inner light are conscious, and those which don't are not.[20] On the other hand, if the idea of such an extra inner light is a confusion, as I take it to be, then we have no obligation to respect any further intuitions which stem from it.

If the line between conscious and non-conscious states is not sharp, shouldn't we expect to find borderline cases in our own experience? Yet when we look into ourselves we seem to find a clear line. Pains, tickles, visual experiences and so on are conscious, while the processes which allow us to attach names to faces, or to resolve random dot stereograms, are not. True, there are "half-conscious" experiences, such as the first moments of waking, or driving a familiar route without thinking about it. But, on reflection, even these special experiences seem to qualify unequivocally as conscious, in the sense that they are like something, rather than nothing.

However, I don't think that this dis credits my claim that the boundaries of consciousness in general are vague. For I think there is a special reason why we are able to draw a sharp line in our own case. Namely, that in our own case we can simply note which states are introspectable, re-creatable in imagination and memory, and otherwise accessible in first-person ways. States which are so accessible we count as conscious, and those which are not we consider non-conscious.

What exactly is the rationale and s tatus of this decision procedure? This is a tricky question, to which I shall return in the next section. But whatever view we take on this question, note that the decision procedure in question will not work for all beings. For once we move beyond the case of humans, to those many animals and other possible organisms who lack the ability to think about their own cognitive states, then the decision procedure in question ceases to apply. So it will be of no help in deciding whether the states of sharks, for example, or octopuses, are conscious.

So I think we should accept that sometimes it will be a vague matter which states of which beings are conscious. It would be a mistake to conclude from this, however, that consciousness is unimportant or unreal. Any number of genuine and important properties are vague. Consider the difference between being elastic or inelastic, or between being young or old, or, for that matter, between being alive and not being alive. All these distinctions will admit indeterminate borderline cases. But all of them involve perfectly serious properties, properties which enter into significant generalizations, are explanatorily important, and so on.

4.9 Consciousness is Anthropocentric

In this section I want to raise some more serious doubts about consciousness, doubts which suggest that consciousness is not only vague, but downright indeterminate.

The last section was premised on the assumption that consciousness involves some kind of physical or structural complexity; the corollary was simply that consciousness, like other kinds of complexity, will therefore admit borderline cases. But what if there isn't any specific kind of complexity common to conscious states, vague or otherwise?

It will be helpful to approach this possibility by returning to the suggestion, made in the last section, that in practice we decide which huma n states are conscious by considering whether they can be thought about in first-person ways. Now, there are two different ways of looking at this decision procedure. One would be to regard it as a test for the presence of some property that can be independently specified, such as appearing in the short-term buffer memory, say. On this way of conceiving the matter, consciousness is a property that can be independently specified, and first-person accessibility is an empirical symptom of the presence of this independently specifiable property. But there is a rather more plausible way of understanding the decision procedure, which analytically ties the test of first-person accessibility to our notion of consciousness. That is, suppose that our notion of consciousness starts with the test of first-person accessibility, and that the reference of this notion is simply fixed as that feature which is common to all those states which can be thought about in first-person ways. Fr om this point of view, first-person access isn't an empirical symptom of some independently specifiable property, but the hook by which we pick out that property in the first place.

This alternative, however, leaves open the possibility that there isn't any such property in the first place, vague or otherwise, After all, the class of states which we humans can think about in first-person ways is extremely heterogeneous. As well as pains, itches, tickles, and the various modes of sense experience, there are emotions, cogitations, and moods. There seems no obvious reason, on the face of it, why there should be any structural or other physicalist property common to this whole genus. Each species within the genus may share some common physical or structural characteristic which renders it explanatorily significant. But why suppose that there is some further such characteristic, common to members of all these species, which binds them all together?

What about the property of being first-person accessible itself? This is a kind of structural property, and therefore physicalistically acceptable; and it is unquestionably common to all those states which can be thought about in first-person ways. But this property is ill-suited to provide an analysis of consciousness. For, even if first-person accessibility provides a reference-fixing description, our notion of consciousness seems clearly to be a notion of some other property which is responsible for first-person accessibility, not just the concept of first-person accessibility per se.[21]

This is why most people think it obvious that higher mammals, like cats, and bats, and human infants, have conscious states, even though these animals are not capable of thinking of their own states in first-person (or any other) ways. These animals may not have first-person access to their own cognitive states. But their sensory and other states seem so closely similar to our own in every other respect that it seems natural to conclude that they must share the property that underlies the first-person accessibility of our own conscious states, whatever that property might be.[22]

However, to repeat the question, what if there is no such property? What if there isn't anything physically or structurally in common to all our first-person accessible states? We may still feel it is uncontroversial that other higher mammals are conscious, because of the close overall similarity between their states and our own. But once we start considering beings that are less closely allied to us, like fish or toads, not to mention Proxima Centaurians and other extra-terrestrials, t hen we are left with nothing to go on, and it becomes quite indeterminate how the notion of consciousness should apply to their states.[23] The problem here isn't just be the kind of vagueness discussed in the last section. At that stage I was

assuming we knew what kind of organizational complexity was at issue. The only problem was how much of it fish and Proxima Centaurians needed to qualify as conscious. But now we are facing the possibility that there is simply no fact of the matter about what kind of physical or structural features you need to qualify as conscious, let alone how much.

Even this needn't make us reject talk of consciousness altogether. Maybe consciousness isn't an explanatorily important property, the kind of property that enters into laws and serious explanations. But the concept can still be useful in characterizing humans and closely related beings. We might draw an analogy with concepts like good-looking, or witty. These are perfectly useful concepts, and indeed ones which play an important role in human affairs. But nobody would think that they cut nature at the seams, or that it made any significant sense to apply them to beings like fish or toads or Proxima Centaurians.

This view of consciousness may seem to have awkward moral consequences. For questions about consciousness often have moral significance. Whether fish are conscious, for example, seems crucial to the issue of how we should treat them. But if there is no fact of the matter as to whether they are conscious, then doesn't it follow that that there is no right and wrong about how to treat them?

I agree that the position I have reached does have unexpected moral consequences. But I don't think that this shows there is anything wrong with the position. Rather, the position helps us to think better about certain moral questions. I take it that the consciousness of fish and similar beings can only be morally important if there is a definite fact of the matter. If there isn't a definite fact of the matter, we will do better to base our decisions about fish directly on information about the organization of their brains and nervous systems, and not on the supposed further issue of whether this physical organization makes them conscious. Indeed, the idea that this is a further issue of moral importance here seems to me not only theoretically misguided but morally dangerous.
Perhaps we might be persuaded by the physical facts that it is wrong to injure certain beings, even though we felt unsure, prior to addressing this moral question, whether they should be deemed conscious. In such a case, should we count them as conscious because we regard them as objects of moral concern? I am sure that we would do so in practice. It may seem odd to hold that certain beings might be conscious because they are morally significant. But the thought isn't that how it is for them depends on the moral conclusion -- merely that the moral conclusion would give us a motive for refining the indeterminate notion of consciousness in such a way as to include them in the category of conscious beings.

4.10 Pains, Shapes and Colours
In this final section of this chapter I want to return to such specific mental states as pains, tickles, visual experiences and emotions, and consider whether these states are determinate, even if consciousness is not. For nothing in the last section rules out our identifying these specific mental states with specific physical or structural properties, thus making it definite which beings have them, even if there is no way of doing this for the overall genus of consciousness. In the terms used earlier, perhaps there are physical equivalents for the determinates like pains, sensory experiences, emotions, and so on, even if there is none for the determinable property of consciousness itself.[24]

Apart from its intrinsic interest, this possibility would make a difference to the moral issues touched on at the end of the last section. It wouldn't matter too much if there is no principled

basis for deciding whether fish are conscious, if there is a fact of the matter on whether they feel pain.

However, when we investigate this issue, we shall see that there are problems about projecting even such specific conscious states as pain or colour experience onto beings other than humans or higher mammals. For once we abandon the seductive picture which identifies these states with different kinds of inner light, as I have argued we must, then we must face up to the possibility that there is nothing else to decide whether some alien being has the same experience as you have when you see something red.

In a sense such specific states as pains and colour experiences raise a converse problem to that rais ed by the generic property of consciousness. In the case of the generic property, we started with those states which the test of first-person accessibility identifies as conscious, and asked what physicalistically acceptable property might tie them together. The problem was that there may not be any such property, since the different species of human consciousness are so various. On the other hand, if we start with the states we identify as pains, or experiences of red, and so on, the difficulty isn't so much that they may share no physical features, but that they seem to share too many.

Let me explain. It seems likely that human beings who share pains, or colour experiences, or other sensory states, will do so because they have determinate physical properties in common. So far, so good for physicalism. But the trouble is that it also seems likely that such physical commonalities will appear at a number of different levels of abstraction. For example, it may be that two human beings who are both in pain will both have certain kinds of nerve cells firing. But, if so, then they will also share further properties, such as the functional property of having-some-property-which-plays-a-certain-causa l-role. The problem for physicalism is to decide which of these competing properties pain is identical with.

Lycan (1987) has emphasized that there are likely to be a large number of different levels at issue here, starting with very strictly physical levels, which are describable only in the language of fundamental physical science, through physiological levels, and on to various functional levels, which will themselves be distinguished by the fine-grainedness of the causal role they involve. I think Lycan is quite right about this. But for my present purposes nothing will be lost if we revert to the familiar philosophical oversimplification, and pretend that there are only two competing levels at issue, which we can take to be the physiological level ("C-fibres firing", to adopt the conventional philosophical shorthand for the physiology of pain) and the folk-psychological functional level ("a state which mediates between bodily damage and the desire to avoid the cause thereof").

As I observed in chapter 1, there is an obvious rationale for identifying mental states with functional states rather than physiological ones. Namely, that the choice of physiological states would have the "chauvinist" implication that beings with different physiologies, like toads, perhaps, or silicon-based Proxima Centaurians, certainly, could not share our mental states. Yet it seems unreasonable to conclude that Proxima Centaurians cannot believe that the universe is expanding, or, for that matter, that they cannot feel pains, just because they are made of silicon and not carbon.

Yet in the case of conscious mental states, states that it is like something to have, there are also strong contrary intuitions in favour of the equation with physiological states. These intuitions are best elicited by spectrum-inverting thought-experiments. Imagine that you have

your retina altered at birth so that you respond physiologically to green objects in the way other people respond to red objects. After the operation you are then raised normally, so that you learn to call red objects "red", post letters in post boxes, eat red and not green tomatoes, and so on. In consequence, the stat e produced in you by red objects plays the same causal role as normal people's experiences of red. But the physiology of this state will be like the physiology of normal people's experiences of green. What will it be like when you see a red object? A widespread intuition is that it will be like most people's experience of green. According to this intuition, the subjective nature of your colour experience is fixed by what physiological processes are taking place in your brain, and not by what causal role those processes play.[25]

So there seem to be two conflicting intuitions: the anti-chauvinist intuition that wants the Proxima Centaurian to share our mental states, and so equates those states with functional states; and the spectrum-inverting intuition that wants people with abnormal retinas to see red where we see green, and so favours the equation with physiological states.

David Lewis (1980) has developed a theory which aims to accommodate both these conflicting intuitions. In Lewis's view, experiences go with physiology for similar beings, but with functional role for different kinds of beings. Lewis considers pain rather than colour experience. He imagines a human (a "madman") who is spectrum-inverted with respect to pain. The madman is arranged so that the physiological state which realizes pain in normal humans is produced in him, not by bodily damage, but by moderate exercise on an empty stomach; and it doesn't cause him to writhe or try to alter the state, but rather to snap his fingers and think of mathematics. Lewis takes it that the madman will share the experience of pain with normal humans. So pain goes with physiology for humans.

But Lewis does not therefore think that an extraterrestrial being (a "Martian") cannot feel pain. He takes it that a Martian will feel pain just in case it is in the physiological state that realizes the functional role of pain in normal Martians. So a normal human and a normal Martian who both feel pain will share the functional state of being-in-some-state-with-the-relevant-causal-role. Pain goes with functional role for normal beings from different species. (Within the Martian species it goes with physiology again: there could be a mad Martian who feels a pain, not because any of its states play the functional role of pain, but because it is in the physiological state which plays that role in norm al Martians.)

The attractions of Lewis's theory are obvious. It accommodates the intuition that the experiences of spectrum-inverted people depend on their physiology, but avoids the chauvinist consequence that beings of other species cannot share our experiences.

It does, however, have an odd consequence. Imagine that Martians and humans are similar enough to interbreed, in virtue of the fortunate fact that their genes are effectively identical; the only substantial exceptions are the genes that direct the development of the pain mechanism, where, as it happens, the Martian genes are dominant.

So a Martian-human hybrid would have its pain mechanism realized by Martian rather than human physiology. Now imagine that such a hybrid exists, and that its pain mechanism is activated. Is the hybrid in pain? If we count it as a Martian, then it will be: for it will be in the physiological state that realizes the role of pain in normal Martians. But if we count it as a human, it won't be: for, although it is in a physiological state that plays the functional role of pain, this isn't the state that plays that role in normal humans.

Lewis is not unaware that his theory has this kind of consequence. Although he does not consider such an extreme case, he does observe that attributions of experiences will depend, given his theory, on which populations we assign individuals to; and he admits that such questions of classification will not always admit of hard-and-fast answers.

Still, even if Lewis is aware of it, this consequence is still pretty odd. Surely, one feels, whether a given being is in pain is a determinate matter, quite independent of what population we might choose to classify it under. (The hybrid's state isn't going to stop hurting, just because the Earth Government changes its immigration regulations to allow that a single human parent qualifies you as human.)

Odd as this consequence is, I don't think that it should lead us immediately to dismiss Lewis's theory. It is possible that our conviction there is a fact of the matter about alien pains stems from the antipathetic fallacy and the associated picture of extra inner lights. For, if pain were an extra inner light, separate from the physics of the brain, then it would in principle be determinate which brains were illuminated by it. But if there is nothing there, apart from the physics of the brain, then it may be indeed be arbitrary how to classify beings whose brains are like ours in some respects, but not in others.

I have illustrated this possibility with respect to pain, as this is the case that Lewis focuses on. But in fact pain is a somewhat unconvincing example of the possibility. While I do think that there are some sensations whose possession is an indeterminate matter, I don't think that pain is one of them.

This is because I do not think that the intuitions in favour of identifying pains with physiological states carry much conviction to start with. Let us go back to Lewis's madman. According to Lewis, the madman's pain is caused, not by injury, but by moderate exercise on an empty stomach. And it doesn't make him writhe or want to alter his state, but simply to snap his fingers and think of mathematics. Given all this, it doesn't seem to me to make much sense to say the madman is in pain. The madman may share the physiology of normal humans in pain. But if this physiological state causes the madman no discomfort, if he lacks all inclination to make it go away, then I'm inclined to say that it doesn't hurt, that it's just not a pain.

The concept of a conscious pain, it seems to me, is the concept of being-in-a-state-which-disposes-you-to-certain-sorts-of-behaviour. Something just isn't a pain unless your initial reaction is to get rid of it. If this is right, then pains must be equated with functional states, rather than physiological ones. So "madmen" and "mad Martians" are not in pain, even though they share the physiology of their normal conspecifics. And this now removes the earlier indeterminacy: the human-Martian hybrid is unequivocally in pain, however we classify it, for its state plays the functional role of pain.

This disambiguation may of course still leave us with a penumbra of vagueness. Even if pain is firmly tied to functional role rather than physiology, there may remain an indeterminacy about how complex this functional role has to be before it qualifies as pain. But vagueness is a different issue, as we saw earlier. Our current concern is what kind of physical or structural complexity pain should be identified with. We can have a definite answer to this question even if we are vague about how much of that complexity is needed..

Which other sensations are like pain in being conceptually tied to behaviour? Sensations like these will be unequivocally identifiable with functional rather than physiological states, and in consequence their ascription to beings other than ourselves will be determinate, up to the boundaries of vagueness.

There is good reason to regard visual experience of shapes as like pain in this respect. A number of recent works have focussed on such experiences, and their arguments strongly support the vie w that visual experience of shape goes with functional dispositions to behaviour rather than with the physiology of the normal viewer. A test case would be a person who is in the physiological state that normally goes with seeing something square, but tries to draw the shape in question by making circular arm movements. Intuition strongly favours the view that this person must have the conscious experience as of seeing something circular, and thus supports the identification of the experience with functional role rather than physiology.[26]

Indeed, in the case of spatial perception, there seems to be direct empirical evidence in favour of a functional over a physiological identification. I am thinking here of the well-known psychological experiments in which subjects wear glasses with "inverting lenses".[27] When they first wear the lenses, subjects faced with an upright drinking cup, say, will have both the physiology, and the dispositions to behaviour, that normally go with an upside-down cup. Accordingly, we can all agree that at this stage the subjects see the cup as upside down. But after a while such subjects learn to adjust their behaviour, so that they come to behave in the way appropriate to upright cups, even though they still have the physiology that normally results from upside-down cups. And at that stage they then say that the cup "looks the right-way up" again. This obvious fits with the thesis that conscious spatial perception is tied to behaviour rather than physiology.

In fact this experiment is less straightforward than it seems. For the inverting-lens experiment doesn't so much test the thesis that spatial perception is tied to behaviour (after all, I am treating this as a conceptual truth), as the conjunction of this thesis with the further assumption that subjects can tell what kind of experience they are having, even after they have been turned into "spatial madmen". To confirm this, note that somebody who holds that spatial perception goes with normal physiology, rather than with normal behavioural function, can accommodate the inverting-lens experiment simply by arguing that retrained subjects can no longer be relied on to report accurately which how things look to them.[28]

Now consider colour experiences. In this case it seems unlikely that there is any conceptual tie between seeing something as red, say, and behaving in any particular way. When I introduced the colour-spectrum-inverting thought experiment earlier in this section, I said that after the operation you would "call red objects 'red', post letters in postboxes, eat red and not green tomatoes, and so on". Most of the behaviour involved in this functional characterization (saying "red", using red postboxes) depends on nothing more than social convention, and so can scarcely be part of what it is to see red. (We don't want to say that you can't see red unless you know the English word "red".) And the non-conventional behaviour associated with seeing red (eating tomatoes and similar fruit) still seems too thin and topic-specific to tie down the experience.[29]

So colour experience s are different from pains and spatial perception. I do not want to deny that such experiences as seeing red have a characteristic functional role. After all, common sense criteria, which define the functional role for red, are clearly sufficient in practice to

decide which human beings are experiencing red. (In this connection we should not forget the central fact that red objects normally cause red sensations.) But, by contrast with pains and spatial perceptions, colour experiences do not have a stock of non-conventional desires or actions to call their own. And, because of this, it seems unconvincing to argue, as we did for pains and spatial perceptions, that colour experiences are determinately tied to functional roles, rather than to physiology. Where there is a direct link between a conscious experience and something we non-conventionally do, then it seems natural to hold that this functional link fixes the nature of the experience. But with experiences which lack any such intrinsic tie to action, there seems to be no corresponding rationale for holding that functional role, rather than physiology, determines the experience.

So I conclude that with colour experiences (and similarly for tastes and smells[30]) there is a real indeterminacy about how to project our categories beyond the case of normal humans. As long as the physiology and the functional role continue to go together, then there is no problem. But when we have one without the other, as with the subject of the spectrum-inverting operation (the "colour madman"), -or a Martian who comes to earth and learns to make our colour discriminations, then I don't think there is any fact of the matter about whether they have the same experiences as us.

There is still David Lewis's strategy, which decides such cases by seeing whether the difficult individuals share the physiology of the functionally normal members of their group. But then, as we saw, it may be indeterminate which group we should consider the difficult individuals to be part of. Lewis's strategy does place some extra constraints on our ascriptions of colour experiences to difficult cases. But, by making such ascriptions depend on assignments to groups, Lewis in the end only hides the underlying arbitrariness of experiential classifications under the cloak of a different arbitrariness.

I realise that some readers will think it ridiculous for me to suggest that it is an arbitrary matter whether or not colour madmen are counted as have the same experiences as the rest of us. (Surely either they do or they don't). But let me recall a point I made at the end of the last section. I am not suggesting that how it is for the colour madman will depend on how we classify his experience. Of course it won't. My claim is only that it is indeterminate whether the madman's experience is the same kind of experience as our experience of red. That is, I don't think that there's anything lacking in the colour madman. It's just that the notion of sameness of colour experience breaks down when we come to such cases.

No doubt some readers will find even this absurd. Even if I am not saying that we can alter feelings by linguistic fiat, isn't it bad enough for me to be saying that experiential comparisons are indeterminate? Take one of the madman's colour experiences. Now imagine what it's like to see a bright red tomato. Surely the madman's experience is either like that, or it's not. What could be simpler?

But I don't think it is that simple. The reason it seems simple is that we naturally suppose that, when we have (or imagine) a visual experience, we switch on an inner light. And so we all we need to do is compare that shade of inner light with the shade illuminating the madman's mind. But there isn't any such inner light. There are just the physical and structural features of the relevant brains, some of which we share with the madman, and some of which we don't. So our conviction that either the madman must feel the same or feel different is based on a false picture. Wittgenstein had a good analogy: "You surely know what 'It's 5

o'clock here' means; so you also know what 'It's 5 o'clock on the sun' means. It means simply that it is just the same time there as it is here when it is 5 o'clock." (1953, §350.)

1. This follows standard practice in this area: see Horgan (1984, pp 147-8) and Tye (1986, p 1).

2. Moreover, most of the arguments between dualism and physicalism arise in exactly the same way between dualism and any more general non-physicalist "objectivism" about conscious mental states. I shall formulate the issues as a matter of physicalism versus dualism, however, since I think there are good arguments -- namely, those presented in chapter 1 -- for preferring physicalism to other kinds of objectivism. Even so, much of what follows should also be of interest to objectivists who are not physicalists.

3. An immediate qualification is needed here. For we can obviously imagine complex experiences, like seeing a unicorn, as long as we've previously experienced the elements separately. And we can perhaps imaginatively extrapolate to intermediate experiences, like imagining a colour which is spectrally between others we have previously experienced. But these possibilities are clearly consistent with the general thesis that the brain needs to acquire the materials for the replicas from previous experiences, and so in accord with the fact that we can't imagine experiences of a radically unfamiliar kind, like seeing colours at all, or echolocating, until we have actually had those experiences.

4. The view that Mary acquires new abilities rather than new knowledge is urged by Lewis (1988) and Lemirow (1990).

5. Note that the alternative non-physicalist account, in terms of phenomenal property P, does nothing at all to explain why the exercise of our recreative abilities should in some sense make us re-experience the original mental state. Thinking of or remembering something as an event with some property P can in general have any experiential nature, or none at all. Of course, it could be argued that, in the particular case of some phenomenal property P, thinking of or remembering an event with that property involves recreating in your brain a copy of the experience characterized by the property. But, once this last move is made, then it becomes unnecessary to bring in the phenomenal property P to explain Mary's new imaginings and memories in the first place -- for now we can simply exp lain these imaginings and memories directly, by appealing her recreative abilities.

6. For this terminology, see Mellor (1992, p 11).

7. Lewis (1983, pp 131-2)

8. Again a qualification is needed to accommodate the fact we can recognize novel complex experiences, as long as their components have previously been experienced: in such cases the brain doesn't need an original complex experience to form a complex template, but only the originals of the component experiences to form templates of the components.

9. Cf Peacocke (1989, pp 67-9).

10. Isn't Jane ruled out by Wittgenstein's private language argument? Well, she'd better not be, if Wittgenstein's argument is any good, since Jane is clearly possible. I don't think there is in fact any tension here. I take the moral of Wittgenstein's argument to be that there must be

room for error in people's judgements about their experiences, not that those judgements must necessarily be expressed in a language used by a community. And I see no reason to suppose that Jane cannot make mistakes about her own experiences.

11. This suggestion is central to the response made to Jackson's argument in Horgan (1984).

12. Of course, if Mary isn't a physicalist, then she will be disinclined to make this identification, and will no doubt maintain that the first-person concept she shares with Jane refers to a phenomenal attribute, whereas her scientific concept refers to a physical phenomenon.

13. Won't this realization involve some new information, of a kind Mary couldn't have had before her experience? After all, someone who discovers that a = b, where [a] and [b] express two modes of presentation of the same object, will generally acquire the information that the property invoked by [a] is co-extensional with the property invoked by [b]. So won't Mary acquire the new information that the property of having such-and-such neurones firing is co-extensional with the phenomenal with the phenomenal property of red? However, this argument assumes that Mary's first-person mode of presentation of the experience of red invokes some phenomenal property. In contrast, I have just suggested that this is an indexical construction. If this is right, Mary no more acquires new non-physical information than someone who suddenly realizes that it is noon now.

14. Or, if they do involve secondary experiences, as when we think about somebody being in pain, say, by thinking about the visual aspect of their behaviour or b rain state, then they will be different secondary experiences, secondary version of visual experiences, not secondary versions of pain experiences.

15. In line with this, Lycan calls the fallacy the "stereoscopic fallacy".

16. In a generous gesture of help to his physicalist opponents, Nagel points out that the fallacy in question provides an answer to Kripke's modal argument against mind-brain identity. Kripke (1972) appeals to the principle that identity statements involving rigid designators are necessarily true, and then challenges physicalists, who identify mind and brain, to account for the apparent contingency of mind-brain identity statements. Nagel's suggestion is that, instead of looking for some non-rigid way of reading the terms in these statements, which is how we account for other apparently contingent identity statements, like "water = H20", physicalists should simply explain the appearance of mind-brain contingency by reference to the fallacy that makes us so convinced that mind and brain are different to start with. On this suggestion, we won't explain away the appearance of contingency by finding some non-rigid reading which is violated in other possible worlds, as we do with the other cases. Rather, we simply account for the appearance of contingency by explaining why we are so disinclined to accept mind-brain identities in the first place. I agree with Nagel that this is the right way for physicalists to respond to Kripke's argument.

17. Ned Block offered this story to me; I don't know where it originated.

18. Cf McGinn (1982, pp 13-14).

19. An alternative physicalist response to the intuition that consciousness is not vague would be to seek some physicalist characteristic A which does provide a sharp dividing line. But this strategy strikes me as unlikely to succeed.

20. Thus McGinn, ibid: "The emergence of consciousness must rather [unlike the emergence of life] be compared to the sudden switching on of a light . . ."

21. No doubt the idea of an inner light as such a property is partly responsible for our having this notion of consciousness. But it would be a pity, I think, to build the inner light itself into our notion of consciousness.

22. This is of course the standard objection to self-monitoring theories of consciousness like Armstrong's.

23. Chris Hughes suggested to me that the relevant question is whether the states of toads and similar beings would be first-person accessible, if they occurred in beings who could introspect, imagine, and so on. But I doubt this really removes the indeterminacy. Exactly which counterfactual possibilities are we supposed to consider? Is the question whether we humans could introspect toad states, if they occurred in us? Or are we supposed to consider super-toads, who stand to toads as we do to monkeys? But then what is supposed to stop us considering super-trees, say, or super-stones?

24. If so, couldn't we just disjoin the determinates to get a physical equivalent for the determinable? But we still lack a principle to generate all instances of the genus. We might be able to cover all human determinates by brute enumeration. But, in the absence of any property equivalent to consciousness as such, there will be nothing to decide which states of the Proxima Centaurians should be included in the disjunction.

25. Note how this thought experiment differs from the traditional version, in which physical identicals have inverted spectra. I take this traditional version to be discredited by the general arguments for physicalism. The modern version, by contrast, involves people with different physiologies. So it doesn't present a problem for physicalism as such, but only for functionalism.

26. See in particular McGinn (1989, pp 58-94), Davies (1992). It should be said that this literature is more concerned with whether spatial experiences have broad or narrow content s than with their phenomenal identity as conscious states. But the two issues are connected. See Davies, op. cit., sect III. Davies's concern with the issue of content leads him to distinguish carefully between internal inclination to behaviour, and actual external behaviour (with phenomenology going with the former, and content with the latter). But from our perspective these are alike matters of functional role.

27. See Gregory (1977, ch 12).

28. Some readers might feel it would make more sense for sensory states to be incorrigibly tied to introspective reports, instead of to further links to behaviour. But it seems wrong to rule out introspective mistakes in this way. Apart from anything else, there are brain abnormalities that seem to affect introspective abilities rather than anything else. Morphine is a good example: it makes people say that the pain is still there (even though they don't mind it); but, if Daniel Dennett (1978, p 208-9) is right, these people are not in pain on anybody's

account, since not only do they fail to display pain behaviour, but they also lack the normal physiology of pain.

29. Janet Levin (19xx) takes the contrary view that even experiences of colour, taste and smell might be definable by their links with non-conventional behaviour. In line with this, she suggests that spectrum-inverting operations with these modalities might turn out like the "inverting lens experiment": at first the madmen's responses will involve both the "wrong" physiology and the "wrong" behaviour ; but after a while they will adjust their behaviour to make it "right"; and then things will seem consciously "right" to them once more. This argument raises a number of issues. Central is whether the inversions would lead to systematically inappropriate behaviour of a kind that could be remedied by a systematic (rather than piecemeal) rewiring of our behavioural responses. I agree with Levin that if this were so, then it would be appropriate to describe the rewired people as having the "right" phenomenology again. What I doubt is whether there are such systematic links between the relevant experiences and behaviour to start with.

30. Sounds raise yet further issues, which I shall not pursue. Note that the categorization of experiences in terms of their links with behaviour does not coincide with the division between primary and secondary qualities. I have argued that a constitutive link to behaviour is present both in experience of shape, which is a primary quality, and painfulness, which can be thought of as a (hyper-)secondary quality.

CHAPTER 5 RELIABILISM, INDUCTION, AND SCEPTICISM

5.1 Introduction

At the end of chapter 3 I pointed out that the teleological theory of representation has radically anti-verificationist consequences. The contents of belief are fixed by their consequences for action, not by the circumstances that lead believers to adopt them. So it is perfectly possible that a judgement should have a given truth condition, and yet human beings be systematically prone to form this belief when it is false.

Such a realist[1] account of representation might be thought to open the door to scepticism: if truth-conditions transcend evidence, then what assurance do we have that our beliefs are free of error? In this chapter I want to show how this sceptical threat to knowledge can be met from within a realist perspective. Accordingly, in what follows I shall take the implications of chapter 3 as read, and assume without further argument that judgements about the natural world answer to non-verificationist truth conditions. My focus here will rather be on the notion of knowledge, and on how a proper understanding of this notion enables us to give an adequate response to scepticism, even within the framework of a realist theory of representation. At the beginning of the next chapter, however, I shall return to the general debate between realism and anti-realism, and compare my overall realist attitude to representation and knowledge with the anti-realist alternative

In more detail, the plan of this chapter is as follows. Sections 5.2-8 will defend a reliabilist account of knowledge. Such reliabilist theories are nowadays fairly widely accepted; but a distinctive feature of my defence will be its appeal to the point of the concept of knowledge, rather than to intuitions about test cases. In sections 5.9-13 I shall then show how this reliabilist account of knowledge provides an answer to the traditional sceptical problem of induction. Sections 5.14-17 will then generalize this answer and address some other arguments for scepticism.

5.2 Knowledge and the Project of Enquiry

Let me start with a question raised in Chapter 2 of Bernard Williams' Descartes (1978). Williams asks: why do human beings want knowledge? He takes it as given, as I shall, that humans want true beliefs. But, as we all know, a belief can be true and yet not be knowledge, as when it is a mere hunch or some other lucky chance. So the point of the question is: why do we want our beliefs to be known, in addition to being true?

Williams' answer goes as follows. Human beings are prone to false beliefs. So, if our desire for true beliefs is not to be idle, we will need to exercise ourselves to achieve it. It is no good, however, to start checking through all your beliefs with the intention of discarding the false ones. To have a belief is to take that belief to be true. So once you have formed your beliefs, internal inspection will not serve to distinguish the true from the false ones. Insofar as you are prone to error, the damage will already have been done.

The only effective way for us to ensure that our beliefs are true is to block the error at source, by bringing it about that the processes by which we acquire beliefs in the first place are ones that generally yield true beliefs. So Williams argues that the desire for true beliefs itself generates the desire that our beliefs should issue from processes that generally produce truths. And then, finding it independently plausible that beliefs produced by such processes should count as knowledge, Williams has an answer to his original question as to why we should

want knowledge: our desire for knowledge derives from our desire to avoid error, in that attaining knowledge is the only effective means by which humans can avoid error.

I want to draw something more ambitious from this analysis. I think that, in addition to explaining why we should want knowledge, Williams' story also shows us what knowledge is. Williams takes it as given, from outside his analysis, that beliefs generated by truth-producing processes will count as knowledge. But I think that his story also explains why we have this concept of knowledge, why we pick out beliefs generated by a truth-producing process as knowledge, as an especially good kind of belief. My idea here is that our concern to avoid error makes us especially interested in the state we need to get into as a means to avoiding error, and that this is why we call that state "knowledge" - the state, to repeat, of having acquired a true belief from a process which generally produces true beliefs.[2]

5.3 Certainty and Reliability

The above remarks prompt an immediate question: how truth-productive does a belief-forming process need to be in order to be an effective means of avoiding error, and therefore to qualify as a source of knowledge? In particular, is it enough that it merely be reliable, in the sense that it generally delivers true beliefs as a mat ter of contingent fact in this world? Or does it need in addition to yield certainty, in the sense that it should be impossible for a belief issuing from that process to be false?

Much traditional philosophical thinking assumes that knowledge requires certainty. But from the point of view of my remarks in the last section it is not clear why certainty should be necessary. Knowledge, I have suggested, is the state that we need to get into if we are to succeed in avoiding error. But we will succeed in this aim as long as we have belief-forming processes which are reliable in this world. That such processes would lead us astray if things were different does not mean that they will lead us astray, as things are. This line of thought suggests that the traditional demand for certainty may be a mistake, perhaps fostered by an over-optimistic view of what human thought can achieve, but inessential to knowledge itself.

I shall return to the idea that the demand for certainty may be a mistake in section 5.5 below. But first, in the rest of this section and the next, let me say a bit more about the contrast I have drawn between reliability and certainty. Note that I have defined certainty objectively, rather than psychologically: the issue is whether it is in fact impossible for a given belief-forming process to produce a false belief, not whether the subject is aware of this, nor whether it yields some feeling of absolute security. It is this objective notion that matters to the arguments of this chapter. However, there are obvious links between it and subjective requirements on knowledge. For, as Descartes so forcefully argued, the only plausible source of certainty in the objective sense derives from various operations of the conscious mind -- in particular, from introspection and intuition. And so, if we can achieve knowledge with objective certainty, then we will also, as it happens, have "subjective warrants" available, in that we will always be able to tell introspectively that our knowledge has come from these putatively infallible conscious sources.

From the reliabilist point of view, by contrast (henceforth I will use "reliabilism" to mean the view that only reliability is required for knowledge), any subjective requirements on knowledge are gratuitous. For, in order for a belief-forming process to be reliable, there is no need for its reliability, or even its existence, to be available to consciousness. According to reliabilism, we will know, say, that there is a table in front of us, just in case the unconscious visual processes that give rise to such perceptual beliefs generally deliver true beliefs,

whether or not we are aware of this. There is therefore no pressure, given reliabilism, to reconstruct such perceptual knowledge as first involving some infallible introspection of some sensory idea, and then some intuitively compelling inference from this idea to the presence of a table. If the demand for certainty in knowledge is unmotivated, then so too is this reconstruction of perceptual knowledge as involving infallible inferences from infallible introspections.

5.4 Knowledge and Normativity

It is sometimes felt that reliabilist epistemology changes the subject. As I have just pointed out, reliabilism implies that whether or not we know will often hinge on matters, such as the reliability of some visual process, which lie quite outside our consciousness. But this seems to imply that we are at the mercy of nature, that we cannot do anything to affect whether or not we know. And this t hen makes reliabilist epistemology seem a quite different subject from the traditional version (henceforth I shall use "traditional" to refer to views according to which knowledge requires certainty or subjective warrants[3]). For surely a central concern of traditional epistemology was the normative question of what we should do in order to ensure that our beliefs are knowledge.

However, this reaction to reliabilism involves a fallacy. It is true that traditional conceptions of knowledge offer advice on how to achieve knowledge: roughly, you should consciously monitor your thought processes, and avoid any which are not necessarily infallible. And it is true that reliabilism does not concur in this advice. But this is not because reliabilism has stopped offering advice on how to know, but simply because reliabilism offers different advice.

Where traditionalists advise aspirant knowers to monitor what goes on in their conscious minds, reliabilists will simply advise them to take whatever steps are needed to bring it about that their beliefs come from reliable processes. Such steps may well call for us to influence processes which lie outside consciousness, but that is no reason to conclude we cannot succeed. After all, most of the things we influence lie outside consciousness, like our environments, our physical health, and so on. Similarly, there is no reason why we cannot influence non-conscious aspects of our belief-forming processes, by such means as rote learning, adjusting the working of instruments we rely on, and so on[4]. (Reliabilists will allow that conscious monitoring is one way to improve the reliability of our belief-forming processes. But it is not the only way.)

To guard against a possible misunderstanding, let me emphasise that I am not suggesting that it is a requirement on knowledge that knowers must take active steps to bring it about that their beliefs are knowledge. I am aware that I began this chapter by identifying the concept of knowledge as the state someone concerned to avoid error (a "concerned enquirer" henceforth) wants to get into as a means to avoiding error. But it does not follow from this that the only way to be a knower is to take active steps to get into that state. F or you may already be in the requisite state, not because you did anything to make your belief-forming processes reliable, but simply because they were reliable to start with. Such passive knowers will already be in the state concerned enquirers aim to get into, even though they do not themselves share the concern to get into that state.[5]

5.5 Rationales versus Intuitions

I have just argued that reliabilism does not abandon the traditional normative issue of how best to acquire knowledge. In this section I want to consider a rather different argument for thinking that reliabilism changes the subject. This appeals, not to considerations of

normativity, but directly to intuitions about knowledge. Many philosophers take it to be intuitively obvious that subjective warrants are part of the concept of knowledge. And so they conclude that reliabilism, which dispenses with such requirements, must be wrong.

My response to this is that intuitions are not the only way to evaluate a theory of knowledge. There is a vast contemporary literature which aims to decide between reliabilism and traditional theories solely by appeal to intuitions about ingenious test cases.[6] Unfortunately, however, these intuitions weigh on both sides, and the literature based on them is notoriously indecisive. By contrast, my approach in this paper has not appealed to intuitions, but has tried to identify an underlying concept of knowledge, by locating the role it plays in our thinking, by trying to understand why knowledge is such a matter of concern to human beings.

My suggestion has been that knowledge is tied up with our desire to avoid false beliefs: it is the state a concerned enquirer needs to get into as a means to achieving this desire. Given this identification of the concept of knowledge, so to speak, we can then investigate more detailed conceptions, or theories, let us call them, of what that state is, more detailed theories of exactly what state an active enquirer needs to get into as means to avoiding error. The theory I am defending is that the requisite state is acquiring a belief from a reliable process. The theory that certainty and hence subjective warrants are required is a different theory, but still a theory, in the terminology I am using, of the same concept of knowledge.

This is why I think that any intuitions which may favour the traditional theory over the reliabilist alternative are indecisive. If I have shown that the reliabilist theory is the right theory, in that acquiring a belief from a reliable process is indeed what is needed as the means to avoid error, then I have therewith shown that the traditional theory and the intuitions that support it are mistaken.

Perhaps there remains a gap here. Suppose it is granted that I have identified the underlying concept of knowledge successfully, and have shown that reliabilism is the theory of knowledge that best fits it. An obvious question which then arises is why anybody should have had contrary ideas about certainty and subjective warrants in the first place. For, as I have told the story, reliabilism, as opposed to the traditional theory, follows pretty quickly from the concept of knowledge, thus making it mysterious why anybody should ever have thought anything more was needed. We cannot simply rest with the suggestion, offered briefly in passing earlier, that traditional ideas about knowledge may have been fostered by excessive optimism about what can be achieved. For the possibility of achieving something does not explain why we should want it, if it is not already desirable.

I shall return to this issue in section 5.8. But first let me comment briefly on the similarities between the strategy outlined in this section and that defended by Edward Craig in "The Practical Explication of Knowledge" (1986). At a detailed level, Craig's views differs from mine: he offers a third-person account of the concept of knowledge, by contrast with my first-person account, arguing that knowledge is the state our informants need to be in, for us to avoid error, not the state that we ourselves need to be in. But at the level of general strategy, Craig and I are in accord, in that he too seeks to offer an account of the point of the concept of knowledge, and to use this account to explain the nature of the concept, rather than simply trying to identify the concept from intuitions alone.

As to our differences, my objection to Craig's line would be that he is in effect focusing on the special case in which we succeed in avoiding error by acquiring beliefs from informants

who succeed in avoiding error. I accept that this special case may well have been of primary significance in the historical development of the everyday concept of knowledge, in that worrying about your informants' reliability calls for rather less sophistication than worrying about your own reliability (cf. Craig, op. cit., p 215). But, even so, Craig's third-person focus seems to me to have the disadvantage of cutting the link with the traditional normative issue of what we should do to avoid error. That is, even if it is unfaithful to the history of the concept of knowledge to view the desire for good informants as a special case of a general desire to have good belief-forming processes, I would argue that the more general perspective I have adopted nevertheless has the advantage of showing how the concept of knowledge relates to familiar philosophical worries about knowledge.

5.6 Knowledge and Percentages

How much reliability should a reliabilist require for knowledge? I shall consider two dimensions to this question. First, I shall consider whether we need 100% reliability, or whether some lesser percentage, such as 95%, say, is enough. Second, I shall ask over what range of possibilities the relevant percentage is to be assessed.

On the percentage question, it would be a mistake to think that the rejection of certainty has already decided this question in favour of something less than 100%. For "certainty", as I have been using it, implies that a belief-forming process cannot go wrong, will deliver 100% true beliefs in all possible worlds. This is a much stronger requirement than 100% reliability in this world. So we can reject certainty and still uphold a requirement of 100% contingent reliability. On the other hand, even the latter seems a fairly strong requirement. So perhaps we should consider arguments in favour of some lesser percentage.

On the second question, about the range of possibilities, the argument so far has shown it would be a mistake to require this degree of reliability to hold up across all possible worlds. But, as we shall see, there may remain reasons for wanting it to hold up across a t least some counterfactual situations.

Let me deal with the percentage question first. I shall return to the question about the range of possibilities in the next section. Up to a point, it is possible to by-pass the percentage question. Suppose a given belief-forming process delivers beliefs which are true 95% of the time. Then the appropriate output from that process would not be a full belief in the first place, but a 0.95 degree of belief. After all, if you believe it is going to rain tomorrow on the basis of a 95% reliable method of forecasting, you would be ill-advised to bet a million pounds to a penny, or indeed to stake anything more than nineteen to win one, on its raining tomorrow. So if knowledge unqualifiedly requires belief, as I have implicitly been assuming throughout, then this in itself seems to call for belief-forming processes which deliver truths with 100% reliability.

Still, perhaps it is a bit quick to assume that knowledge requires strictly full belief. After all, in everyday discourse we certainly refer to beliefs of high, but less than strictly full, degree as "beliefs" simpliciter, and to that extent we should expect the notion of knowledge also to encompass sufficiently well-founded beliefs of high, but not full, degree. But, having said this, there is then an obvious answer to the question of how well-founded such a belief of high, but not full, degree needs to be to qualify as knowledge. For, under the present suggestion, everyday discourse has certain standards, perhaps varying from context to context, of how firmly a belief has to be held to qualify as a belief simpliciter. So why not simply incorporate those standards into our analysis of knowledge, and say that for a belief to

be knowledge it should come from a process whose reliability is at least sufficient to warrant the degree of confidence required for the belief to qualify as a belief simpliciter in the first place, and not just as what even everyday discourse would consider as a partial belief?

It will be helpful for what follows to observe that, while it is certainly true that we often allow beliefs of less than strictly full degree to qualify as knowledge, there is also a practical sense in which it is always better to get a belief from 100% reliable processes. To put it simply, the reason is that such beliefs will then be true, and so decisions informed by them will succeed with probability one, whereas, if those beliefs came from less than 100% reliable processes, then the actions they informed would be less likely to succeed.

Actually, this puts the point rather too simply, in that even if your beliefs are proof against error, they may still not be informative enough to tell you how to achieve some result; and, even if your beliefs are informative enough, you may fail to draw the inference correctly. But these two caveats would apply equally even if the same set of beliefs came from less than 100% reliable processes, and so do not affect the point that it is always preferable, from the point of view of achieving your desires, to get a full belief from a 100% process, rather than a less than strictly full belief from a less reliable process.

It should also be admitted that in many cases the extra costs of getting 100% reliability will not be worth the extra probability of success, in which case we will do better to settle for a partial belief. This is no doubt why everyday discourse does not make strictly full belief a precondition of knowledge in general. But this merely calls for a yet further qualification, and still does not affect the underlying point that, when costs are equal, full belief from 100% reliable processes is always better.[7]

5.7 Nearby Possible Worlds

Let me now turn to the second question raised at the beginning of this section: what range of possibilities is relevant to the reliability of belief-forming processes? At first sight it might seem to follow from my overall argument that reliability in the actual world is all that matters. After all, as I observed earlier, reliability in this world is all that we need in order to avoid error. However, there are good reasons why knowledge calls for more than merely this-worldly reliability.

Let us return to the idea that knowledge is the state concerned enquirers need to get into in order to avoid error. It is true that concerned enquirers have no interest in reliability in non-actual worlds as such. Nevertheless, in acquiring processes which are reliable in this world, concerned enquirers will inevitably acquire processes which are reliable in a range of non-actual situations as well.

The reason is that, if you are an concerned enquirer, you will not be able to anticipate the future in enough detail to be able to tell exactly when you are going to use any given belief-forming process, and so will not know exactly which truths that process needs to deliver in order to be reliable in the actual world. Instead, you will inevitably have only limit ed knowledge about the general nature of the world and your particular situation in it[8], information which will narrow down the range of circumstances you may in future find yourself in, but it certainly will not tell you exactly what they will be. So, in aiming for reliable belief-forming processes, you will inevitably be constrained to aim for belief-forming processes which will reliably deliver true beliefs across the entire range of possible circumstances that your current information leaves it open you may end up in. Since not all

the possibilities in that range will become actual, you will inevitably be aiming to get into a state which would deliver true beliefs in various nearby possible worlds, as well as in the actual one.

So my overall approach to knowledge accommodates the requirement that knowledge should have a certain degree of counterfactual reliability. Note, however, that this is still a long way short of requiring reliability in all possible worlds, or even reliability in all causally possible worlds. For the information already possessed by concerned enquirers will still in general be enough to rule out the possibility of their being in most possible worlds (such as the world where you are manipulated by Descartes' evil demon) or even in most causally possible worlds (such as the world where you are a brain in a vat).[9]

5.8 The Attractions of Certainty

I promised to return to the question of why certainty should be intuitively plausible as a requirement for knowledge. A serious answer to this question would include an historical dimension, examining the development of Western epistemological notions, with particular reference to the mediaeval distinction between demonstration and o pinion, to the seventeenth and eighteenth century struggles to find a place for the newly emerging scientific knowledge within this distinction, and, perhaps most important of all, to the religious dimensions which so animated the participants in these de bates. However, any such historical investigation is beyond the scope of this book. Instead let me offer a possible philosophical explanation of the pull of certainty, not as a competitor to an historical account, but as a possible complement.

I have just argued that knowledge requires not only reliability in the actual world, but reliability across all worlds which are possible, relative to the information open to concerned enquirers. However, in discussing this issue of counterfactual reliability, I have so far implicitly been taking it for granted that our notional concerned enquirers are aiming for full beliefs from 100% reliable processes. But, as we saw earlier, in many practical contexts it will often be more efficient to settle for less than full beliefs, delivered by belief-forming processes of appropriately high, but less than perfect, reliability.

Now, an enquirer who was concerned to acquire such a less than full belief would be entitled to ignore, when assessing the reliability of the relevant belief-forming process, not only all worlds which are impossible relative to his or her current information, but also any worlds which fall below an appropriate threshold of probability relative to that current information. For clearly the fact that the process would be unreliable in such unlikely circumstances does not give a concerned believer sufficient reason to withhold a high degree of belief from its deliverances.

Add to this last point the consideration, elaborated in section 5.6, that, although it is often perfectly sensible to settle merely for a high degree of belief, it is always better, especially where it is important that your actions will not fail, to get full beliefs from 100% reliable processes. Putting these two points together, it follows that your knowledge will get better the more possible circumstances with any non-zero probability your belief-forming processes are reliable across.

This does not of course amount to a good argument for thinking that the best thing would be to acquire beliefs from processes that are reliable across all possible worlds. Even if it is a good thing to be reliable across all worlds with non-zero probability relative to current knowledge, this is a long way short of reliability across all possible worlds. For, as I have

said, most possible circumstances will be downright impossible relative to the information available to any concerned enquirer. Nevertheless, one can see how it would be easy to slide, from the thought that you need ideally to guard against any possibilities that your information leaves with any non-zero probability, to the thought that you need ideally to guard against any possibilities whatsoever.

5.9 The Problem of Induction? What Problem?

I turn now to the problem of induction. Let us suppose, for the sake of the argument, that the general form of induction is simple enumeration. (I do not really think this is a good model for inductive inferences. But it will help the exposition to assume so for a while.) So, for example, from the premises, that N ravens have been black so far, we conclude that all ravens are black. Schematically,

(1) Fa1 & Ga1
.
.
FaN & GaN

_____

All Fs are Gs

The traditional complaint about this form of inference is that it is logically invalid. The conclusion does not follow logically from the premises. It is logically possible that the premises be true, but the conclusion be false. For reliabilists, however, this complaint has no force. Since a form of inference can well be contingently reliable without being logically guaranteed, reliabilists can simply respond to the traditional complaint by arguing that the illogicality of inductive inferences is no reason to deny that such inferences yield knowledge.

Perhaps it is worth pausing briefly to explain how the notion of reliability applies to inferences. Though I have not treated this explicitly so far, the appropriate notion is obviously conditional reliability: the conclusion should always be true in the actual world, if the premises are. (This will then ensure, for a reliabilist, that reliable inferences will transmit knowledge, that they will yield known conclusions when applied to known premises. For if the premises are known, in the sense that they are true and reliably arrived at, then any conclusion derived from a conditionally reliable inference will also be true and reliably arrived at, and so known.)

It is tempting to leave the problem of induction here, with the observation that the logical invalidity of induction does not mean its conclusions are not knowledge. However, I suspect that most readers will be unpersuaded by this quick way with inductive scepticism, even if they are persuaded by the general arguments for reliabilism. So in the next X sections I shall consider whether there are any further reasons why a reliabilist should worry about induction. Accordingly, I shall now take it as given that reliabilism is the right account of knowledge in general; the issue to be considered is whether any sceptical doubts about induction still arise within this assumption.

5.10 Is Induction Reliable?

One possible worry about the simple reliabilist response to the problem of induction sketched in the last section is that it seems little different from the "analytic justification of induction" proposed by Edwards (1949) and Strawson (1952, Ch. 9). Yet it is now widely agreed that inductive inferences cannot be shown to be legitimate simply by observing, as the "analytic

justification" does, that most people would characterize induction as a central case of "rational" thinking. For such facts of common usage leave it open that there may be underlying requirements for a form of reasoning to be rational, which are not in fact satisfied by induction, are that most people may therefore be in error in holding induction to be rational.

However, the reliabilist response to induction is quite distinct from the analytic justification. Reliabilists do not accept a form of reasoning as rational just because it is widely regarded as "rational", but only insofar as it satisfies the underlying requirement of reliably delivering truths.[10] In particular, reliabilists will deem induction to be rational, and its conclusions therefore knowledge, not because it is called "rational", but because they believe that it is in fact a reliable method of getting new truths out of old ones.

However, this now points to an obvious problem. That induction reliably generates truths is itself a substantial contingent claim. Yet no support has so far been offered for this. We reliabilist friends of induction seem simply to be taking it for granted that induction is a reliable method of inference, and then concluding, in virtue of our general reliabilism about knowledge, that induction yields knowledge. But what basis do we have for the initial assumption that induction is reliable?

Some reliabilists are inclined to respond, at this stage of the proceedings, that we do not need to know that we know in order to know.[11] I think this is the wrong move. It is perfectly true, of course, that ordinary non-philosophical knowers do not need to know that they know. But the present demand for a defence of the claim that induction yields knowledge is not being made of ordinary knowers who are using inductions, but rather of us philosophers who are talking about induct ions, and in particular about the question of whether inductions yield knowledge. We reliabilist friends of induction are explicitly claiming that inductive inferences yield knowledge because they reliably yield truths. Given this, it is perfectly reasonable for someone to challenge us to provide support for this claim.

Of course, if we fail to meet this challenge, this will not necessarily show that induction does not yield knowledge. To lack any grounds for accepting the reliability of induction is not yet to have grounds for denying it. But such a stand-off would be a failure for us friends of induction, and a success for our sceptical challengers. The point at issue is whether induction yields knowledge, that is, given reliabilism, whether induction reliably generates truths. We friends of induction say yes, our sceptical challengers ask for support for this claim. If we can't answer them, then they will have succeeded in showing we aren't entitled to our stance.

So to uphold induction as a source of knowledge we need to show that inductive inferences are reliable. However, now we are clear about this need, I do not think it is hard to satisfy. The obvious way to find out whether induction is reliable is to examine such evidence as bears on the matter. When people make inductions, do their conclusions turn out to be true? There are plenty of past examples of people making inductions. And when they have made inductions, their conclusions have indeed turned out true. So we have every reason to hold that in general inductive inferences yield truths. That is:

(2) When person1 induced, from N observations of A going with B, that All As are Bs, this conclusion1 was true

When person2 induced, from N observations of C going with D, that All Cs are Ds, this conclusion2 was true

When personN induced, from N observations of L going with M, that All Ls are Ms, this conclusionN was true

---

Whenever someone induces, their conclusion is true.

5.11 The Legitimacy of Normal Methods

Let me first put to one side two obvious worries about the premises of this argument. First, aren't there plenty of past examples of unsuccessful inductions with false conclusions, as well as successful ones with true conclusions? Second, how can we know that even the successful inductions are successful, given that observation of the past will only show, for example, that As have been Bs so far, not that all As are Bs? I shall deal with both these points in due course (in section 5.17 and footnote 14 respectively). But for the moment it will helpful to ignore them, and attend instead to the move from the premises to the conclusion of (2). For it will not have es caped the notice of most readers that this is itself an inductive inference, of just the kind whose reliability we are presently concerned to investigate.

However, is there anything wrong with this? It is not as if the discussion so far has identified some flaw in induction, of a kind which would imply that it ought to be eliminated from the battery of procedures by which we normally arrive at our beliefs. In particular, we have agreed that the logical invalidity of inductive inferences in itself casts no discredit on induction. Given this, when a certain question of fact is raised -- namely, are inductive inferences always reliable? -- what is more natural than to try resolve this question by means of our normal procedures of investigation, which include, as it happens, our inductive procedures?

I know that to some philosophical sensibilities this will seem unduly complacent: surely we aren't entitled to any methods of investigation, until we have demonstrated their worth. But where are we supposed to start? We certainly need to begin with some methods of thought, lest we lapse into philosophical catatonia. Many philosophers, I realize, will want to follow Descartes, and restrict our initial methods to introspection and intuition. But Descartes' rationale for this restriction was that it promised certainty, and we have already agreed that this is an unnecessary desideratum on our methods of thought. It would seem equally sensible to continue with our normal methods of thought, at least until we uncover some reason to distrust them. And these methods will include induction, since, to repeat, we have not as yet been given any reason to distrust induction.

5.12 Varieties of Circularity

Still, even if nothing has as yet been shown wrong with induction in general, it may well be felt that there is something wrong with the inductive argument (2) in particular. For isn't (2) a circular argument, and therefore illegitimate? This objection needs to be treated carefully. It is true, as we shall shortly see, that circularity of a certain sort is present in (2). However, provided we keep firmly in mind the specific argumentative task to which (2) is directed, we shall also see that this circularity is not damning.

As a first step, we need to distinguish between "premise-circularity" and "rule-circularity".[12] An argument is premise-circular if its conclusion is contained among its premises. An argument is rule-circular if it reach es the conclusion that a certain rule of inference is reliable

by using that self-same rule of inference. Clearly premise-circularity is a vice in an argument. The point of an argument is to take us from old beliefs, which we already accept as premises, to some new belief as a conclusion. But if the conclusion is already contained in the premises, then the argument will fail in this primary task. However, argument (2) is clearly not premise-circular. It is a genuinely expansive argument, whose conclusion, that all inductions yield true conclusions, manifestly outruns its premises, that N inductions so far have done so.[13]

On the other hand, argument (2) is rule-circular. Even if the claim that induction is reliable does not appear a mong its premises, it does use an inductive inference to reach its conclusion that induction is reliable. I have a number of comments to make about the rule-circularity of (2). But first let me make a wider comment, not about argument (2) in particular, but about rule-circularity as such: namely, that it can scarcely be a general requirement, on all legitimate forms of inference, that it be possible to show that they are reliable in some non-rule-circular way. For this would disqualify even deduction as a legitimate form of inference. (While it is possible to demonstrate that deductive inferences are reliable -- indeed necessarily reliable -- by means of the standard semantic soundness proofs, these demonstrations themselves unquestionably employ deduction.) So the fact that induction can only be shown reliable in a rule-circular way, as in (2), certainly does not in itself yield any immediate reason to distrust induction.
But this is merely to repeat the point that we have as yet been given no good argument for distrusting induction. Our current concern, however, is whether (2) takes us beyond this, and gives us a positive basis for trusting induction, despite the fact that it is admittedly rule-circular.[14] Let us recall the context of argument in which (2) was put forward. We agreed, on general reliabilist grounds, that induction does not need to be logically valid to yield knowledge, but will yield knowledge just in case it is reliable. However, the sceptic then pointed out, we cannot just take the belief that induction is reliable for granted. To which we responded that we are not taking this for granted, but have a good argument, based on empirical evidence, for the conclusion that induction is reliable, namely, argument (2). It seems to me that, in this specific context, the context of showing a sceptic who accepts reliabilism that we are not just helping ourselves to the belief that induction is reliable, (2) does just the job it is required to do.

5.13 Who Needs Persuading?
Perhaps the best way of showing this is by detailing some of the task's argument (2) is not intended to fulfil. For a start, we should recognize that argument (2) would be no good for persuading people who do not make inductions to start making them. While the conclusion of (2), that inductions are reliable, would certainly be a good reason for such people to start inducing, if they accepted it, they clearly will not be persuaded to do so by (2), for the route from (2)'s premises to (2)'s conclusion requires just the kind of inductive inference that they eschew. In particular, then, (2) will be no good for persuading people who have already reflected on the reliability of induction, and have been persuaded, for whatever reasons, that they ought to stop performing inductions, that they ought to start again.

However, in the present context of argument, this is no demerit in (2). Argument (2) is not addressed to people who avoid inductions. We may yet discover good reasons for avoiding inductions, and indeed in the next few sections I shall examine some possible such reasons, but right now we are assuming that nothing has yet been shown wrong with induction, and are considering whether, given this, argument (2) can show us whether induction is reliable.

So (2) should be thought of as addressed to people who have not yet been given any reason to distrust induction. And (2) ought surely to persuade such people at least of its conclusion.

I realise that many readers will feel that, if (2) is a good argument in defence of induction, then it ought to be capable of persuading any intelligent being, with whatever epistemological habits. But this is an extremely strong demand, and it is not at all clear why we should accept it. The only plausible rationale, once more, seems to stem from the assumption that knowledge requires certainty, together with the assumption that the only kinds of belief-forming processes which can plausibly deliver certainty are conscious operations whose logical infallibility is introspectively available. Together these assumptions imply that any source of knowledge ought in principle to be recognizable as such by any conscious beings, in virtue of their introspective abilities; and hence these assumptions imply that a good argument for the legitimacy of some source of knowledge ought to persuade all people whatsoever, however wrong-headed their starting position. However, once we reject the assumption that knowledge requires certainty, then this whole line of reasoning falls away, and the strong demand that a good defence of induction ought to persuade any conscious being is left without any obvious means of support.

These last remarks bear on the question of "counter-inductive" arguments for "counter-induction". It is often observed that inductive arguments for induction, like (2), have counter-inductive mirror images. Counter-inductivists, when they observe that a number of As are all Bs, conclude that the next A will not be a B. When it is pointed out that this is illogical, they can respond, "So what? Illogical it may be, but this doesn't show that it is not in fact a good way of reliably reaching true conclusions". And when we say, "All right. But what basis do you have for supposing that counter-induction does in fact deliver true conclusions?", they reply, "Ah, so that's what you're worried about. Let us then look at the evidence that bears on the question. On a large number of occasions in the past people have counter-induced, and have been led to false conclusions. So we conclude -- counter-inductively -- that the next time we counter-induce we will get a true conclusion".

There is room to dispute whether this is in fact a perfect mirror image of (2) (Cf Van Cleve, op. cit., footnote 16.) But let that pass. The more important point is that, even if counter-inductivists can mirror (2), this does nothing to discredit (2) itself. I have already conceded that (2) is not going to persuade people who do not make inductions to start making them. Counter-inductivism now simply gives us a further example of people who have abnormal inferential dispositions, and who will not therefore be persuaded by (2). Except that the parable of the counter-inductivist s adds an extra twist, namely, that counter-inductivists will be persuaded, by their mirror of (2), to the conclusion that their abnormal counter-inductive dispositions are reliable. But all this leaves (2) untouched. We should not expect it to perform the impossible task of knocking imaginary non-inductivists out of their non-inductivism -- its task is only to allow normal people, like ourselves, to resolve the issue of whether induction is reliable.

By this stage, some readers may be feeling that argument (2) does not do very much. Indeed, if it only works for people who already make inductions, is it really doing anything at all? My answer is that it is not supposed to do very much. Nearly all the serious work was finished before (2) came on the scene. Most importantly, the general arguments for reliabilism have already shown that the logical invalidity of induction is not a problem. Argument (2) is just supposed to show that, given that there is nothing problematic about induction, then there is no barrier to our concluding that it is reliable, and hence that it yields knowledge.

Trained philosophers naturally expect a "justification of induction" to do something to rehabilitate induction, in response to an argument that there is something wrong with it. But (2) is not meant as a "justification" in this sense. So we should not condemn it for its failure to be one.

5.14 The Strategy Generalized

As a first step towards generalizing the anti-sceptic al strategy outlined in the last few sections, let us be more realistic about induction. I have already noted one way in which the above discussion of induction has involved an idealization, namely, in respect of the assumption that in our experience all past enumerative inductions have been successful. This assumption is of course manifestly false. There are plenty of good examples of enumerative inductions leading to false conclusions, from Russell's chicken who expected to be fed every day, to the Newtonian physicists who expected acceleration always to be inversely proportional to rest mass.

In any case, apart from such direct evidence, there is also a principled argument to show that simple enumerative induction cannot possibly be a reliable method of inference. I refer to Goodman's "new problem of induction". Goodman (1954) shows that there are far too many ways of classifying events, far too many As and Bs, for every instance of schema (1) to yield a true conclusion. Indeed, Goodman shows how to construct, for every instance of (1) that might yield a true conclusion, an infinity of other instances which will then yield false conclusions.

These are good arguments against enumerative induction. It is important to recognize, however, that they are quite independent of the traditional objection to induction. They do not just make the point that enumerative induction is logically invalid. On the contrary, they show that enumerative induction is not just invalid, but downright unreliable.

The moral, for us reliabilists who want to resist scepticism about induction, is that we had better not take our stand on simple enumerative induction as schematized in (1). Rather, we need somehow to show that our actual inductive practice has a more sophisticated structure, perhaps involving restrictions on the As and Bs which are candidates for projection, and perhaps limited in the degrees of belief which we extend to its conclusions.

I shall say a bit more about such an alternative model of inductive inference shortly. But first let me observe that such a model will open the way to the anti-sceptical strategy outlined above once more. Imagine that we can show that our actual inductive practice is more sophisticated than simple enumerative induction, and that it therefore cannot be discredited as unreliable by either Goodman's new problem or past performance. And imagine, furthermore, that when we investigate the reliability of our inductive methods, using existing metho ds of investigation, including those inductive methods themselves, we find ourselves able to conclude that it is reliable. Then this defeats scepticism about our inductive practice. As before, neither the fact that this practice may be invalid, nor the fact that its reliability might only be discoverable in a rule-circular way, will be a barrier to our concluding that it yields knowledge.

In this chapter I have been concentrating on induction. But the anti-sceptical strategy I have used can be generalized to apply to our belief-forming methods in general, including such non-inferential methods as perception and memory. If the only objection to them is that they

are not certain, in the sense that it is possible that they should yield false beliefs, then this is no reason to believe that they are not reliable. And if, moreover, investigation shows that those methods are reliable, then, any rule-circularity notwithstanding, we will be in a position to conclude that they yield knowledge.[15]

5.15 Non-Enumerative Induction

I now want to outline a more realistic model of induction, with the intention of showing that induction is not in fact discredited as unreliable by either Goodman's argument or by past performance.

To start with Goodman, note that any solution to Goodman's problem is likely to lead us to view induction in terms of elimination rather than enumeration. For any solution will involve some limitation on the As which are possible candidates for being associated with any given B. But if we have such a limited range of possible As, then it should be possible for us to find experimental data which will identify the actual antecedent of B by eliminating the other candidates, rather than by providing repeated instances of the relevant generalization.

Within the context of deterministic assumptions, J.S. Mill's methods of induction show what kinds of data are required to reach such eliminative conclusions, given various kinds of assumptions about limited ranges of possible antecedents. Much contemporary science, it is true, does not assume determinism. But there are probabilistic analogues of Mill's methods, which use techniques of analysis of variance and multiple regression to discriminate, among the factors which might in principle be correlated with some effect B, those which are genuinely rather than spuriously correlated with it.[16]

Mill's and related methods are nowadays little discussed by philosophers. This is unfortunate, for Mill's methods are clearly far more in accord with actual scientific practice than the standard philosophical model of induction by enumeration of instances. Science does not need a a large number of repeated observations to establish that copper melts at 1084°C, or that chickenpox is caused by a herpes virus, or that water is H20. Rather, since there are only so many possible melting points, or infectious agents, or combinations of elements, a few relatively simple observations will suffice in each such case to discriminate the actual law from the initially possible alternatives.

No doubt part of the reason philosophers have been uninterested in Mill's methods is that Mill himself does not offer any satisfactory response to sceptical doubts about induction. But the argument of this chapter shows how we can defend Mill's methods against such sceptical doubts: namely, by showing that those methods are a reliable source of true beliefs, and so of knowledge. Moreover, the argument of this chapter shows that there is no reason why such a demonstration should not be rule-circular: what we want is some route to the conclusion that those methods are reliable, but not necessarily a route that avoids those methods themselves.

In the rest of this section I will sketch out one possible route to the conclusion that Millian methods of eliminative induction are reliable.[17] But before I do so, let me observe that a demonstration of the reliability of Mill's met hods will also constitute an implicit answer to Goodman's new problem of induction. For, in order for Mill's methods to be reliable, only a certain limited range of As can possibly associated be with any given B. So any demonstration of the reliability of Mill's methods will need to show that these As are indeed the only candidates for projecting along with B. That is, if Mill's methods work, then there must be general reasons why only some sorts of generalizations -- like "all emeralds are

green" -- are on the cards, and that others -- "all emeralds are grue" -- are not. A demonstration of the reliability of Mill's methods will thus show why green is projectible with respect to emeralds and grue is not.

Let us look a bit more closely at Mill's methods. It is important, in thinking about the reliability of these methods, not to view the kind of background assumption which tells us, say, that B has a deterministic antecedent, and that A1, . . ., An are the only possibilities, as a premise to a Millian inductive inference, to which we conjoin the further, observational premises that A1, . . ., An-1 have been found without B, to conclude that An is the actual antecedent.[18] For this would just make Millian induction a special case of deduction, and moreover it would leave the scientists who engaged in Millian inferences with an undischarged premise, namely, the premise that one of the A1, . . ., An is the deterministic antecedent of B.

Rather, we should think of the observation of A1, . . ., An-1 without B as the sole premise to a Millian inductive inference. We don't need to suppose that the scientists themselves know why this inference works (though they probably will); all that matters, in the first instance, is that they are disposed, on observing A1, . . ., An-1 without B, to conclude that An is the actual antecedent.

The further thesis that B has a deterministic antecedent, and that A1, . . ., An are the only possibilities, need only come in at the philosophical meta-level, when we address the question of the reliability of the Millian inductive inference. It is we philosophers, who want to ascertain whether the Millian inference is reliable, who need to know that one of the A1, . . ., An must be the deterministic antecedent of B, not the scientists who actually make this inference.

So the scientists who make Millian inferences are not necessarily guilty of helping themselves to undischarged premises. The only premise they need is that A1, . . ., An-1 have been found without B, and that they can get from observation. On the other hand, the complaint of undischarged premises can reasonably be levelled at a philosopher, like myself, who explicitly argues that such Millian inductive inferences are reliable, on the grounds that one of the A1, . . ., An must be the deterministic antecedent of B. For I at least then owe some account of my basis for this latter claim.

When we were still thinking of induction as simple enumeration, this was the point where we turned induction on itself, and used an inductive inference to arrive at the conclusion that induction is reliable. If we could do this in the present context, then we would once more have an answer to sceptical questions about induction. However, it is not so obvious that we can make the same move within the context of an eliminative approach to induction.

Consider the inference by which medical scientists establish that chickenpox is due to a herpes virus. We can construe the scientists as inferring this from a set of Mill-style observations about the presence and absence of various viruses in people with and without chickenpox. And we can explain the reliability of this inference on the grounds that viral agents are the only possible causes of infectious diseases that do not respond to antibiotics.

But now what about this latter claim, that viruses are the only possible causes of antibiotic-resistant infectious diseases? We need this claim to explain the reliability of the medical scientist's Millian inference. But where does it come from? The trouble is that we can't get it

from the Millian inference in question itself, since this form of inference can only tell us which viruses are the causes of which antibiotic-resistant infections, not which agents are responsible for antibiotic-resistant infections in general. That is, the Millian inference in question isn't self-supporting in the way enumerative induction is.

Perhaps we can still establish the general claim that viruses always cause antibiotic-resistant infectious diseases by another type of eliminative induction, different from the first. What we would need would be a type which uses suitable observations to eliminate all the other possible agents apart from viruses as the cause of such diseases, not one which eliminates all other viruses apart from the herpes virus as the cause of chickenpox.

Now, it is plausible enough that there is such a reliable mode of eliminative induction. But then the same problem arises again. For the reliability of this new mode of eliminative induction will now rest on some further fact, such as that only a certain range of invasive physical agents that disrupt the biochemistry of the sufferers are possible candidate cause s of diseases. And so the philosopher who wants to assert the reliability of this new form of eliminative induction has a new undischarged premise to cope with.

A kind of regress threatens. Eliminative inductions seem to fall naturally into a hierarchy, with the reliability of each being explicable only with the help of assumptions which derive from a form of inference higher in the hierarchy. We can explain the reliability of the procedures which show that a herpes virus causes chickenpox by invoking the assumption that antibiotic-resistant infectious diseases are always due to viruses. But this assumption in turn needs to be established by a different procedure, whose reliability depends on some such assumption as that all diseases are due to invasive physical agents that disrupt the biochemistry of the sufferers. And perhaps this assumption too can in turn be established by an eliminative induction, which uses relevant observations to discriminate between this assumption and other physically possible models of disease. But this then leaves us philosophers with the task of explaining how we came by the assumption that the physical possibilities are the only possibilities.

We can expect to find this kind of hierarchy repeated in different areas of science. Perhaps the fact that such sequences will characteristically go from subject-specific assumptions to general assumptions about physical possibility offers a way out of our problem. If the regress is to stop, at some point we will need to find a form of eliminative induction which can establish its own reliability, as enumerative induction did earlier. I would like tentatively to suggest that such a form of eliminative induction might be found at the level of basic physical science. At this level the task is to discover the limits of physical possibility itself. Basic physics aims to decide between different theories of force, matter and spacetime, between different theories of what is physically possible. Now, it is plausible that at this level the inductive strategy used by physicists is to ignore any such theories which lack a certain kind of physical simplicity. If this is right, then this inductive strategy, when applied to the question of the general constitution of the universe, will inevitably lead to the conclusion that the universe is composed of constituents which display the relevant kind of physical simplicity. And then, once we have reached this conclusion, we can use it to explain why this inductive strategy is reliable. For if the constituents of the world are indeed characterized by the relevant kind of physical simplicity, then a methodology which uses observations to decide between alternatives with this kind of simplicity will for that reason be a reliable route to the truth.

It should be emphasized that this story does not depend on any a priori notion of simplicity. To this extent the term "simplicity" is perhaps a misnomer. The account simply depends on the existence of certain general features which characterize the true answers to questions of fundamental physical theory. Far from being knowable a priori, these features may well be counterintuitive to the scientifically untrained. Thus circular motion is not especially "simple", in the relevant sense, compared to the kinds of motion that results from inverse square force laws. Discontinuity is not "simple", notwithstanding the fact that everyday experience shows us sharp boundaries between physical objects and their surroundings. Explanations in terms of observable causes is not "simple", compared with explanation by microscopic hidden mechanisms.

5.16 The Historical Contingency of Knowledge

It follows from this last observation that the methods of inference which enable us to find out about the physical world are not native to human thinking. At some point in human history people acquired the ability, which they did not have before, to focus specifically on certain kinds of explanations of physical phenomena, and to ignore others.

This might seem to raise a problem for my overall response to scepticism. My suggestion has been that we should check our existing methods of thought to confirm that they are reliable sources of true belief and hence of knowledge. And, in answer to the objection that these checks themselves employ those same existing methods of thought, I have observed that we have as yet been given no reason to distrust those methods.

However, some readers may feel that this strategy is markedly less plausible when applied to methods of thought which are historically happenstantial, rather than to those which are an innate part of the human cognitive endowment. Demonstrations of reliability that depend on such innate methods of thought at least have the virtue of persuading all actual human beings; it is only purely though-experimental individuals, like the counter-inductivists, who will be unpersuaded by such demonstrations of the existence of human knowledge. But demonstrations of reliability by appeal to historically contingent methods of thought like the preference for "physically simple" hypotheses will fail to persuade many actual human beings as well -- namely, all those individuals in the historical, geographical, or cultural distance who have not adopted the contingent methods of thought in question.

Still, does this matter? That other people foreswear some method of thought is not in itself a conclusive reason for distrusting that method of thought, nor, therefore, for distrusting vindications of human knowledge that depend on it. Maybe our reasons for thinking that we have knowledge, and that the sceptic is therefore wrong, will fail to persuade various actual human beings, as well as the imaginary counter-inductivist. But this scarcely shows they are not good reasons. After all, there is plenty of evidence to show that many methods of thought which are native to human thinking are untrustworthy, and that we therefore do better to replace these native habits by historically contingent alternatives.

Perhaps a kind of generalization from variability lies behind the widespread feeling that historically contingent methods of thought are epistemologically suspect. I have in mind the following line of argument: different people in different times and places have adopted many different procedures for deciding questions about, say, the causes of observable phenomena; since these different procedures standardly deliver mutually inconsistent answers, only one of which can be true, most of these procedures must be unreliable; so it is highly probable that

any given such procedure must be unreliable; in particular, therefore, it is highly probable that our currently preferred procedure is unreliable.

However, once it is spelt out, the weakness in this line of argument is apparent. For it presupposes that such a set of historically varying procedures forms a homogeneous category, whose overall inadequacies therefore detract equally from all members of the category. However, there may well be relevant differences between members of the category, differences which block the inference from "most of the procedures in the group are unreliable" to "this particular procedure is probably unreliable". For instance, it may be that some currently preferred procedure has been adopted as a result of critical reflection, rather than mere deference to tradition; perhaps it makes use of controlled experiment, rather than mere hearsay; maybe it involves mathematical precision, rather than mere guesswork. These features do not of course provide a conclusive demonstration that the procedure in question is a reliable source of truth. But they are surely sufficient to invalidate the argument that, since various procedures which lack these features are unreliable, our preferred procedure must be unreliable too.[19]

5.17 The Pessimistic Meta-Induction

The argument from relativism considered in the last section needs to be distinguished from the well-known "pessimistic meta-induction from past falsity". The "pessimistic meta-induction" also calls in question the reliability of the inductive strategies of modern science. But instead of maintaining that there is nothing to choose between these strategies and the incompatible alternatives which have been practised by historically and culturally distant humans, the pessimistic meta-induction focuses directly on the output of modern scientific method, and argues straight off that this scientific method must be unreliable, since it characteristically issues in false beliefs.

After all, the pessimists can point out, we now take it that Newtonian physics, the phlogiston theory of combustion, the theory that atoms are indivisible, and so on, are all false. So doesn't it immediately follow that method by which these theories were reached cannot be a reliable route to the truth?[20]

This pessimistic meta-induction is unquestionably an important argument, which indicates that caution is necessary in scientific theorizing. However, it raises many questions of detail, and it would extend our discussion unduly to treat it fully here. I shall restrict myself to a few brief comments, which I hope will suffice to show that, even if the pessimistic meta-induction advises caution, it does not mean that we should withhold belief entirely from all scientific claims.

The basic flaw in the pessimistic meta-induction mirrors that in the argument from variability discussed in the last section: it lumps into one homogenous category items that deserve separate treatment. The argument from variability lumps together all human thought processes. Similarly the pessimistic meta-induction lumps together all scientific theories.

In so doing, the pessimistic meta-induction ignores important differences between scientific theories, differences which matter to the question of whether the historical record casts doubt on their truth.. For a start, the tendency to theoretical falsity is much more common in some areas of science than others. Thus it is relatively normal for theories to be overturned in cosmology, say, or fundamental particle physics, or the study of primate evolution. By contrast, theories of the molecular composition of different chemical compounds (such as that

water is made of hydrogen and oxygen), or the causes of infectious diseases (chickenpox is due to a herpes virus), or the nature of everyday physical phenomena (heat is molecular motion), are characteristically retained once they are accepted. So the testimony of past form counts against some kind of theories more than others. Past scientific failures indicate caution about the thesis that quarks and leptons are the ultimate building blocks of matter. But they give us no reason to doubt that water is made of hydrogen and oxygen.

Nor need we regard this differential success-rate of different kinds of theories as an inexplicable historical datum. It may simply be a result of the necessary evidence being more easily available in some areas than others. Paleoanthropologists want to know how many hominid species were present on earth three million years ago. But their evidence consists of a few pieces of teeth and bone. So it is scarcely surprising that discoveries of new fossil sites will often lead them to change their views. The same point applies on a larger scale in cosmology and particle physics. Scientists in these areas want to answer very general questions about the very small and the very distant. But their evidence derives from the limited range of technological instruments they have devised to probe these realms. So, once more, it is scarcely surprising that their theories should remain at the level of tentative hypotheses. By contrast, in those areas where adequate evidence is available, such as chemistry and medicine, there is no corresponding barrier to science moving beyond tentative hypotheses to firm conclusions.

I do not necessarily want to suggest that there is a level of evidence which will ensure sure-fire inductive inferences, inferences that are 100% reliable in the actual world. Maybe even the best achievable evidence will on occasion lead scientists astray. However, by distinguishing between well supported and badly supported theories, we can at least avoid tarring all scientific conclusions with the failures of poorly supported speculations.

If even the best evidence is less than sure-fire, then we ought never to accord strictly full belief to scientific conclusions. Rather, we ought to tailor our degree of belief to the reliability of similarly evidenced conclusions, in the way indicated in section 5.6 above. But as long as the success-rate of well evidenced inferences is high, this degree of belief can still be close to one. Speculations based on meagre evidence may often turn out to be false. But this is no reason to think all inductive conclusions will suffer the same fate.

1. This terminology derives from Dummett, who introduced the term "anti-realist" for theories of meaning according to which truth is not evidence-transcendent. (Cf Dummett, 1978, p 146 and passim.)

2. Gettier cases show that we need an extra requirement. Not only must you acquire a true belief from a process that generally produces true beliefs, but the truth of your belief must not be an accident relative to its coming from that process. This extra requirement is a natural upshot of the hypothesis that knowledge is a state which is a means of avoiding error; for if the truth of your belief is a result of a Gettier-style accident, then your avoidance of error won't be a result of your embodying a truth-conducive process, in the sense that it won't be because the process generally produces truths that you have avoided error. Rather, it will be because of some lucky fluke. Cf. Williams, op. cit., pp 43-4. Further analysis is needed, of course, to make the relevant notion of accident precise.

3. What of the many philosophical views, especially contemporary views, which favour subjective warrants, but not certainty? I have two excuses for running the two requirements

together. First, I can think of no good rationale, as opposed to intuitions, in favour of the demand for subjective warrants, except as a corollary of the desire for certainty. Second, most of my arguments will apply as much to former demand taken separately as to the latter.

4. For more on the extent to which we can alter our non-conscious belief-forming processes, see my Reality and Representation (1987, ch 7.4, 7.6).

5. Perhaps passive knowers should have at least this much in common with concerned enquirers: their belief-forming processes should not just happen to be reliable, but should be present because they are reliable. This suggestion will rule out purely fortuitous reliability, but will allow in reliability due to evolution, learning, and education, alongside the case where concerned enquirers adopt processes because they consciously recognize the reliability of those processes. This suggestion also indicates a possible answer to the delicate question of how to individuate belief-forming processes for purposes of assessing their reliability: namely, as a first approximation, we should individuate them by the same characteristics as are needed to explain their adoption. Cf Papineau (1987, pp 136-8).

6. See Shope The Analysis of Knowing: A Decade of Research (1983).

7. I am here assuming beliefs about non-chance matters. Where genuine non-unitary chances are involved, the best degree of belief about any outcome will, of course, be different from one. But even here it will always be better to get true beliefs of full degree about chances, belief-forming costs apart, for such beliefs will then ensure that in general your decisions maximize your objective chances of success.

8. Some readers may be unhappy with these assumptions about the knowledge available to concerned enquirers. There are two possible worries here, one about a possible circularity on my part, the other about a possible circularity on the part of concerned enquirers. If you are worried about a circularity on my part, let me observe that my present concern is not to define knowledge, so much as to identify the role the concept plays in our thinking: I think it is helpful in this task to consider the predicament of a concerned enquirer who already has some knowledge; it would be a further task, which I shall not attempt here, and which would indeed preclude mention of the knowledge of concerned enquirers, to specify necessary and sufficient conditions, in non-epistemological terms, for someone to be a knower. If, on the other hand, you are worried about some kind of circularity on the concerned enquirer's part, on the grounds, perhaps, that any seriously concerned enquirer ought to start by assuming nothing, then your worry should be assuaged by the points made in sections 5.11-13 below.

9. Craig, op. cit., pp. 218-21, argues similarly from the limited informational situation of an enquirer to a counterfactual requirement on knowledge. But he is concerned with the kind of counterfactual reliability we want of our informants, given our limited information about the particular situation at hand, whereas I am interested in the counterfactual reliability we want of ourselves, given our limited general information about the situations we are going to be in. The precise degree of counterfactual reliability required for knowledge is a complex issue, which I shall not pursue any further here, except to observe that this is an area where one good theoretical rationale seems to me to be worth a thousand delicate intuitions.

10. This claim perhaps deserves further discussion. Some reliabilists would hold that, while reliability suffices for knowledge, some kind of extra subjective warrant is needed for

rationality. In my view, however, the arguments about knowledge go through for rationality. For a defence of the analogous point about "justification", see Goldman (1979).

11. Cf Van Cleve (1984) p. 559, 562. Much of my following defence of induction is influenced by Van Cleve's important article. However, in the passages referred to, Van Cleve seems to deny that reliabilists need to defend the reliability of induction, on the grounds that reliability is an "external" requirement. This seems to me an unfortunate slip, given that evaluations of reliabilist defences of induction are highly sensitive to prior judgements of exactly what the reliabilist needs to do.

12. Cf Braithwaite (1953, pp 276-7), Van Cleve, op. cit., p 558.

13. There is a problem about the notion of premise circularity: if "contained among the premises" just means logical implication, then all deductive arguments will be premise circular. Some philosophers, most notably Descartes and Mill, take this to show that deduction is uninformative. The majority prefer to understand "containment" more strictly. This debate is irrelevant to our current concerns, however, since everybody will agree that the non-deductive argument (2) is not premise circular.

14. If it does, then this will answer the second of the questions raised about the premises of (2) at the beginning of this section, namely, the question as to how we know that the conclusions ("All As are Bs") of past inductions are true. The answer is, by induction. For, if the distinction between rule- and premise-circularity legitimates the inductive move from (2)'s premises to (2)'s conclusion, then it will also serve to legitimate antecedent inductive moves from instances of As being Bs to "All as are Bs". Cf Van Cleve, op. cit., pp 560-1.

15. One of the aims of my Reality and Representation (1987) was to show how we might investigate all of our standard methods of belief-formation, as we might investigate any other natural phenomena, and discover that those methods are by and large reliable sources of truth.

16. These probabilistic analogues of Mill's methods involve two inferential steps: first from sample data to objective correlations, then from these correlations to causes. While this second step can be 100% reliable, the first cannot avoid the uncertainties of statistical inference. (Cf Papineau, 1993.) I have hopes for a reliabilist-style account of statistical inference, but there is no question of going into this here.

17. This will expand the discussion in Papineau (1987, pp 196-8).

18. This simple method of difference is just one of Mill's methods. For a full catalogue see the Appendix to Mackie (1974).

19. In Reality and Representation my response to the argument from variability was that there is no variability among self-certifying batteries of belief-forming procedures, that is, batteries of procedures which generate conclusions about their own reliability (1987, ch 10). I still think this argument can be made to work, but it now seems rather a big hammer to crack a small nut. From my present perspective, self-certification is merely one example of the kind of relevant difference which can block the generalization from variability.

20. Doesn't this argument undermine itself, by casting doubt on the present-day scientific theories whose truth it assumes when judging past theories to be false? But the argument can

be cast as a reductio: if current theories are true, then past theories are false; so, by the pessimistic meta-induction, current theories are false; so current theories are false. (Cf Jardine, 198x, p x.)

CHAPTER 6 MATHEMATICS AND OTHER NON-NATURAL SUBJECTS

6.1 Introduction

The arguments of the last chapter, together with those in chapter 3, amount to a realist package. In chapter 3 I argued for a realist theory of content, a theory which implies, as I observed in section 3.13, that there is no conceptual guarantee that our procedures for forming beliefs should yield beliefs which are true. Such a theory establishes a sense in which our beliefs answer to an independent world: the claims made by our beliefs conceptually outstrip the basis on which we form those beliefs.

This gap between the basis for judgements and their content creates room for sceptical threats. The theory of knowledge developed in the last chapter, however, shows how such threats can be dealt with. For even if there is no conceptual guarantee, issuing from the theory of content, that judgement and truth should co-vary, they might still co-vary as a matter of a posteriori fact. The arguments of the last chapter showed how this might work for the general run of our beliefs about the natural world: there may be good empirical evidence that our belief-forming procedures are reliable producers of truth, even if there is no conceptual guarantee of this. So the last chapter in effect added a realist theory of knowledge to the realist theory of content developed in chapter 3, by showing how, even though truth conceptually transcends evidence, they may still co-vary as a matter of empirical fact.

So realism, as I shall use the term henceforth, has two components. First, it requires a realist theory of content, according to which our beliefs answer to an independent world. And then it deals with the resulting sceptical threat by means of a realist theory of knowledge, according to which truth and judgement co-vary as matter of a posteriori fact, even if not of conceptual necessity.

So far, however, I have focused exclusively on "natural beliefs", in the sense of beliefs about the natural world. What about non-natural beliefs, such as beliefs about mathematics, or morality, or about modal questions of necessity and possibility? In chapter 3 I explicitly excluded such non-natural beliefs from the scope of the teleological theory of content, on the grounds that the teleological theory requires beliefs that are relevant to the success of actions, and it is debatable whether non-natural beliefs satisfy this requirement.

So it is unclear whether our realist theory of content, in the form of the teleological theory, applies to non-natural beliefs. Moreover, it is equally unclear whether a realist theory of knowledge has any grip on the non-natural realm: it seems odd, to say the least, to suppose that mathematical or other non-natural beliefs should derive their warrant from a posteriori reliable methods of belief-formation, from procedures whose reliability is testified by empirical evidence.

Non-natural beliefs thus seem to fall outside the scope both of our realist theory of content, and of our realist epistemology. Given this, it seems appropriate to explore a different route to the vindication of non-natural beliefs, and to see whether they can be vindicated in an anti-realist manner instead.

I am here using "anti-realist" in the sense made popular by Michael Dummett, to refer to analyses of content which imply that truth does not transcend evidence, that there is no gap between a judgement being properly arrived at and its being true. Even if such an anti-realist view of content is inappropriate to beliefs about the natural world, it may still be the right

account of mathematical and other non-natural beliefs. And this would then deal with any epistemological difficulties that may threaten such non-natural judgements: for if there is no conceptual gap between truth and evidence, then there is no need for any further explanation of why our practices for making such judgements should be thought to yield truths.[1]

6.2 Content and Knowledge

Before proceeding to the detailed consideration of mathematical and other non-natural judgements, it will be helpful to digress briefly and make some observations about anti-realist theories of content and knowledge in general, as applied to natural as well as non-natural beliefs. As I have just observed, anti-realist theories of content have the epistemological attraction of promising to dissolve sceptical threats: if your theory of content tells you that there is no conceptual room for properly arrived at judgements to be false, then you can stop worrying about the possibility of error. This is not of course an argument for anti-realism: that certain philosophical problems would disappear, if judgements about trees were really just judgements about sensations, is not in itself a good reason for thinking that trees are sensation s. Accordingly, anti-realist philosophers, from Berkeley on, have offered independent arguments for thinking that the content cannot outstrip evidence. Still, there seems no doubt that the epistemological implications have always operated as a strong motive for adopting anti-realist theories of content. (After all, the arguments for idealist, verificationist, and other anti-realist theories of content are scarcely conclusive; and the conclusions, certainly as applied to non-natural beliefs, are pre-theoretically highly implausible; it follows, I think, that something other than the arguments is needed to account for the widespread acceptance of anti-realist theories of content.)

On this question of motive, the epistemological attractions of anti-realism of content are all the greater if you aspire to certainty as a requirement for knowledge. For this strong traditional demand on knowledge adds weight to scepticism, and so to the desirability of a theory of content that promises to block scepticism at source. This link between certainty and anti-realism isn't inescapable: Descartes, for instance, managed to uphold certainty as a requirement on knowledge without abandoning a realist theory of content. He did, however, need God as a guarantor of certain knowledge. Without God, it is difficult to uphold certainty except by appeal to anti-realism. For the only plausible strategy for achieving certainty without God is the anti-realist tactic of collapsing the world into the mind by arguing that the contents of claims about the world don't extend beyond what introspection and logical analysis guarantees. (Note how this anti-realist move then aims to satisfy the demand f or certainty, the demand that we arrive at beliefs in ways that necessarily deliver truths, by arguing that the truth of our beliefs is conceptually guaranteed by the ways we arrive at them.)

Of course, anti-realism of content often has difficulty delivering on its anti-sceptical promises. In order for anti-realism of content to be all plausible, it needs to allow at least some distance between appearance and reality, needs to allow that there is something more to a judgement being true than that it is taken to be true on some specific occasion. So anti-realism will quickly move away from the claim that the presence of a table, say, is just a matter of your current sense impressions, and allow that it also depends on what impressions you will have in a moment, and on those that other people will have, and in the end on enough evidence to ensure that the relevant judgement won't be overturned by further discoveries. But of course the more anti-realism moves along this dimension, the less effective it becomes as a response to scepticism: it is one thing to be certain about what

sensations you are currently having; it is quite another to be certain that some current judgement will never be overturned by future discoveries.

But we can leave these problems to proponents of the anti-realist programme.[2] For it matters little whether or not they can fulfil their epistemological promises, given that we have good reason, as was shown in 3.13 above, for rejecting the anti-realist theory of content in the first place.

Moreover, once we abandon the demand for certainty, and therewith the requirement that our methods of thought should necessarily produce truths, there is then much less initial pressure for a theory of content which makes truth a conceptual upshot of our belief-forming procedures. The reliabilist alternative to the demand for certainty asks only for methods that are contingently reliable for truth. So from the reliabilist perspective there is no need to find a conceptual link between our doxastic practice and truth. It is quite enough if we can defend a realist theory of knowledge, a theory according to which the reliability of our methods is an empirical matter.

6.3 Anti-Realism versus Fictionalism for Mathematics

Let me now return to the issue of non-natural beliefs. As I said, it seems unlikely that we will be able to develop a realist theory of knowledge for the non-natural realm which will defend the reliability of our belief-forming procedures on a posteriori grounds. So perhaps here at least we should vindicate our beliefs by showing, in anti-realist spirit, that truth does not conceptually outstrip the basis on which we make such judgements.

In what follows I shall concentrate on mathematical judgements. It would be unreasonably ambitious to aim for detailed accounts of moral and modal judgements as well. But at the end of the chapter I shall return briefly to morality and modality, both for purposes of comparison, and to offer a few promissory thoughts.

My procedure, in connection with mathematics, will not be to aim for some general semantic theory, akin to the teleological theory I developed for natural judgements, which will explain "aboutness" for mathematical judgements. The welter of existing controversy which surrounds any philosophical discussion of mathematical judgements effectively precludes any such direct approach. Instead I shall proceed indirectly, by asking about the epistemological consequences of mathematical meaning, rather than about mathematical meaning itself. In particular, I shall ask directly whether an anti-realist epistemology is defensible for mathematics: is truth, for mathematical judgements, nothing more than evidence, nothing more than being warranted by proper mathematical procedures?

In due course I shall conclude that this anti-realist view of mathematics is unacceptable. This will implicitly establish that mathematical judgements have a realist semantics, in the sense that truth, for mathematical claims, conceptually transcends the basis on which we make such claims.

This then threatens scepticism. I have argued that, in the case of claims about the natural world, the corresponding sceptical threat is blocked because our judgemental practices are reliable for truth as a matter of a posteriori fact, even if not by conceptual necessity. I shall briefly consider whether any analogous strategy will work for mathematics. But this line of thought will come to nothing.

So we will be left with a sceptical -- or, more familiarly, "fictionalist" -- attitude to mathematics. Hartry Field (1980, 1989) has done much to explain how such a position can work. A detailed explanation is best left till later. But in outline the fictionalist attitude will combine:

(a) a literal understanding of mathematical claims, as referring to abstract objects like numbers, sets, and so on, with

(b) a rejection of belief in such claims, and

(c) an acceptance of such claims as fictions which are useful for various pragmatic purposes.

The resulting position is closely analogous to the instrumentalist attitude to scientific theories adopted by Bas van Fraassen (1980). Van Fraassen's sceptical instrumentalism avoids the contortions of earlier anti-realist brands of scientific instrumentalism, in that he takes scientific theories at face value, as literally referring to unobservables like atoms and electrons, and abandons any attempt to re construe scientific theories as merely making claims about observables. He then combines this literal understanding of scientific claims with a refusal to uphold those claims as true. Van Fraassen's view is that we shouldn't believe scientific claims about unobservables, but should simply "accept" them as useful instruments for making predictions, summarizing data, and so on.

I don't agree with Van Fraassen about scientific theories about unobservables, as the arguments at the end of the last chapter will have made clear. But I do think the analogous position is right for mathematics: we should understand mathematical claims at face value, but should only accept them as useful instruments, not believe them.

6.4 If-Thenism

I shall start, as I said, by asking whether an anti-realist theory of mathematical knowledge is defensible. Is there an intrinsic link between evidence and truth for mathematical claims? At first sight such a view may seem highly plausible. After all, it is a familiar thought that, where judgements about the natural world answer to independent facts, there isn't anything more to mathematical truth but provability: that what makes it correct to say that there is no greatest prime number, say, or that the real numbers are non-denumerable, is not that there is some independent world in which these facts obtain, but simply that these claims can be established by recognized methods of proof. (My initial concern in this chapter is with pure mathematics, like arithmetic and analysis; the application of mathematics to the empirical world will be discussed in due course.)

However, familiar as it is, this anti-realist view of mathematics faces difficulties. Let me focus on the notion of proof. What exactly are the "recognized methods of proof" in any given mathematical subject area? An initial answer might be that in any such area mathematician s start with certain basic assumptions (which might or might not be formally recognized as "axioms") about the natural numbers, or the reals, or groups, or non-Euclidean spaces, or whatever, and then use logic to derive further conclusions as theorems from those basic assumptions. Let us grant, for the time being, the appropriateness of logic for this purpose. (I shall return to the epistemological status of logic at the end of this chapter.) This leaves us with the axioms. And here there is an obvious difficulty, namely, that the axioms haven't themselves been proved. Rather they are the point at which mathematicians start

proofs. So, on the face of it, it seems that mathematical proofs only establish that if certain assumptions are true, then certain other claims, the theorems, are also true.

There is a view in the philosophy of mathematics according to which mathematical assertions should be understood as expressing precisely such hypothetical claims. This view is called "if-thenism". Now, if "if-thenism" were true, then the existence of a mathematical proof would indeed conceptually guarantee the truth of the corresponding mathematical claim, and mathematical anti-realism would be vindicated. However, it seems clear that "if-thenism" is simply wrong about what mathematical statements actually mean (cf Resnik, 1980, ch 3). Number theorists don't just hold that if there is a number 0, and if every number has a successor, and . . . so on for the rest of Peano's postulates, then there is no greatest prime number. On the contrary, they hold that 0 does exist, and that every number does have a successor, . . . and consequently that there definitely isn't a greatest prime number.

We could of course understand "if-thenism" not as an account of what mathematical statements do mean, but rather as an account of what they should mean. On this interpretation, "if-thenism" would be recommending that we should revise our understanding of mathematical statements, precisely so as to ensure that our methods of mathematical proof suffice to establish those statements. But this then makes my point clear: namely, that, as currently meant, mathematical statements lay claim to more than mathematical proofs establish, thus undermining the anti-realist equation of mathematical truth with mathematical proof.[3]

6.5 Postulationism

The difficulty raised for mathematical anti-realism in the last section was in effect that mathematical practice seems just to assume the axioms from which it starts proofs, and does nothing to establish those axioms. But perhaps anti-realists could respond that the peculiarity of mathematics is precisely that its basic assumptions don't need any further proof, on the grounds that the requisite mathematical objects will automatically be available to satisfy any consistent set of mathematical assumptions.

I shall call this attitude towards mathematics "postulationism", since it implies that no further justification is needed for a mathematical theory than the consistency of its postulates. At first sight postulationism might seem to make mathematical existence unacceptably dependent on human activity, with mathematical objects somehow springing into existence as mathematicians formulate assumptions. But we needn't understand the postulationist theory in such a strongly anthropocentric way. Rather the idea could be that there is a timeless Platonist realm in which there are abstract objects satisfying any possible set of consistent mathematical axioms, whether or not anybody has yet thought of those axioms.[4]

This would of course mean that there are an awful lot of mathematical objects -- as well the familiar objects of standard mathematics, there will also be such non-standard objects as all the different kinds of numbers modulo n, and all the shapes in all possible geometries, and all the operators in all possible vector spaces, and indeed all kinds of things that have never been thought of and never will be. But perhaps there's nothing wrong with that. Large universes are scarcely alien to mathematics.

A more substantial objection to postulationism might be that mathematicians themselves make a distinction between those branches of mathematics whose existence claims they take seriously and those whose they don't. The complaint here would in a sense be the mirror

image of the claim levelled against "if-thenism". Where "if-thenism" says that all mathematics is meant hypothetically, "postulationism" seems to imply that all mathematics can be asserted unconditionally. But this then means that "postulationism" is open to a mirror image of the objection made to "if-thenism": since there are branches of mathematics in which mathematicians do restrict themselves to hypothetical attitudes, considering the axioms as assumption s whose consequences are worth exploring, rather than as claims to be believed, it is wrong to read all mathematics as unconditionally assertible.

However, I think that postulationism has a reasonable answer to this complaint. For there is a natural way for postulationism to distinguish between those branches of mathematics which it appropriate to understand hypothetically and those which it appropriate to understand unconditionally, a way which seems to line up accurately with the way practising mathematicians make this distinction. Postulationists can appeal to the distinction between sets of axioms which are categorical, in the technical sense of determining a unique model, up to isomorphism, and those which are not so categorical. The axioms of group theory, for instance, are not categorical, in that quite disparate sets of objects, of different cardinalities, can form groups. Peano's postulates, by contrast, are categorical (in second-order logic), in that any set of objects and relations which satisfy them can be placed in a structure-preserving one-to-one correspondence. So the natural move for postulationism is to argue that mathematical objects are available to satisfy every consistent and categorical set of axioms; non-categorical axiom sets, by contrast, do not guarantee the existence of any mathematical objects, and so should be read hypothetically, as saying merely that if there are any objects which . . ., then . . .

This suggestion accords well with actual mathematical practice. Mathematicians certainly seem to be committed, as I have already observed, to the numbers 0, 1, and all their successors, as existing objects. By contrast, it makes little mathematical sense to talk about the identity element mentioned in the axioms of group theory. Yet, even within group theory, once we add enough special assumptions to the general axioms of group theory to give us categoricity, then, in line with the current suggestion, we do find mathematicians talking unconditionally about the simple group of order 68, the elliptic modular group, and so on, as if these special groups, at least, were as real as the number one.

So postulationism can answer the charge that it makes all mathematics unconditionally assertible. It simply restrict its ontological commitment to those abstract objects required to satisfy categorical mathematical theories. This then implies, in accord with existing ma thematical practice, that only such categorical theories should be unconditionally asserted, and that non-categorical theories should merely be embraced hypothetically.

There is, however, another rather more telling objection to postulationism. According to postulationism, as now understood, all objects that can consistently and categorically be postulated thereby exist. But how then does postulationism differ from a fictionalist attitude to mathematical objects? If everything that can consistently and categorically be thought to exist thereby does exist, then won't Sherlock Holmes exist, and Santa Claus, and the Wicked Witch of the North?[5] Perhaps it is natural to slip into unreflective acceptance of categorical truths about numbers and sets and simple groups. But if all this really amounts to is accepting what follows from assumptions agreed by mathematicians, why is it any different from accepting what follows from our agreed assumptions about Santa Claus?

The postulationist might object that the comparison is not fair. If we take claims about Sherlock Holmes and Santa Claus literally, then these claims are about people who inhabit the same spatiotemporal world as we do. And on this literal reading these claims can be shown to be false. Maybe the stories are internally consistent, but they aren't consistent with the totality of our beliefs about the world. Which is why we don't in fact accept these claims literally, and why we don't accord Shylock and Santa Claus real existence, but only fictional "existence", that is, non-existence. By contrast, claims about mathematical objects aren't about spatiotemporal objects. So nothing forces us to regard them as literally false, nor to regard the objects they mention as mere fictions.

But I don't think that this is good enough. It still seems to me that the postulationist story as told so far gives us no reason to view mathematical existence as anything more than fictional non-existence. Maybe the definite reasons which force us to adopt fictional attitudes to explicit fictions don't carry over to the mathematical case. But, even so, the postulationist hasn't told us anything more about what's involved in mathematical existence, other than that a consistent and categorical story can be told about the objects in question. Such a story guarantees fictional "existence". But if mathematical existence is to amount to more than the non-existence of fictional "existence", then there must be something more at issue than an internally consistent story, for abstract objects just as for concrete ones.

6.6 Reductionism

The last two sections have presented mathematical practice as an entirely "internal" business, in which assumptions are accepted and consequences drawn therefrom. As long as we stick with this internalist picture, it will be difficult to avoid fictionalism, for lack of any account of what makes the acceptance of basic assumptions anything but arbitrary. But should we accept this internalist picture in the first place? After all, isn't there an essential relationship, at least for the central branches of mathematics like set theory and arithmetic, between abstract mathematics and activities like classifying and counting ordinary non-abstract objects?

Consider these two sentences:

(1) John is tall

(2) John is a member of the set of tall people.

Again, consider these two:

(3) (Ex)(Ey)(Rx & Ry & x*y & (z)(Rz -> z=x v z=y))

(4) The number of rhinoceroses in England = 2,

where "R" abbreviates "is a rhinoceros in England".

Clearly there is an intimate relationship between the first and second members of these pairs. On the surface at least, the second sentence in each case mentions an abstract object -- a set, a number -- whereas the first member is free of any such reference. But, despite these surface differences, there is no doubt in each case that any non-philosopher who understands both sentences and accepts one will automatically accept the other.

The relationships illustrated by these pairs hold out the promise of grounding our knowledge of abstract mathematical objects. Given the close affinity between the two sentences in each pair, it is hard to see how there can be epistemological problems about the latter platonist claims, given that there obviously aren't any about the former nominalist ones.

However, to develop this idea we need a more precise analysis of the relationships illustrated by the above pairs. In this section and the next I shall consider two possible such analyses. I shall argue that both analyses run into difficulties, and that in both cases we are in the end forced back to fictionalism.

The first analysis -- let's call it the "reductionist" analysis -- is that the quantificational sentence (3) gives the real meaning of the apparently platonist sentence (4). (It will be convenient to focus on the second, arithmetic example, as it brings out the issues more clearly.) On this view, there isn't any real reference to the abstract object 2 in sentence (4) in the first place. (4) is just a stylistic variant of (3), and no more commits us to abstract numbers that talk of doing things "for the sake of such-and-such" commits us to sakes.

If we restrict our attention to pairs like (3) and (4), then this argument has a high degree of plausibility. That there's one rhinoceros in England, and another one, and no more, indeed seems to be just what is meant by saying that "the number of rhinoceroses in England equals two". So in this kind of case the apparent reference to a numerical object, two, can quite happily be viewed as a mere figure of speech. (4) commits us to rhinoceroses, but not numbers. What is more, if (4) is equivalent to (3), it is easy to see how we could establish that it was true, by counting or some equivalent procedure.

So far this deals with statements that attach a number to a non-numerical concept like " rhinoceros in England". But what about statements of pure arithmetic, like

(5) 2 + 3 = 5?

Here too the reductionist has a plausible line. To simplify our notation, let us abbreviate (3) above as (E2x)(Rx), and understand further numerical quantifiers (Enx) analogously. Then the natural reductionist move is to read (5) as really saying

(6) (V)(W)[(E2x)(Vx) & (E3x)(Wx) & -(Ex)(Vx & Wx) -> (E5x)(Vx v Wx)].

That is, we can read (5) as saying merely that if there are two Vs and three different Ws, then there will be five things which are V or W. The numerals here just indicate the kind of quantification involved, and don't refer to numbers.

It is by no mean s implausible that (6) gives the real content of (5). What is more, (6) is a logical truth, and so, assuming still that we can give a satisfactory account of logical knowledge, the rendering of (5) as (6) accounts for our ability to know such truth of simple arithmetic as 2 + 3 = 5.[6]

This then offers the model for a reductionist account of arithmetic in general. First of all reductionists parse away apparent references to numbers as abstract objects in favour of quantificational constructions. And then they aim to show that the truths of arithmetic reduce to truths of logic.[7]

However, this programme runs into difficulties when we come to more complicated arithmetical statements, such as "there is no greatest prime number". There is, it is true, a quantificational version of even this statement, which once more is free of any commitment to numbers as such. But now the quantificational version is extremely complex, involving not just second-order, but third-, fourth- and fifth-order quantifiers, and it becomes much harder to see in exactly what sense the quantificational version is equivalent to the original number-theoretic claim.

Perhaps the reductionist need not be unduly worried by such complexity. Why shouldn't the surface structure of arithmetical statements conceal hidden logical articulation? But there are further problems. Consider again the simple arithmetic truth 2 + 3 = 5. A couple of paragraphs ago I allowed that this was plausibly equivalent to a second-order logical truth.

But suppose that what we're counting is not rhinoceroses, say, but numbers themselves, as in "If there are two numbers which are F, and another three numbers which are G, then there are five numbers which are F or G". In line with the reductionist programme, this will come out as a fourth-order logical truth. And similarly there will be yet higher-order "versions" of 2 + 3 = 5. The reductionist seems to be forced to say that "2 + 3 = 5" is ambiguous, hiding a number of distinct "real" contents behind its surface structure. But this is surely unacceptable. It's one thing to say that surface structure is misleading as to the hidden content of arithmetical statements. It's another to maintain that straightforward arithmetical statements don't have an unequivocal real content at all.[8]

To leave the example of arithmetic for a moment, it is worth pointing out that similar problems of ambiguity will arise if we attempt to apply the reductionist programme to mathematics in general. The natural strategy here would be (a) to reduce other branches of mathematics to set theory, (b) appeal to the affinity illustrated by (1) and (2) above to argue that the apparent reference to sets conceals the real quantificational content of the reduced mathematical theories, and (c) to aim to show that the reduced theories all are logical truths.

Two problems of ambiguity face this programme. To start with, there is the point that the simple notion of set will correspond to different types at different levels of logic, analogously to the above way in which numbers come out differently at different logical levels. And there is also an additional difficulty, because of the familiar point that there are in general many alternative ways of reducing branches of mathematics like real analysis, say, to set theory, all of which preserve the relevant logical structure, but which give different set theoretical surrogates for given statements of analysis. For both these reasons the thought that logical reduction gives the "real" content seems to lead to the unattractive conclusion that straightforward mathematical claims conceal hidden ambiguities.

In the face of such problems, defenders of the reductionist programme tend to shift position, and allow that in the end that mathematical statements, as meant by mathematicians, do after all essay reference to simple mathematical objects like numbers, and that because of this such statements are both free of ambiguity and psychologically manageable.[9] They thereby limit their reductionism to the moderate position that the legitimacy of mathematical statements derives from the availability of logically true quantificational surrogates which don't refer to abstract objects.

However, this move takes away the distinctive claims of reductionism. You can't have it both ways. Either mathematical statements really do mean the same as their quantificational

surrogates, or they do not. If they do, you are stuck with the problems of ambiguity mentioned above. If they do not, then the fact that we should believe the quantificational surrogates doesn't establish that we should believe the mathematical statements.

Of course there is still room to argue that the affinities between mathematical and quantificational statements show why it is harmless, and useful, to accept mathematical claims. But this is different from showing that it is right to believe those claims. Indeed this position is not significantly different from fictionalism. Th e reductionist is now arguing that it is legitimate to "accept" mathematical claims because they can in principle always be replaced by logically true quantificational surrogates. But, as we shall see, fictionalists hold a very similar view, though for somewhat different reasons, in that they hold that our "acceptance" of mathematical claims about abstract objects is all right because in principle mathematics doesn't allow us to do anything that we couldn't do by logic alone.[10]

The fictionalist, however, goes on to insist that since these mathematical claims, which commit us to abstract objects, are not equivalent to logical claims, which do not, and since we have no epistemological warrant for this extra commitment to abstract mathematical objects, we ought to stick to the "acceptance" of mathematical claims, and eschew belief. Similarly, once somebody of reductionist sympathies admits that mathematical claims do refer to abstract objects, and so are not equivalent to quantificational surrogates, however significant those surrogates may be for understanding why references to abstract objects are useful, then the reductionist has ceased to offer an argument for believing mathematics.[11]

6.7 Neo-Fregeanism

I have just argued that, once we allow that mathematical claims commit us to abstract objects, then we cannot continue to view them as equivalent to quantificational claims, on the grounds that the latter do not commit us to abstract objects. Crispin Wright, in Frege's Conception of Numbers as Objects (1983), disagrees, for arithmetic at least, on the grounds that quantificational claims do commit us to abstract objects like numbers.

Let us return to the equivalence:

(3) (Ex)(Ey)(Rx & Ry & x*y & (z)(Rz -> z=x v z=y))

(4) The number of rhinoceroses in England = 2.

Wright agrees with the reductionist that (3) and (4) mean the same. But he thinks that (3) gives the real meaning of (4), rather than the other way around. That is, he thinks that the surface form of the quantificational (3) is misleading, and that we ought to recognize that underneath its surface it commits us not just to rhinoceroses, but to the number two as well.

Since he holds that arithmetic does commit us to abstract objects, Wright needs a non-reductionist epistemology for arithmetic. To this end, he introduces an equivalence between the following two schemas (which is in effect a generalization of the equivalence of (3) and (4)):

(7) The Fs can be put into a one-to-one correspondence with the Gs.

(8) The number of Fs = the number of Gs.

Wright calls this equivalence "N=", and he proceeds to show that it implies all of Peano's postulates, and hence all of arithmetic, in the context of second-order logic.[12] The resulting system, which is closely modelled on Frege's Grundlagen, treats the numbers themselves as objects in the range of first-order variables. It uses second-order quantification, but, unlike the reductionist programme, nothing higher. However, where the reductionist programme promises to account for all arithmetical knowledge as purely logical knowledge, Wright needs to add N= to logic, since there is no question of justifying statements which commit us to numbers as objects by pure logic alone.

The success of Wright's programme thus hinges crucially on the status of N= itself. Wright takes this to be a conceptual truth, despite the fact that (8) refers explicitly to numbers but (7) does not. His line here is the same here as with (3) and (4). He thinks that the reference to numbers a s objects in (8) is indeed to be taken at face value. But he doesn't think that this undermines the conceptual equivalence of (8) with (7), because he thinks that (7) itself commits us to numbers as objects, even if its surface structure conceals this fact.

In support of Wright's view that the reference to numbers in (8) should be taken at face value, we can observe that the numerical expressions appearing in (8) certainly seem to function like genuine singular terms in these, and other, contexts. They can flank identities, they allow existential generalization, and so on. This creates a strong prima facie case for reading these terms referentially,[13] and provides a serious challenge to anybody who wants to de fend a non-referential interpretation (a challenge which, as the last section showed, the reductionist, for one, is unable to meet).

Yet, once we accept this referential reading of (8), then, given the conceptually equivalence of (8) with (7), it immediately follows that somebody who asserts (7) is already committed to numerical objects, even if it doesn't look like it.

Or so at least Wright argues. The difficulty with this line, however, is that by urging the genuineness of the numerical singular terms in (8), Wright thereby undermines the analytic equivalence between (7) and (8).

Recall that the schemas at issue are:

(7) The Fs can be put into a one-to-one correspondence with the Gs.

(8) The number of Fs = the number of Gs.

I am entirely happy to agree with Wright that instances of (8) are genuine identity statements which commit us to numbers as objects. However, this claim surely takes away any original reason we had for accepting the analytic equivalence of (7) and (8). For on the face of it, where (8) commits us to numbers as objects, (7) does not.

In Frege's Conception of Numbers as Objects, Wright does not really ad dress this objection. This is because he takes his main opponent to be the reductionist, and accordingly takes N= to be agreed as an analytic truth on all sides. Wright's concern is then to merely to show that, once it is agreed that N= is analytically true, his reading of N= is superior to the reductionist reading. I agree that his reading is superior to the reductionist reading, if we assume (7) and

(8) are analytically equivalent. But my point is that, once we move to Wright 's reading, then we ought to question whether (7) and (8) are equivalent, as asserted by N=, in the first place.

After all, the most natural way to read (8) is as increasing our ontological commitments, beyond what is required by (7). If we adopt this reading, then we will agree with Wright that (8) involves genuine commitment to numbers as objects. But we will deny, precisely for this reason, that (7) is analytically equivalent to (8). After all, it certainly doesn't lo ok as if (7) requires us to believe in numbers as well as everyday objects. We can certainly imagine a community, for instance, who understood statements like (7), but who had no notion of a numerical object.

Wright would object that the possibility of such a community isn't conclusive: the crucial issue is whether, once the community has acquired the notion of a numerical object, it is then in a position to recognize that N= is analytically guaranteed. However, we can agree that this is the crucial issue, yet still insist that the onus is on a Fregean like Wright to produce some argument for the analytic equivalence of (7) and (8). For, as before, at first sight the relevant statements certainly seem to differ markedly in ontological commitment.

Wright holds, plausibly enough, that N= will play a central part in any adequate introduction to the concept of number. But this doesn't suffice to make N= an analytic truth. Consider an analogy. Some such thought as that electrons are negatively charged objects orbiting the nuclei of atoms is no doubt essential to any adequate introduction to the concept of an electron. But that doesn't make it an analytic truth that, if there are atoms, then there are electrons. For a commitment to electrons is an extra ontological commitment, over and above any commitment to atoms, as is shown by the example of late nineteenth-century chemists, who believed in atoms, but not in electron s. What is analytically true is this: if there are any small, negatively charged entities orbiting the nuclei of atoms, then those objects are electrons. But this is not enough to derive, from claims about atoms, claims about electrons. For that we need extra evidence that there actually are small, negatively charged entities orbiting the nuclei of atoms.

Yet this is what Wright seems to think we can do for numbers. Let us grant Wright that any adequate introduction to the concept of number will contain the information that the same number attaches to equinumerous concepts. Still, it doesn't follow that N= is an analytic truth. For, just as with the example of electrons and atoms, numbers may involve an extra ontological commitment, over and above that required by equinumerous concepts. What is certainly analytically true is that, if there are any numbers, then the same number will attach to equinumerous concepts. But this in itself doesn't suffice to take us from premises about equinumerous concepts to conclusions about numbers. In order to make that move, we need some independent argument for supposing that numbers actually exist.

6.8 Mathematical Anti-Realism Rejected

In the absence of any such argument, I think we ought to reject Wright's neo-Freagean account of arithmetic. And, more generally, I think that we ought also now to reject the overall anti-realist approach to mathematics. The initial objection to this anti-realist approach, made in section 6.4, was that mathematical evidence, in the form of proofs, only seems to establish hypothetical claims, while mathematics itself consists of unconditional assertions.

In answer to this objection, I offered the anti-realist various ways of discharging the axioms assumed in mathematical proofs: first, we considered if-thenism, which read all mathematical

claims hypothetically; then I discussed postulationism, which argued that, for any consistent and categorical set of assumptions, the abstract objects exist which make them true; after that came reductionism, which claimed that mathematical assumptions could all be construed as truths of pure logic; and finally there was Wright's neo-Fregeanism, which claimed that the crucial assumptions necessary to introduce abstract objects are simply analytic truths. None of these strategies has proved defensible, and it is difficult to think of any other a priori argument for the view that the assumptions standardly made by mathematicians about abstract objects are automatically true. It seems to me that is now time to conclude that no such anti-realist defence of mathematics is available, and to accept that the content of mathematical claims does indeed conceptually outstrip the grounds on which mathematicians make them.

I realise that some readers will find this difficult to stomach. Surely, they will say, the truth conditions of mathematical judgements cannot possibly transcend our grounds for asserting those judgements. Are not the contents of such judgements fixed by the grounds which our practice authorizes as sufficient for their assertion? So what could possibly make it the case that a mathematical statement stands for something that goes beyond such grounds?

But my thesis is precisely that the contents of mathematical judgements are not fixed by the grounds our practice recognizes as sufficient. I am denying that such an anti-realist model of meaning is acceptable for mathematics. It may be helpful, in this connection, to consider the fate of anti-realist thinking in the analogous context of the interpretation of scientific theories. In the first half of this century many philosophers were attracted to the view that theoretical terms in science were a disguised shorthand for describing complexes of observational circumstances. This was, of course, an absurdly counter-intuitive view. It is scarcely credible to suppose that scientists who talk about "electrons" are in fact talking about the behaviour of oil drops, tracks in cloud chambers, and so on, and not about the small negatively charged objects which orbit the nuclei of atoms. But philosophers had difficulty seeing how scientists could be talking about small negatively charged objects. Since the authorized grounds for applying the term "electron" are always observable circumstances, what could possibly justify us in interpreting the scientists as making some further insecure reference to invisible entities?

Frank Ramsey (1931) explained how scientists manage to refer to unobservables. "Electron" does not have its meaning fixed just by association with the observable symptoms of electron behaviour. It also gets its meaning from its role in a theory which postulates the existence of small particles which orbit atomic nuclei and are responsible for those observable symptoms. Ramsey showed how statements about electrons can be read as existentially quantified statements, which say that there exist particles which are small, negatively charged, orbit atomic nuclei, have certain observable symptoms, . . .

In effect, Ramsey shows that talk about scientific unobservables derives from our ability to make existential claims about object which are not immediately accessible. I suggest that this same ability makes it possible for mathematical cl aims to answer to proof-transcendent states of affairs. In the case of arithmetic, say, we have a theory which postulates the existence of objects with certain properties, namely, just those properties which flow by logic from N=. We call these putative objects numbers. But the basis of our ability to make claims about numbers, namely, our power of existential generalization, is independent of any further abilities we may have to prove such claims.

It must be allowed, of course, that we have an established discursive practice of making arithmetical claims, and that a central role in this practice is played by N=. But the existence of this practice does not justify N=, nor the arithmetical claims which follow from it. For, as we have seen, N= is not analytic, but a synthetic claim, which inflates our ontology by postulating entities we are not otherwise committed to. As such it cannot be justified just on the grounds that it is part of an established dis cursive practice. Analytic truths are justified by facts about linguistic usage. But synthetic claims require some other warrant.

One last throw is available to mathematical anti-realists. They can deny that "existence", in the context of mathematical discourse, is to be understood in the same way as in other areas of discourse, and thereby hope to argue, for example, that the "existence" of numbers is analytically guaranteed by facts of equinumerosity. But this move is not only unattractively ad hoc -- since there is no other reason, apart from the threat of scepticism, for suspecting mathematical existence claims of equivocation -- but it is also likely to prove a two-edged sword -- since the anti-realist will still have to explain what "existence" means for mathematical objects, and why it is different from fictional non-existence.

The argument of this section has in effect shown that mathematical discourse falls within the scope of the teleological theory of content after all. For I have now argued that mathematical terminology can be introduced, a la Ramsey, by existential quantification into theoretical contexts. This means that mathematical discourse rests on no special vocabular y, but simply on the existential quantifier we use in general discourse. A corollary is that the semantic realism of mathematical discourse is just a special case of the general semantic realism which emerges from the teleological theory.

6.9 The Putnam-Quine Defence of Mathematics

If mathematical claims have a realist semantics, they face a threat of scepticism. Still, perhaps this threat can be met. After all, scientific theories are able to cope with the threat of scepticism. Perhaps mathematics can be defended against scepticism in the same way as scientific theories are.

This might seem a faint hope. However, there is a well-known line of argument, propounded by Hilary Putnam (1971), and originally due to Quine, which seeks to vindicate claims about abstract mathematical objects by arguing that such claims play an ineliminable role in scientific theories about the natural world. So far I have been taking mathematics to consist of pure mathematics, such as arithmetic and analysis. But references to abstract mathematical objects, and in particular to the natural and real numbers, are also regularly made in the applied sciences, as when we say, "the number of planets = 9" , or "the distance-in-metres between two particles = 5.77". The Quine-Putnam point is that if scientific claims of the latter kind are epistemologically warranted, as surely they often are, and if those claims commit us to real n umbers as objects, as they certainly seem to, then claims about real numbers must be epistemologically warranted too.

An extension of this line of argument promises to vindicate, not just mixed statements of applied mathematics, as in the above examples, but also the axioms on which pure arithmetic and analysis are based, such as that every number has a successor, or that every set of reals has a least upper bound. For these assumptions are presupposed in the mathematical calculations which we use to derive predictions from scientific theories, as when we add together the numbers of stars in different galaxies, say, or divide forces by masses, and so can

be argued to be confirmed, along with the rest of such theories, when such predictions prove successful.

From the point of view of my general epistemological framework, this argument amounts to a realist defence of mathematics. Quine and Putnam would no doubt not think of it in this way, given their generally pragmatist attitudes to scientific truth. But, on my account, scientific theories have realist contents, yet qualify as knowledge because the methods by which we choose them are reliable for truth as a matter of empirical fact. So if mathematic al theories qualify as knowledge as part and parcel of scientific theories, then they will share this realist epistemological status with scientific theories.

One difficulty facing the Quine-Putnam argument is that it seems unable to account for the difference between the research methods of pure mathematicians and natural scientists. Where scientists actively seek to vindicate their theories by experimental means, pure mathematicians seek a priori proofs. The idea of a pure mathematician trying to establish some mathematical principle by experiment seems silly. Yet the Quine-Putnam argument seems to imply that pure mathematics and natural science share the same epistemological status.

This objection, however, is less conclusive than it looks. It is certainly true that the Quine-Putnam argument implies that basic mathematical principles are supported by observations. But this doesn't imply that there will be some specific experiment that bear s on each such principle. Rather, as with the basic laws of motion, large numbers of observations will contribute to the support of mathematical principles in a holistic manner, by confirming the overall theory in which they play a part. As to the r ole of a priori proof in mathematics, we can accept that basic mathematical principles depend on observational support, without denying the importance of exploring the purely logical consequences of those principles. In line with this, we might view the overall scientific enterprise as containing a division of labour: the scientists conduct experiments which will shape the overall tree of scientific theory; while the mathematicians explore the purely logical consequences of the assumptions that lie at the root.

Perhaps there remains something counterintuitive in the idea that the axioms of Peano arithmetic have the same epistemological status as Newton's laws of motion. I shall not pursue this issue any further, however, since there is a rather more telling objection to the Quine-Putnam argument, elaborated in Hartry Field's Science without Numbers (1980).

6.10 Field's Fictionalism

Field argues that the crucial premise of the Quine-Putnam argument, that pure mathematics is an inextricable part of natural science, is unwarranted. There are two parts to Field's claim here. First, he argues that we can say everything we want to say about the natural world in "nominalist" terms, that is, without mentioning abstract objects. When it comes to arithmetic, for example, his claim is that we can always describe the natural world using quantificational statements like "(E2x)(Fx)" or "F and G are equinumerous", and can thus avoid any commitment to numbers as objects. And in the case of geometry, to take another example, he argues that talk which commits us to distances as real numbers can always be dispensed with in favour of claims about relations of congruence between different spatial intervals. Field argues that similar procedures will allow all claims about the natural world to be understood as free of commitment to abstract objects.

Second, Field argues that whenever we use abstract mathematics to facilitate inferences between such "nominalist claims" -- and Field admits that abstract mathematics often enables us to find a simple route through inferences that would otherwise be impossibly complex -- we could in principle always make the inferential step by logic alone. As Field puts it, mathematics is conservative with respect to inferences from nominalist premises.

In line with these arguments, Field concludes that there is no good argument f or believing the claims of abstract mathematics, and that we should therefore reject these claims. This doesn't mean we should simply away throw all mathematical claims as complete rubbish. As I have just observed, Field accepts that mathematics is often immensely convenient for making inferences, and accordingly recommends that we should adopt the fictionalist view that mathematics is a useful pretence. But the point remains that this kind of usefulness provides no basis for belief, but only for the kind of attitude that we have towards the statements in a fictional narrative, or -- perhaps a better analogy -- the kind of attitude that we have towards the Coriolis force, or towards the Newtonian theory of gravitational forces in Euclidean space. We don't believe in these theories or the entities they postulate, but we know they will work satisfactorily enough when certain purposes are at hand, and then we find it convenient temporarily to pretend that they are true.

In a sense Field's motivations are similar to those of the reductionist discussed earlier -- he too wants to do without abstract objects. But instead of arguing, implausibly, that our existing mathematical discourse is free of commitment to abstract objects, his strategy is instead to admit that mathematics does commit us to abstract objects, but show how we could in principle manage without mathematics.

There are a number of technical difficulties facing Field's programme, both in respect of the first claim, that natural science can be "nominalized", and especially in respect of the second claim, that, within such a nominalized science, mathematics won't ever underpin any inferences that logic can't (cf Malament, 1982; Shapiro, 1983; Chihara, 1990). But for the most part I propose to skip these technicalities here (though I shall touch on some of them in section 6.15 below). After all, there are strong prima facie reasons for expecting them to be surmountable: it would surely be surprising if descriptions of the natural world of space and time required essential reference to abstract objects outside space and time; and it would be almost as surprising if logically possible combinations of natural facts were incompatible with standard assumptions about abstract mathematical objects, as would be the case if mathematics were not conservative with respect to nominalist premises.

Suppose then that we agree that abstract mathematics can be successfully extricated from the rest of science, in the way Field has in mind. It is worth considering a bit more carefully exactly why the rejection of mathematical beliefs is supposed to follow. After all, what Field has shown is that we can do without mathematical beliefs, in the sense that our scientific beliefs do not require mathematical beliefs. But this is scarcely the same as showing that we ought to do without mathematical beliefs. Why shouldn't we still retain mathematical beliefs, in addition to scientific beliefs?

Consider the analogous issue as to whether we ought to have beliefs in scientific unobservables, in addition to beliefs about observables. Craig's theorem shows that such unobservable claims are extricable from the observational claims, analogously to the way that Field shows mathematical claims are extricable from "nominalist" claims. In this sense, Craig's theorem shows how we can do without belie fs about scientific unobservables.[14]

However, few philosophers nowadays would want to infer from Craig's theorem that we ought to do without such beliefs, that our beliefs about scientific unobservables are all unwarranted.

To conclude that we ought not to adopt a given belief, just because we don't have to, is to betray an overdeveloped taste for desert landscapes. A better principle would be that we ought to adopt as many beliefs as we can, on matters that are of interest to us, as long as these beliefs can be reached by reliable methods, and so can be expected to be true.

In line with this principle, and following the points made at the end of the last chapter, I would offer the following explanation of why we are entitled to believe theories about scientific unobservables: the procedures that lead us to believe such theories are reliable routes to the truth, since nature generally prefers the kind of simplicity recognized by scientists to the kind of complexity that would obtain if reality were exhausted by the observable phenomena.

However, if this is our reason for upholding scientific theories, then why can't we defend mathematical theories in the same way? W hy shouldn't we argue that scientific theories which incorporate mathematics display a greater degree of basic simplicity than nominalized theories which don't incorporate mathematics?

This wouldn't be the same as the Quine-Putnam argument that mathematics is inextricable from scientific theories. The idea would rather be that the addition of mathematics to scientific theories is justified in the same way as the addition of unobservable claims to observable claims. This is a different kind of realist defence of mathematical theories. Instead of arguing that the leap to mathematical knowledge is part of the leap to scientific knowledge, we are now accepting that mathematical knowledge requires an extra leap, but suggesting that it might be achieved by the same technique as the leap to scientific knowledge.

However, if we examine this idea more closely, it doesn't really work. We can all agree, fictionalists included, that the incorporation of pure mathematics into scientific theories adds to the ease with which they are manipulated by human beings. But this is not the kind of simplicity at issue. What guides us in our choice of scientific theories is not computational simplicity but physical simplicity, the kind of simplicity manifested by the atomic theory of matter, or the relativistic theory of gravitation. These theories show that the world contains fewer independent phenomena than we might initially suppose, and as such accord with the general pattern of physical simplicity which allows us to gain knowledge of the natural world. But the incorporation of pure mathematics into scientific theories does not add to this kind of simplicity, but subtracts from it: it requires that we should recognize, in addition to the nominalized world of distances, forces and other interrelated natural quantities, a world containing real numbers, and sets, and other purely mathematical objects. This might make it easier to do calculations, but it receives no backing from principles of scientific theory choice.

Field (1980) makes this point in terms of a contrast between "intrinsic" and "extrinsic" scientific explanations. Mathematized scientific theories commit us to "extrinsic" explanations, to the kinds of explanation which explain a body's acceleration, say, as depending inter alia on the body's connection (via a mass-in-some-units relation) with a real number outside space and time. A nominalized version of this theory, by contrast, will

explain accelerations in terms of "intrinsic" masses which do not commit us to real numbers. It seems obvious that a theory which needs to invoke relations to real numbers to explain accelerations has less physical simplicity that one which does not.

So, to sum up, the arguments in this section and the last show that pure mathematics cannot be given a realist defence, either on the Quine-Putnam grounds that they are inextricable from scientific theories, or on the grounds that their addition to scientific theories adds to physical simplicity. Since I can think of no other prospects for a realist defence, and since we have already decided against anti-realist defences, I conclude that we ought to adopt a sceptical fictionalism about mathematics.

Bob Hale (1987, ch 5.II) has argued, against Field, that mathematical fictionalists cannot coherently maintain

(a) that mathematics is false

(b) that the truth of mathematics is possible, and

(c) that mathematics, if it were true, would have no consequences in the nominalist world that do not already follow from nominalist truths, as is required by Field's conservativeness claim.

Hale's thought is that if mathematic s is only contingently false, then its falsity ought to make some nominalist difference, ought to manifest itself by the failure of certain nominalist consequences.

However, I don't see why fictionalists cannot maintain the conjunction of (a), (b) and (c). Fictionalists are certainly committed to (a) and (b). By definition they think mathematics is false. And, since they take mathematics to be making meaningful existence claims, they do not think that its truth is ruled out by logic or concepts alone. But there seems no obvious reason why they should not believe (c) too, and deny that the contingent falsity of mathematics need show up in the nominalist world. After all, the fictionalist objection to mathematics is not that we can detect definite symptoms of its falsity in the natural world, but rather that neither this world nor anything else provide us with any grounds for believing it.

If there is an objection here, it is that lack of grounds for belief might seem only to warrant suspension of belief, rather than active disbelief. In the face of this objection, the fictionalist could simply acquiesce, and agree that we ought to combine our fictionalist "acceptance" with a neutrality of belief, rather than an outright rejection of mathematics. But this seems unnecessarily weak-kneed. On our current understanding, mathematical theories invite us to inflate our ontology by adopting synthetic bridge principles. If there are no positive grounds for these principles, other than that the entities they posit are possible, then surely the appropriate attitude is disbelief rather than neutrality. If someone urges that there are little green men on the first planet of Proxima Centauri, but by way of evidence offers us only that these men are possible, then surely we ought to reject this claim outright, rather than afford it the courtesy of agnosticism.

6.11 Morality and Modality

Many of the above arguments about mathematics are parallelled for morality and modality. In both these areas anti-realism seems initially promising: that is, it seems initially plausible that

truth, for either moral or modal judgements, might be conceptually guaranteed by the availability of some discursively authorized way of establishing these judgements.

However, when we turn to the details of our agreed discursive practices, they seem to hinge on certain crucial assumptions, analogous to the arithmetical N=, which provide a bridge between non-moral or non-modal claims, on the one hand, and distinctively moral and modal claims, on the other. For example, in morality we have bridge principles such as that killing people is b ad, and that causing happiness is good; in modality we have bridge principles such as that anything provable from no premises is necessary.[15]

But, now, when we ask for the warrant for these claims, epistemological difficulties arise. Reductionist readings, which say that "good" means nothing more than "increased happiness", or that "necessary" means nothing more than "provable from zero premises", seem clearly not to do justice to the intended meaning of the moral and modal terms. Yet once we allow that moral and modal claims take us beyond claims about happiness and provability, then we can no longer assume that the crucial bridge principles are simply analytic truths guaranteed by the meanings of the terms involved. And this undermines the prospects for an anti-realist defence of moral and modal claims. For if the bridge principles are substantial synthetic claims, we cannot simply maintain that they are somehow automatically guaranteed by the structure of our discursive practice.

On the other hand, there seems little likelihood of a realist vindication of such judgements, a vindication which accepts that the content of moral and modal claims take us beyond what our discursive practices automatically guarantee, but which nevertheless argues on a posteriori grounds that these practices reliably generate truths. So the only option left seems to be some version of sceptical disbelief.

These moves are perhaps most familiar in the case of morality. Thus Hume long ago pointed to the non-analyticity of moral bridge principles, by observing that you can't infer an "ought" from an "is"; and G.E. Moore's analogous objection to the "naturalistic fallacy" is that it always makes meaningful sense, given any natural description of some situation, to query whether it satisfies any further moral description (Moore, 1903). These doubts about the analyticity of moral axioms make it difficult to provide any anti-realist epistemology for morality. On the other hand, once we do allow that the content of moral claims transcends the moral evidence, there seems little hope of a realist demonstration that this evidence is nevertheless a reliable guide to the moral truth. (Moore's intuitionism can be seen as a desperate attempt to provide such a realist epistemology.) So we seem pushed towards scepticism, and indeed we find this position explicitly defended in, for example, J.L. Mackie's error theory of ethics (Mackie, 1977).

The same moves are also discernible, if somewhat less familiar, in the philosophy of modality. Thus Simon Blackburn (1986, pp 120-1) argues that any attempt to build an analytic bridge between non-modal premises and modal conclusions must fail to do justice to the distinctively modal character of those conclusions: how can the fact that our language happens to contain certain proof procedures guarantee that certain things are necessary? Yet, if such bridge principles aren't analytic, it is difficult to see how modal anti-realism can work. On the other hand, a realist epistemology for modality seems even less likely than for morality. So again we seem pushed towards some kind of fictionalist scepticism.

6.12 The Non-Doxastic Alternative

However, there is an alternative to scepticism about morality and modality. So far I have been taking it for granted that moral and modal claims are expressions of belief. But perhaps we shouldn't read them in this way in the first place: instead of reading them as beliefs about some distinctive kind of non-natural fact, perhaps we should read them as expressing a distinctive non-doxastic attitude towards natural facts. Thus, as a first shot, there is the option of reading moral judgements as expressing some kind of impartial approval.[16] And, similarly, there is the option of reading modal judgements of necessity as expressing our unqualified commitment to certain forms of argument.[17]

If such a non-doxastic account of moral and modal claims is right, then the arguments for scepticism fall away. It will no longer be an immediate objection to the relevant bridge principles, for instance, that they are synthetic assertions that lack empirical evidence. For, on the construal now being considered, the bridge principles will not be assertions at all, but rather prescriptions or permissions about adopting certain non-doxastic attitudes when certain natural facts obtain. That "killing people is bad" will no longer license inferences to moral beliefs about acts of killing, but simply prescribe a negative moral attitude to such acts. Similarly, that "propositions provable from zero premises are necessary" won't entitle us to any beliefs about necessity, but just endorse our unqualified commitment to syntactically valid arguments. There will of course remain room to debate the appropriateness of these prescriptions and permissions. But the lack of empirical evidence for the corresponding assertions will cease to be an obvious objection.

This line of thought thus points to the possibility of a position about morality and modality, which would be manifestly different from scepticism, in that it would recommend upholding our existing moral and modal judgements, rather than rejecting them. But this would not be because it recommended believing them, on either realist or anti-realist grounds, but rather because, on a proper understanding of their significance, questions of belief would simply not be at issue.

This non-doxastic suggestion about morality and modality raises some obvious questions. What justification is there for denying that moral and modal claims are expressions of belief, especially given that they share many features of normal belief-expressing assertions? And, if is right to read moral and modal claims in this way, then why isn't it right to read mathematics similarly non-doxastically, and thereby avoid scepticism about mathematics?

In the next two sections of this chapter I shall try to answer these questions. A final section will then say a bit more about how a non-doxastic interpretation might work for modality. This issue, like the analogous issue about morality, about which I shall say little further, is far too large a topic to treat properly here. But since the fictionalist account of mathematics needs various modal claims, it will be appropriate to finish this chapter with some brief suggestions on the topic.

6.13 Why are Morality and Modality Non-Doxastic?

My suggestion is that moral and modal utterances be understood as expressing some attitude other than belief. However, I can imagine that some readers will be dubious about the force of this idea. What exactly is the significance of denying moral and modal judgements the title of "belief"? In particular, what real difference is there then between my position, and that of an anti-realist who says that truth, for a moral or modal belief, is something conceptually guaranteed by the availability of an authorized derivation within our existing practice? After

all, the point of the non-doxastic option is supposed to be that it allows us to continue upholding moral and modal judgements. But if I thus allow that our existing practice for making such judgements is entirely appropriate to the content of those judgements, and that we are therefore quite entitled to the moral and modal judgements that we make, then what makes me different from the anti-realist who argues on just these grounds that we should uphold moral and modal beliefs?

This query can be made even more pressing if it can be shown, as arguably it can, that the structure of moral and modal discourse closely matches the structure of discourses which do express beliefs, in respect of conforming to the canons of logic, allowing conditional constructions, accepting that currently accepted views may come to be overturned, and so on.[18]

However, I think that there is a good reason why it would be wrong, despite all these similarities, to assimilate moral and modal judgements to normal beliefs. Namely that, on our present assumptions, moral and modal judgements behave differently from beliefs in epistemological contexts. I argued in section 6.11 that, if moral and modal judgements were beliefs, then the bridge principles which take us from natural premises to moral or modal conclusions would be unwarranted. It is only when we view moral and modal conclusions as non-doxastic that we are free to read those bridge principles as acceptable prescriptions, rather than unevidenced beliefs. So we need to deny the title of belief to moral and modal claims, if we want to continue upholding them.

It might seem as if this reason for denying the title of belief to moral and modal claims is purely a philosopher's reason, which doesn't reflect anything in the cognitive workings of ordinary people, but only philosophical anxieties about epistemological "justification". However, I don't think that questions about justification can be sliced off from the first-order reality of psychological states in this way.

By way of analogy, consider the case of a scientist who uses a theory about unobservables for making observational predictions, accounting for observable results, designing experiments, and so on. According to Bas van Fraassen (1980), somebody can do all this, and yet be an instrumentalist in not believing what the theory says a bout unobservables. Paul Horwich (1991) has queried the cogency of Van Fraassen's position here, on the grounds that, once scientists "think with" a theory to the extent that Van Fraassen allows, then there remains no substance to the thought that they do not believe the theory.

However, I think that substance can be given to this thought, contra Horwich, by taking into account the scientists' epistemological dispositions. Imagine that it is pointed out to certain scientists that, despite its usefulness, a given theory does not satisfy the epistemological requirements normally imposed on beliefs about things we are not directly acquainted with. The scientists' subsequent response will disclose whether or not they believe the theory: if they are unperturbed by the theory's admitted epistemological shortcomings, then we can infer that their attitude is not belief, but merely instrumentalist acceptance; on the other hand, if they are unable to view the shortcomings with equanimity, and accept that they require a change of attitude, then this shows that their original attitude was belief.[19]

The point of these last remarks about scientific theories has been to show how belief is distinctive among psychological attitudes in answering to certain epistemological obligations. And so, to return to the main thread of argument, if moral and modal attitudes do not answer

to these obligations, then this is itself a good reason for placing these attitudes outside the category of beliefs. This now explains why the non-sceptical position about morality and modality I have outlined in this section is genuinely different from anti-realism. Anti-realism needs to show that our practice is adequate for generating moral and modal beliefs. I agree that our practice is adequate, but say that this is precisely because moral and modal claims aren't beliefs. (If they were beliefs, they would then be subject to certain damaging epistemological requirements, which would then leave us with scepticism, not anti-realism.)[20]

It is perhaps worth emphasizing that in denying the title of belief to moral and modal judgements, I do not want to suggest that they are unimportant, or that the difference between correct and incorrect such judgements is somehow arbitrary. On the contrary, I fully accept that both moral and modal judgements play a central role in the thinking and action of human beings, and that their doing so requires that they conform to fully impartial standards of correctness. I would also like to emphasize that I not want to disqualify moral and modal judgements from the category of beliefs just because they report on matter which are non-natural, or abstract, or outside the causal world of space and time. I have been happy to concede the possibility of such non-natural beliefs from the beginning of this chapter. Indeed this is precisely my view of mathematical judgements: I take these to be beliefs about non-natural states of affairs (albeit beliefs that we have no warrant for adopting).

6.14 Why isn't Mathematics Non-Doxastic?

This brings us to the next question: namely, if moral and modal attitudes can be saved from scepticism by the non-doxastic option, why isn't the same true of mathematics? Why has my analysis of mathematics led to scepticism, rather than the endorsement of some non-doxastic species of mathematical judgement?

After all, isn't the fictionalist attitude to mathematics itself an instance of such a non-doxastic attitude? Once we embrace fictionalism, then we uphold mathematical judgements -- not as beliefs, it is true , but nevertheless as claims that can properly be accepted-in-the-mathematical-fiction. So why isn't this just another range of judgements that can be upheld because they don't express beliefs?

Well, once we do embrace fictionalism, then the resulting attitude to mathematical judgements does evade scepticism. But the difference between mathematics, on the one hand, and morality and modality, on the other, is that, prior to philosophical argument, most people adopt an attitude of belief to mathematics, whereas I take it that moral and modal attitudes are already different from belief. Fictionalism, as a philosophy of mathematics, is thus advocating a revision of everyday thinking, where the non-doxastic account of morality and modality is happy to leave everything as it is. This is why fictionalism in the philosophy of mathematics a sceptical doctrine: it is sceptical about the beliefs that most people have, whereas the non-doxastic account of morality and modality has no corresponding objection to everyday thought, since it doesn't take every day thought to involve modal and moral beliefs.

Is this contrast justified? Is it true that most people believe mathematical claims, but express different attitudes in their moral and modal judgements? Well, this seems plausible to me, but I do not need to defend the thesis here. For it is an empirical matter, a thesis about the psychology of actual individuals, not a philosophical issue . In the last section I showed that there is a real difference between believing moral, or modal, or mathematical claims, and having a non-doxastic attitude to them. But, given this, the further question of how many people actually hold belief s, and how many hold non-doxastic attitudes instead, is a question

for sociologists, not philosophers. The essential philosophical point can be put hypothetically, in a way that abstracts from the actual psychology of individuals, and so applies equally to mathematics, morality and modality: if people adopt beliefs, then they are wrong, and they should reject those beliefs, in a sceptical spirit, and switch to some less demanding non-doxastic attitude instead.

Still, as I said , it seems plausible that mathematics ordinarily involves beliefs in a way that morality and modality does not, and I shall continue to speak accordingly. It may be helpful to offer a possible empirical explanation for this conjectured empirical contrast. Note that it is entirely natural to view mathematical judgements, whatever their doxastic status, as essaying reference to a distinctive range of objects, like numbers, sets, vector spaces, and so on. With moral and moral judgements, by contrast, it is by no means so obvious there is any intended reference to distinctive objects. So there is a sense in which mathematical thought doesn't need any attitude other than belief to constitute itself as a distinctive mode of discourse: its range of distinctive objects already ensures that mathematical claims have a distinctive content. By contrast, if moral and modal judgements do not refer to any distinctive range of objects, then there remains a question about what gives those judgements their distinctive contents; and a natural answer to the question is that they express attitudes other than beliefs.

This story isn't incontestable. For a start, you might question whether a lack of intended reference to distinctive objects does distinguish moral and modal claims from mathematical claims. And, second, even if you do accept this, it doesn't automatically follow that mathematical claims must be understood as expressing beliefs, and moral and m odal claims other attitudes.

Let me take the latter question first. Even if, as I have been suggesting, moral and modal claims don't have any objects of their own, it might still be possible to understand them as expressing distinctive kinds of belief: for nothing I have said so far rules out the possibility that such operators as "it is right that", and "necessarily", yield beliefs when applied to contents, rather than non-doxastic attitudes. (Though the arguments of 6.11 would then still indicate a sceptical attitude towards these beliefs.) Nor, conversely, does the intended reference to distinctive mathematical objects force us to view mathematical claims as beliefs: after all, in a community of self-professed fictionalists, claims about numbers, sets, and so on, would express fictional acceptance, rather than belief.[21] Nevertheless, even if there is no logical tie, in either direction, between distinctive objects and the adoption of belie fs, the differing involvement of objects still seems to me to offer a likely empirical explanation of why people should have mathematical beliefs, but non-doxastic moral and modal attitudes. For even if it is logically possible to combine objects with the absence of belief, and vice versa, it still seems psychologically plausible that people will adopt the attitude of belief to the objects they are introduced to in mathematics, but non-doxastic attitudes to non-object-involving moral and modal claims.

There is also the former question, about the object-involving contrast: am I right to hold that mathematics involves intended reference to objects, while moral and modal claims do not?

The first part of this claim has already been established: we have already examined accounts of mathematics which represent it as free from commitment to numbers, sets, and other mathematical objects -- namely, if-thenism and reductionism -- and reject ed them, precisely on the grounds that such non-objectual readings are not faithful to the standard content of mathematical claims.[22]

The converse issue, however, is less clear-cut. Thus there is the well-known "possible worlds" interpretation of modal discourse, which takes modal judgements to commit us to a range of non-actual universes. And similarly it is possible to construe moral claims as essaying reference such distinctive entities as rights and wrongs, virtues and vices. I do not myself think these objectual readings of everyday modal and moral claims are compelling, but I shall not argue the point here, for little of philosophical substance hangs on it. The question of whether moral and modal claims involve reference to distinctive objects is once more an empirical issue, a matter of the actual contents of the thoughts of actual individuals. And I have put forward the empirical hypothesis that they do not so refer only in order to explain another empirical conjecture, namely, the conjecture that such claims do not express beliefs, but some other non-doxastic attitude.

My only substantial philosophical contention remains the hypothetical thesis that, if anybody were to ado pt the attitude of belief to moral and modal claims, then the resulting beliefs would be epistemologically unjustified. And to this thesis the possibility of object-involving interpretations of moral and modal discourse poses little threat. Fo r if modal beliefs commit us to possible worlds, or moral beliefs to rights and wrongs, it will surely be harder, not easier, to justify modal and moral beliefs.

6.15 Fictionalism and Modality

In this final section I want to say a bit more about modality, and in particular about the use that fictionalism makes of this notion.

Note first that fictionalism needs to make assumptions about logical consequence, in at least two places. First, and most obviously, the claim that mathematics is dispensable in science rests on the premise that it is conservative with respect to nominalist truths -- that is, that no nominalist conclusions follow logically from nominalist-premises-plus-mathematics that don't follow from nominalist premises alone.

Second, fictionalists need the notion of logical consequence to identify what they mean by "mathematics" in the first place. For the kind of mathematics that fictionalists think is dispensable, though nevertheless useful as a fiction, isn't just any old set of claims formulated in mathematical vocabulary (for that wouldn't be conservative, or useful), but rather standard, or accepted, or good mathematics. So fictionalists owe us so me account of what standard mathematics is, some account of what exactly it is that they are recommending we accept as a fiction. And the obvious account is that standard mathematics consists of all the claims that follow logically from standard mathematical assumptions.

I shall now show that we fictionalists need to understand such claims about logical consequence in modal terms: that is, we need to understand the claim that B follows logically from {A} as equivalent to it is not possible that {A} all be true and B false.

This may not be immediately obvious. Why can't the fictionalist avoid modality by simply appealing to the standard characterizations of logical consequence in metalogic? For example, why not explain consequence semantically, saying B is a logical consequence of {A} if and only if B is true in all models in which all members of {A} are true? Alternatively, why not offer a syntactic analysis, saying that B is a logical consequence of {A} if and only if B can be proved from {A} using a specified set of rules of inference?

The characterizations are unquestionably of great mathematical significance. However, there are reasons why we fictionalists cannot rest with either of them as a philosophical account of logical consequence. Let me take the semantic characterization first. The obvious problem is that models are themselves abstract mathematical objects, and therefore, given the overall argument of this chapter, not something we can have beliefs about.[23] (In the end, of course, we should be able to regard such claims about models, along with other mathematical claims, as useful parts of the mathematical fiction. Bu t at this stage we still face the task of explaining what exactly the mathematical fiction comprises.)

The alternative metalogical characterization of logical consequence, in terms of syntax, specifies a set of rules of inference for moving between sentences with certain syntactic forms. The normal objection to fictionalists accepting this characterization is that rules of inferences and syntactic forms are themselves abstract objects, and so inadmissible from our fictionalist perspective.[24] But this is less than compelling; there seems plenty of room to view such syntactic entities as features of certain physical systems, namely, languages. However, there are worse problems facing a fictionalist who appeals to the syntactic explanation of logical consequence. These arise from the fact that nothing stronger than first-order logic can be completely characterized in syntactic terms. So, for a start, there is a problem about the fictionalist delineation of standard mathematics as whatever follows from the standard axioms. If "follows from" simply means whatever follows by first-order logic, then Godel's theorem tells us that there will be mathematical truths which do not so follow from the standard axioms, contrary to the fictionalist's delineation of standard mathematics.

A related Godelian difficulty faces the fictionalist's claim that standard mathematics is conservative with respect to bodies of nominalist assert ions. For consider, as the relevant body of nominalist assertions, the claim that there exists a geometrical point, and another point, and then another as far away again in the same direction, . . . By such means we can construct a nominalist version of Peano's axioms, which refers to geometrical points instead of the natural numbers. But now there will be a "nominalized version" of a Godel sentence which does not follow logically from these axioms, if "follow from" means by first-order logic. However, this nominalized Godel sentence will follow in first-order logic if we are allowed to add pure arithmetic plus appropriate bridge principles to the nominalist Peano axioms, since pure arithmetic will include the pure version of this Godel sentence. The upshot is that pure mathematics, and in particular pure arithmetic, is not conservative with respect to bodies of nominalist assertions, if by "conservative" we mean that the addition of mathematics to nominalist assumptions gene rates no new conclusions within first-order logic.[25]

So a fictionalist cannot happily rest either with the model-theoretic or with the syntactic characterization of logical consequence. Suppose, however, that we take the modal notions of necessity and possibility as primitive, and define consequence in modal terms in the way suggested above, as the impossibility of the premises being true and the conclusion being false. This then evades the difficulties facing the standard semantic and syntactic characterizations. There is no obvious commitment to abstract objects like models in this definition. And since there is no reason to regard the consequence relation thus defined as restricted to first-order consequence, the Godelian difficulties need no longer apply.

If the arguments of the last few sections are right, it follows that the appropriate attitude to claims of logical consequence will not be belief, but rather a non-doxastic attitude of unqualified commitment to the corresponding forms of argument. Of course, there is nothing

to stop us believing that, whenever the premises of such arguments are true, then the conclusions will be true too, that is, that such arguments are reliable. But the further thought, that these arguments are necessarily reliable, will not itself be a belief.[26]

As I observed in passing earlier, a non-doxastic view of necessity raises the question of whether our normal criteria for making judgements of necessity provide appropriate grounds for the relevant non-doxastic attitude. In the present context, where we are using necessity to explain logical consequence, there are also a number of further technical issues, which I cannot pursue here, about whether these criteria are adequate to the mathematical structure of logical consequence.[27] However let me make just one point. A standard soundness proof for some form of argument, such as is given by the truth table for modus ponens, say, provides an obvious vindication of an unqualified commitment to the reliability of that form of argument. Such a proof can simply be thought of as arguing in the alternative, for all the possible alternative arrangements of semantic values for non-logical expressions which would make the premises true, that the conclusion would be true too.[28] So such a soundness proof provides an immediate basis for belief in the reliability of the form of argument in question. And since the proof hinges on no assumptions save those about the meaning of logical expressions, it also provides obvious grounds for an unqualified commitment to that form of argument.

1. There is a problem of terminology here. In some circles, especially American ones, philosophers like Bas van Fraassen are called "anti-realists", not because they hold that there is no substantial possibility of erroneous belief, but, on the contrary, because they fear that this possibility is actual (cf Van Fraassen, 1980). However, in my terminology, and in contemporary British usage, Van Fraassen is not an anti-realist, but rather a pessimistic realist, that is, a sceptic. To get things straight , we need to distinguish three positions: anti-realism, in the British sense, which denies the conceptual possibility of error; optimistic realism, which admits the conceptual possibility of error, but disputes its actuality; and pessimistic realism, or scepticism, which fears that error is not only possible but actual. In what follows, unqualified uses of the terms "anti-realism", "realism", and "scepticism" should be understood to stand for these three positions respectively. That is, I shall reserve the term "anti-realism" for philosophers who uphold beliefs on the grounds they can't be false; philosophers who reject beliefs because they fear they are false will be called "sceptics".

2. For more on the different varieties of anti-realism and their problems, see Papineau (1987, ch 1).

3. I would make the same point about contemporary "structuralist" or "modal-structuralist" views, which read mathematics as saying only that there exist, or possibly exist, some objects satisfying the axioms, and that therefore the theorems are true of those objects, or of those possible objects, whatever they might be. (Cf Lewis, 199x, pp xx-xx; Hellman 1989, 1990.) Whatever other virtues these views may have (but cf footnote 22 below), they are unquestionably revisionary proposals. The same goes for Michael Resnik's (1981, 1982) more platonist species of structuralism; this also faces extra problems, because of its reification of "structures" (cf Chihara, 1990, ch 7).

4. I have found this position defended more often in conversation than in writing.

5. Are these stories categorical? The plethora of Santa Clauses who appear around Christmas might make us wonder. But such stories can easily be made categorical, by including the

explicit provisos that there is only one genuine Santa Claus, only one genuine Sherlock Holmes, and so on.

6. It is true that (6) is a truth of second-order logic, and some philosophers will feel that this commits (6) to sets, thus reintroducing the epistemological difficulties of abstract objects. But this is by no means uncontentious: in a series of recent papers George Boolos (1975, 1984, 1985) has defended higher-order quantification against the charge of implicit reference to sets. See also Wright (1983) pp 132-3.

7. For an interesting recent version of this generally Russellian approach, see Hodes (1984). Cf also Lear (1982). It is worth distinguishing this "reductionist" approach from the "if-thenism" mentioned in section 2. Both approaches claim that there is nothing more to mathematical knowledge than logical knowledge. But "if-thenism" does so by arguing that mathematical knowledge is always knowledge that if such-and-such axioms hold, then certain theorems follow. The reductionist approach, by contrast, needs to show that the axioms (and so the theorems) are themselves logical truths.

8. Hodes (1984) holds that the construction "the number of . . ." is systematically ambiguous, but gets disambiguated when the gap is filled in. This is reasonably plausible. What is not so plausible is the claim that "2" is ambiguous in "2 + 3 = 5".

9. Cf Hodes op. cit. p 144-6; Lear op. cit. p 188-91.

10. Note however that reductionism, as I have characterized it, is technically more demanding than Field's fictionalism: the reductionist needs to find, for every mathematical claim, some (family of) quantificational surrogate(s) which yields the same inferential power; while Field is only committed to holding that all inferences underpinned by mathematical claims can be made by logic alone, and not to a case-by-case pairing of mathematical claims with quantificational equivalents.

11. It is perhaps unfair to accuse Hodes (op. cit.) of wanting to have it both ways, since he explicitly embraces a kind of fictionalism (p 146), and to that extent explicitly abandons his reductionist ambitions. Lear (op. cit.), on the other hand, does seem to want to have it both ways. He shows how the possibility of holding reduced beliefs which do not involve abstract objects makes it both harmless but useful to work with mathematical propositions that do. But he the n claims that this legitimates belief in the mathematical propositions.

12. This provides a route to knowledge of arithmetic. But what about the rest of mathematics? Well, it is arguable that the rest of mathematics is reducible to set theory. Moreover, there is a plausible set-theoretical analogue to N=, namely, the conceptual equivalence of:

(9) (x)(Fx <-> Gx), and

(10) The set of Fs = the set of Gs.

But of course, as it stands, this is too strong: without some restrictions on what can be substituted for F and G, Russell's paradox will follow. Still, there remain weaker equivalences which are both pre-theoretically plausible and powerful enough to yield set theory. I shall not pursue this line of thought, however, since the criticisms I am about to

make of Wright's account of arithmetic will carry over to any analogous account of set theory.

13. For details of this line of argument, see Hale (1987, ch 2).

14. Of course , theoretical claims aren't conservative with respect to observational claims in the strong sense that extra theoretical premises never augment the consequences of any set of observational premises. But Craig's theorem does show that adding the claims of some theory to the observational consequences of that theory does not augment observational consequences. So to that extent theories are dispensable for drawing observational conclusions.

15. In what follows "necessary" should be understood in the narrow sense of logically necessary. I take it that other kinds of necessity (physical, conceptual, legal, and so on) can be defined in terms of logically necessity (as necessary consequences of physical laws, conceptual laws, legal laws, and so on). In the case of physical necessity, there is of course the extra problem of distinguishing physical laws from accidentally true generalizations. In my (1986a) I advocated a fictionalist view of this distinction. I now think this was a mistake. My current view is that we can distinguish physical laws as consequences of those true generalizations which have sufficient robustness to qualify as causal. But that is another story.

16. As, for example, in Ayer (1936, ch 6).

17. Cf McFetridge (1990, essay VIII).

18. Simon Blackburn (1984, 1986) has coined the term "quasi-realism" to emphasize the structural affinities between normal discourse and moral and modal discourse. As it happens, Blackburn also upholds the "projectivist" view that moral and modal judgements express attitudes other than belief. The question I am currently asking (though Blackburn does not) is why his "quasi-realism" doesn't undermine his "projectivism". Fo r this point, see Wright (1987). See also Hale (1986) for further discussion of Blackburn's position.

19. So I agree with Van Fraassen, and disagree with Horwich, that it is psychologically possible to "think with" a theory, and yet not believe it. But I certainly don't agree with Van Fraassen's further sceptical claim that this is the appropriate attitude to all scientific theories. Cf my remarks about the epistemology of theory-choice in 6.10 above.

20. Why exactly should belief be subject to epistemological requirements not imposed on other attitudes? A short answer is that belief is that attitude which is supposed to represent how things are, as opposed to how they are taken to be. This is why beliefs require evidence, and cannot be made true just by being part of some established intellectual practice.

21. Note that the intended reference to mathematical objects is the reason why fictionalism is the appropriate non-doxastic attitude in mathematics. I have already explained why this non-doxastic attitude will involve a sceptical element: it requires us to reject the mathematical beliefs I conjecture most people to hold. But such a sceptical attitude needn't be fictionalist. To see this, imagine a community who did have moral beliefs whose content derived from a non-object-introducing moral operator of the kind mentioned above. Then, I say, we ought to reject those beliefs in favour of a non-doxastic moral attitude. This attitude would thus be

sceptical about those people's moral beliefs. But it wouldn't be fictionalist, for lack of any moral objects to populate the fiction. In mathematics, by contrast, we have intended mathematical objects to provide our fiction.

22. Even if such non-objectual views are wrong about the meaning of mathematical claims, might they not be defended as revisionary proposals? This is possible, but there is an obvious respect in which the revision proposed by fictionalism is preferable. Fictionalism only requires a revision in attitude, where these alternatives require a revision in content. Apart from this, the explanations of the applicability of mathematics offered by if-thenism and reductionism are technically more demanding than that offered by fictionalism. The extra requirements facing reductionism are those noted in footnote 10 above. The problem facing if-thenism is that it can only account for the applicability of mathematical theories to sets of natural objects which provide models of those theories; yet we often apply mathematical theories to sets of natural objects which are too sparse to yield such models.

23. For other objections to the semantic characterization of logical consequences, which apply even if you don't mind abstract objects, see Field (19xx) and McGee (1991).

24. Cf Putnam, 1971, ch 2; Field, 1984, p 514.

25. This Godelian argument is due to Shapiro (1983). At first sight Shapiro's argument might seem inconsistent with the proof of the conservativeness of mathematics given by Field in Science Without Numbers. However, what Field proves is that mathematics is conservative with respect to first-order nominalized theories. The kind of geo metrical theory needed to mimic Peano's postulates, by contrast, requires more than first-order quantification, which is why it escapes Field's proof. This leaves Field with a problem, however, since he himself requires just this kind of nonfirstorderizable geometrical theory to nominalize physics. (Cf Chihara 1990, ch 8 and Appendix.)

26. This answers a point raised by Hale (1987, p 120).

27. But see Field (19xx) for an investigation of a primitively modal interpretation of logical consequence. I should make it clear that Field does not himself advocate a non-doxastic approach. But it seems to me that many of his points could be adopted by someone who does.

28. As a fictionalist, I don't want to read such soundness proofs as showing that the conclusion is true in all models in which the premises are true. My idea is rather that they assume that (in actuality) either X or Y or . . ., and conclude that (in actuality) either the premises are false or the conclusion true.